Федеральное государственное бюджетное учреждение науки
Институт проблем управления им. В.А. Трапезникова
РОССИЙСКОЙ АКАДЕМИИ НАУК

Научно-производственное объединение
«Информационные и сетевые технологии»

Институт информационных и телекоммуникационных технологий
БОЛГАРСКОЙ АКАДЕМИИ НАУК

РАСПРЕДЕЛЕННЫЕ КОМПЬЮТЕРНЫЕ И ТЕЛЕКОММУНИКАЦИОННЫЕ СЕТИ: УПРАВЛЕНИЕ, ВЫЧИСЛЕНИЕ, СВЯЗЬ (DCCN-2015)

МАТЕРИАЛЫ ВОСЕМНАДЦАТОЙ МЕЖДУНАРОДНОЙ НАУЧНОЙ КОНФЕРЕНЦИИ

(19–22 октября 2015 г., Москва, Россия)

Под общей редакцией д.т.н. В.М. Вишневского

Москва
ИПУ РАН
2015

V.A. Trapeznikov Institute of Control Sciences of Russian Academy of Sciences

Research and development company **"Information and networking technologies"**

Institute of Information and Communication Technologies
Bulgarian Academy of Sciences

DISTRIBUTED COMPUTER AND COMMUNICATION NETWORKS: CONTROL, COMPUTATION, COMMUNICATIONS (DCCN-2015)

PROCEEDINGS OF THE EIGHTEENTH INTERNATIONAL SCIENTIFIC CONFERENCE

(19–22 october 2015 г., Moscow, Russia)

Under the general edition of Dr. of Computer Science V.M. Vishnevskiy

Moscow
ICS RAS
2015

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-2015) = Distributed computer and communication networks: control, computation, communications (DCCN-2015) : материалы Восемнадцатой междунар. науч. конфер, 19–22 окт. 2015 г., Москва: / Ин-т проблем упр. им. В.А. Трапезникова Рос. акад. наук ; под общ. ред. В.М. Вишневского – М.: ИПУ РАН, 2015. – 656 с.
– ISBN 978-5-91450-170-6.

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

- Архитектура компьютерных и телекоммуникационных сетей.
- Управление в компьютерных и телекоммуникационных сетях.
- Оценка производительности беспроводных сетей трансляции мультимедийной информации.
- Аналитическое и имитационное моделирование сетевых протоколов.
- Теория очередей и теория надежности.
- Беспроводные сети IEEE 802.11, IEEE 802.15, IEEE 802.16 и UMTS (LTE).
- Технология RFID и ее применение в интеллектуальных транспортных системах.
- Проектирование протоколов (MAC-уровня) сантиметрового и миллиметрового диапазона радиоволн.
- Интернет, веб-приложения и услуги.
- Интеграция приложений в распределенных информационных системах.

В материалах конференции DCCN-2015, подготовленных к выпуску Козыревым Д.В. обсуждены перспективы развития и сотрудничества в этой сфере.

Сборник материалов конференции предназначен для научных работников и специалистов в области теории и практики построения компьютерных и телекоммуникационных сетей.

Текст воспроизводится в том виде, в котором представлен авторами

Утверждено к печати Программным комитетом конференции

ISBN 978-5-91450-170-6

Содержание / Contents

A TCP-LIKE CONTROL OF NOTIFICATION DELIVERY FOR SUBSCRIPTION OPERATION IN SMART SPACES

D. Korzun[1], *M. Pagano*[2], *A. Vdovenko*[1]
[1] Petrozavodsk State University, Petrozavodsk, Russia
[2] University of Pisa, Pisa, Italy
dkorzun@cs.karelia.ru, m.pagano@iet.unipi.it, vdovenko@cs.karelia.ru

Abstract

The paper studies the subscription operation in smart spaces for the case when the notification delivery to a client is subject to losses. We consider a TCP-like control of the check interval for adaptation to the observable loss rate. We propose an adaptive strategy for a client to manage the check interval based on the number of notification lost. The performance is analyzed and compared with other strategies under simplified assumptions about the notification loss distribution.

Keywords: smart spaces, mobile clients, control of notification check interval

1. Introduction

Information sharing in a smart space employs the subscription operation for content change detection and subsequent delivery of the notification to clients [4, 9]. Both change detection and notification delivery are subject to losses in wireless networked environments [10], widely used in emerging Internet of Things (IoT). In existing solutions, the major role is played by information brokers. In this paper, we study a solution closer to the Internet philosophy, in which the control of notification delivery is partially delegated to (typically mobile) clients. The latter actively request the broker for new notifications, in addition to default passive waiting for incoming notifications from the broker. A key performance parameter in this scenario is the check interval, which should be adapted to the observable loss rate. This issue resembles the congestion control in TCP (Transmission Control Protocol) [1, 2], by which the window size is reduced (multiplicative decrease) in case of losses and incremented (additive increase) otherwise.

This paper continues our research [10]. We consider a TCP-like control of notification delivery for subscription operation for the case of notification losses. In our solution, the client follows an adaptive strategy controlling the check interval based on the number of notifications lost in the latest window. We compare the performance of this adaptive strategy with other strategies under simplified assumptions about the distribution of the notification losses.

The rest of the paper is organized as follows. Section 2 states the notification loss problem for subscription operation in smart spaces. Section 3 introduces

adaptive strategies for a client to control the notification delivery in a TCP-like manner. Section 4 provides early simulation experiments to evaluate the performance of the proposed control strategies. Section 5 concludes the paper.

2. Notification Loss Problem

The pub/sub model is widely used for organizing multi-agent interactions in distributed systems [3]. We consider the use of pub/sub model in smart spaces [9, 6]. A smart space forms a sparse-connected multi-agent system deployed in a networked computing environment, typically with access to the Internet. Such an environment consists of various digital devices, including the growing family of IoT devices. Running on devices, software agents interact over the shared information content. This type of interaction involves a lot of informational sources and destinations in parallel and asynchronously. Information sharing makes the interaction indirect, based on a semantic information broker (SIB). The latter implements a common information storage and serves requests from agents on read/write operations.

The subscription operation specifies a persistent query from an agent (a subscription client) to SIB (a subscription server) for a particular part of the shared content. Whenever the specified part is changed the agent should receive the subscription notification. Changes are due to parallel activity of other agents, which act as publishers in this interaction. An agent may combine the roles of publisher and subscriber (client). SIB monitors all subscriptions of all clients and maps all incoming content changes to the specified interests. Therefore, changes are controlled on the SIB side, and corresponding notifications are sent to the clients. SIB acts as a passive receiver, and we call such subscription notifications passive [10].

We employ Smart-M3 as a reference platform for creating smart spaces [4]. For each subscription, the SIB maintains a persistent network connection (e.g., a TCP connection) established by the client's request [6, 7]. Knowing the set of all subscriptions, the SIB regularly checks that they are alive, removing the subscription if its network connection is lost. Smart-M3 follows the best effort style. A notification *should* be sent to a client if a related change in the content has happened. Some notifications can be unsent by SIB due to its overload or operability faults. SIB does not check delivery for already sent notifications. A notification can be sent although the underlying network connection is broken on the client side. The above properties do not ensure the dependable delivery even if reliable network protocols are used, such as TCP. Therefore, each client needs additional mechanisms for reducing the number of undelivered notifications. The obvious way is augmenting the passive notification delivery with an active control strategy the client performs individually on its own.

Let $i = 1, 2, \ldots$ be the event-based time evolution on the client side, where i is the index of notification events. An event i is either a passive notification (received from SIB) or an explicit check the notification delivery (made by the active client). Denote by t_i and k_i the time elapsed and the number of

losses between i and $i+1$, respectively. Assume that some initial t_0 is always defined. The values for k_i are non-negative integers. We consider the following alternatives for possible distributions of the notification losses.

1. Let the time between consecutive losses follow a uniform distribution $\mathcal{U}\Big\{at_0, bt_0\Big\}$. Hence, the average number of losses in any check interval is proportional to its length t_i

2. Let k_i follow a Poisson process of parameter λt_i. The number of losses during t_i has the probability mass function

$$\mathbb{P}(k_i = k) = \frac{(\lambda t_i)^k}{k!} e^{-\lambda t_i} \tag{1}$$

3. Let k_i follow a two-state alternated Poisson process: the loss rate is $\lambda_1 t_i$ and $\lambda_2 t_i$ with probability p_1 and $p_2 = 1 - p_1$, respectively. The assumptions $0 < p_2 < p_1$ and $\lambda_1 \ll \lambda_2$ describe that the network typically operates with moderate losses (λ_1) while from time to time the network suffers from high losses (λ_2), e.g., due to burst overload.

3. Control Strategies

Let the client have observed no losses during t_{i-1}, i.e., $k_{i-1} = 0$. It indicates that the system state becomes "good". To save its resources, the client increases additively $t_i = t_{i-1} + \delta$ with a fixed parameter $\delta > 0$. The increment is conservative since high increase of t_i is a clear risk for suffering a burst of losses.

Now let the client have observed losses, i.e., $k_{i-1} > 0$. Then, the client reduces t_i to decrease the number of losses in near future. The reduction is multiplicative since the client is interested in fast achieving $k_i = 0$. The reaction should take into account the previous observations, thus we apply the multiplicative average $t_i = \alpha t_{i-1} + (1-\alpha) t_{i-1}/(k_{i-1}+1)$ with a fixed parameter $0 \le \alpha < 1$.

As a result, one can construct the recurrent system that describes an *adaptive strategy* [10] by which the check interval t_i is reduced (multiplicative decrease) in case of losses and incremented (additive increase) otherwise.

$$t_i = \begin{cases} t_{i-1} + \delta & \text{if } k_{i-1} = 0, \\ \dfrac{1 + \alpha k_{i-1}}{k_{i-1} + 1} t_{i-1} & \text{if } k_{i-1} > 0. \end{cases} \tag{2}$$

Note that (2) is valid only for active control of subscription notifications. When a passive notification i is delivered then t_i cannot be set by the client.

Interestingly that this adaptive strategy provides a TCP-like control [1, 2]. One can consider t_i as a TCP congestion window and $\delta = 1$ means the additive increment by one full-sized segment. In the TCP case, observation of congestion makes $k_{i-1} = 1$, and taking $\alpha = 0$ leads to halving $t_i = t_{i-1}/2$ in (2).

In order to evaluate and compare the performance, we consider alternative strategies and present some auxiliary analytical results. The simplest approach is the strategy of a *constant check interval* when $t_i = t_0$ for $i = 1, 2, \ldots$. The mean and variance of the random variable K, which describes the number of losses, depend only on the loss distributions introduced in Section 2. In case of Poissonian losses, we have:

$$\mathbb{E}\Big[K\Big] = \mathrm{Var}\Big[K\Big] = \lambda t_0. \tag{3}$$

Since in many complex systems, randomness can improve the performance, an interesting strategy is *random selection* of the check interval. Intuitively, let t_i be chosen at random close to t_0. In particular, we consider a continuous uniform distribution $\mathcal{U}\Big\{t_0 - \Delta, t_0 + \Delta\Big\}$ for small $\Delta > 0$. For instance, in case of Poissonian losses, taking advantage of the laws of total expectation and variance (e.g., see [11]), we have

$$\mathbb{E}\Big[K\Big] = \lambda t_0 \qquad \mathrm{Var}\Big[K\Big] = \mathbb{E}_T\Big[\mathrm{Var}\Big[K|T\Big]\Big] + \mathrm{Var}_T\Big[\mathbb{E}\Big[K|T\Big]\Big] = \lambda t_0 + \frac{1}{3}\lambda^2\Delta^2, \tag{4}$$

where the mean value is the same as in (3), while the variance is increased due to the variability of the interval length. Roughly speaking, in this case the randomness does not change the average performance indexes, but increases the probability that a higher number of notification can be lost.

Finally, we consider a semi-adaptive approach, in which the check interval is halved in case of losses and set to the initial (reference) value t_0 otherwise:

$$t_i = \begin{cases} t_{i-1}/2 & \text{if } \; k_{i-1} > 0, \\ t_0 & \text{if } \; k_{i-1} = 0. \end{cases} \tag{5}$$

We shall refer this strategy as *multiplicative decrease*. The evolution of the check interval can be described by a discrete-time Markov chain [5]. State i corresponds to a check interval of length $t_i = 2^{-i}t_0$ and the only possible transitions from state i are back to state 0 (in case no losses are experienced) with probability p_i and to the following state $i+1$ (the check interval is halved in case of losses) with probability $q_i = 1 - p_i$, as shown in Fig. 1.

Due to the particular structure of the Markov chain, the state probabilities can be easily written as a function of π_0 (local balance equations):

$$\pi_{n+i} = \pi_0 \prod_{i=0}^{n} q_i \tag{6}$$

with the additional normalization condition

$$\sum_{n=0}^{\infty} \pi_n = 1 \quad \Rightarrow \quad \pi_0\Big[1 + \sum_{n=0}^{\infty}\prod_{i=0}^{n} q_i\Big] = 1. \tag{7}$$

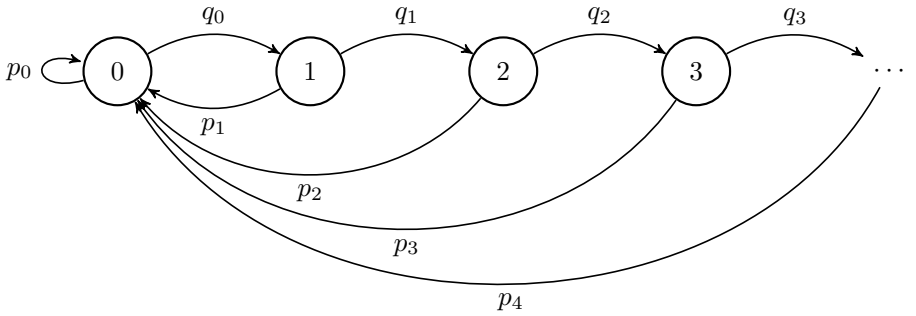


Fig. 1: Markov chain for the multiplicative decrease strategy.

Note that the sum converges under reasonable assumptions over the loss processes. (It is enough to assume that the loss probability goes to zero as $n \to \infty$.) The state probabilities permit to find all the relevant statistics, such as the average check interval

$$\mathbb{E}[T] = \sum_{n=0}^{\infty} \pi_n t_n = t_0 \sum_{n=0}^{\infty} \pi_n 2^{-n}.$$

The numerical values of p_n and $q_n = 1 - p_n$ depend on the loss distribution. In particular, for Poissonian losses we have

$$\begin{cases} p_0 = \mathbb{P}(K = 0 | T = t_0) = e^{-\lambda t_0}, \\ p_n = \mathbb{P}(K = 0 | T = t_n) = p_0^{2^{-n}} \qquad n \geq 1. \end{cases}$$

In the case of uniformly distributed losses, the transition probability $q_n = 1 - p_n$ is proportional to the length of the check interval $t_n = t_0 2^{-n}$, i.e. $q_n = q_0 2^{-n}$, and (6) can be rewritten as

$$\pi_n \;=\; \pi_0 q_0^n \left(\frac{1}{2}\right)^{n(n-1)/2}$$

Taking into account the specific values of p_n and q_n as well as the normalization condition (7), it is possible to calculate the state probabilities (at least numerically). In the special case in which the loss probabilities do not depend on the duration of the check interval (i.e., $q_n = q \;\forall\, n$), it is easy to find a close-form expression for the state probabilities

$$\pi_n \;=\; (1 - q) q^n$$

and calculate the average check interval

$$\mathbb{E}[T] = \sum_{n=0}^{\infty} \pi_n t_n = t_0 \frac{1 - q}{1 - q/2}$$

as well as its variance

$$\mathrm{Var}\Big[K\Big] = \mathbb{E}\big[T^2\big] - \mathbb{E}^2\big[T\big] = t_0^2 \frac{q(1-q)/4}{(1-q/4)(1-q/2)^2}$$

The assumption that the number of losses does not depend on the length of the check interval might be unrealistic, but in any case it provides a lower bound for the average (and a upper bound for the variance) wrt the more realistic assumptions described in Section 2.

Moreover, it is worth mentioning that in practice it is meaningless to use very short check intervals; this corresponds to truncate the Markov chain at state N, remaining in that state in case of further losses (and going back to state 0 with probability p_N as usual). Hence, (6) is still valid for all the states up to $N-1$, while

$$\pi_N = \pi_0 \frac{1}{1-q_N} \prod_{i=0}^{N-1} q_i$$

and the normalization condition (7) involves just $N+1$ terms.

4. Simulation Experiments

The simulation applies the notification loss distributions from Section 2. In all experiments, we take $t_0 = 20$ s and our simulation parameters are summarized in Table 1. Note that every 20 seconds one notification is lost on average for any of the three considered distributions.

Our simulation parameters for the four experimented control strategies are summarized in Table 2. The average check interval is t_0 for the two strategies: constant check interval and random selection. The settings give some odds to these two strategies, which thus can adjust to the average loss rate.

For sake of brevity, in the following we shall discuss only the results for Poissonian losses. An example experiment is shown in Fig. 2. A clear observation is that the adaptive strategy makes the check interval longer compared with the other strategies.

Table 1: Simulation parameters of the notification loss distributions.

Uniform losses		Poissonian losses	Two-state alternated Poissonian losses			
$a = 0$	$b = 0.1$	$\lambda = 0.05$	$p_1 = 0.8$	$p_2 = 0.2$	$\lambda_1 = 0.0375$	$\lambda_2 = 0.1$

Table 2: Simulation parameters of the control strategies for notification delivery.

Adaptive strategy		Constant check interval	Random selection		Multiplicative decrease
$\alpha = 0.3$	$\delta = t_0 = 20$	$t_i = t_0 = 20$	$a = 10$	$b = 30$	factor 0.5

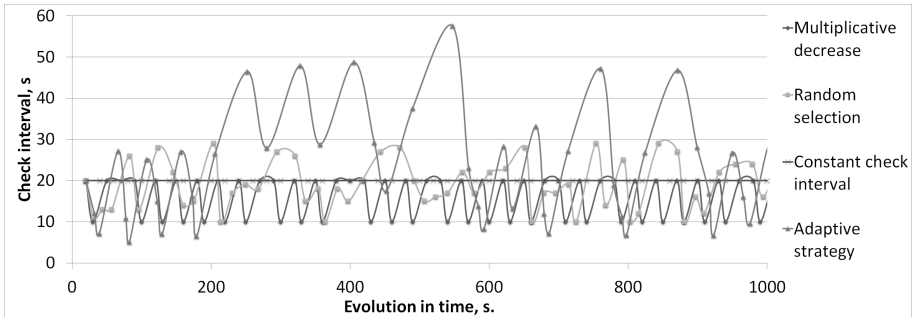


Fig. 2: Notification delivery control in case of Poissonian losses.

The basic efficiency metrics are the average number of losses k_{avg} and the average length of the check interval t_{avg}:

$$k_{\text{avg}} = \frac{1}{n}\sum_{i=1}^{n} k_i, \quad t_{\text{avg}} = \frac{1}{n}\sum_{i=1}^{n} t_i.$$

Intuitively, the client is interested in reducing the losses (i.e., $k_{\text{avg}} \to \min$) for high values of the check interval (i.e., $t_{\text{avg}} \to \max$). Since the two requirements are in opposition, in analogy with the concept of power in computer networks defined as the ratio between the throughput and the delay [8], we use the ratio among the two, such as the following metrics:

$$Q_{\text{tot}} = \frac{k_{\text{avg}}}{t_{\text{avg}}} \to \min \quad \text{or} \quad Q_{\text{avg}} = \frac{1}{n}\sum_{i=1}^{n} \frac{k_i}{t_i} \to \min\,.$$

An example of the performance comparison based on these metrics is shown in Table 3. The early experiments indicate that the control by the multiplicative decrease strategy and adaptation strategy achieves the lowest loss level. The efficiency of multiplicative decrease is due to frequent checks, which consumes resources of the client and the network. On the other hand, the adaptive strategy tries to increase the check interval length (on the cost of higher losses).

Table 3: Performance comparison of the experimented strategies

Metric	Multiplicative decrease	Random selection	Constant check interval	Adaptive strategy
k_{avg} / t_{avg}	0.66 / 15.80	1.04 / 20.79	1.04 / 20.00	1.00 / 24.22
Q_{tot}	0.042	0.050	0,052	0.041
Q_{avg}	0.033	0.051	0.052	0.042

5. Conclusion

This paper has studied the problem of efficient notification delivery for the subscription operation in smart spaces. We considered different assumptions on the notification loss distribution in wireless networked environments. We introduced several control strategies, including a proposal for a TCP-like control for adaptation of the check interval to the observable loss rate. A client can implement such a strategy individually to increase the delivery efficiency. The presented strategies are provided with initial analytical estimates and simulation experiments for performance comparison.

Acknowledgments. The study is part of research project # 14-07-00252 of the Russian Foundation for Basic Research. The work of D. Korzun is financially supported by the Ministry of Education and Science of Russia within project # 1481 of the basic part of state research assignment for 2014–2016.

REFERENCES

1. Allman, M., Paxson, V., Blanton, E.: TCP Congestion Control. RFC 5681 (Draft Standard) (Sep 2009), `http://www.ietf.org/rfc/rfc5681.txt`
2. Callegari, C., Giordano, S., Pagano, M., Pepe, T.: A survey of congestion control mechanisms in Linux TCP. In: Vishnevsky, V., Kozyrev, D., Larionov, A. (eds.) Distributed Computer and Communication Networks, Communications in Computer and Information Science, vol. 279, pp. 28–42. Springer International Publishing (2014)
3. Eugster, P.T., Felber, P.A., Guerraoui, R., Kermarrec, A.M.: The many faces of publish/subscribe. ACM Comput. Surv. 35, 114–131 (June 2003)
4. Honkola, J., Laine, H., Brown, R., Tyrkkö, O.: Smart-M3 information sharing platform. In: Proc. IEEE Symp. Computers and Communications (ISCC'10). pp. 1041–1046. IEEE Computer Society (Jun 2010)
5. Kleinrock, L.: Queueing Systems, vol. I: Theory. Wiley Interscience (1975)
6. Lomov, A.A., Korzun, D.G.: Subscription operation in Smart-M3. In: Balandin, S., Ovchinnikov, A. (eds.) Proc. 10th Conf. of Open Innovations Association FRUCT and 2nd Finnish–Russian Mobile Linux Summit. pp. 83–94. SUAI (Nov 2011)
7. Morandi, F., Roffia, L., D'Elia, A., Vergari, F., Cinotti, T.S.: RedSib: a Smart-M3 semantic information broker implementation. In: Balandin, S., Ovchinnikov, A. (eds.) Proc. 12th Conf. of Open Innovations Association FRUCT and Seminar on e-Tourism. pp. 86–98. SUAI (Nov 2012)
8. Peterson, L.L., Davie, B.S.: Computer Networks: A Systems Approach. The Morgan Kaufmann Series in Networking, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 5th edn. (2012)
9. Smirnov, A., Kashnevik, A., Shilov, N., Oliver, I., Balandin, S., Boldyrev, S.: Anonymous agent coordination in smart spaces: State-of-the-art. In: Smart Spaces and Next Generation Wired/Wireless Networking. Proc. 9th Int'l Conf. NEW2AN'09 and 2nd Conf. on Smart Spaces ruSMART 2009. LNCS 5764. pp. 42–51. Springer-Verlag, Berlin, Heidelberg (2009)

10. Vdovenko, A.S., Korzun, D.G.: Active control by a mobile client of subscription notifications in smart space. In: Proc. 16th Conf. Open Innovations Framework Program FRUCT. pp. 123–128. IEEE, ITMO University (Oct 2014)
11. Weiss, N., Holmes, P., Hardy, M.: A Course in Probability. Pearson Addison Wesley (2005)

ON HYPER-LOCAL WEB PAGES

D. Namiot[1]*, M. Sneps-Sneppe*[2]
[1] Lomonosov Moscow State University, Moscow, Russia
[2] Ventspils University College, Ventspils, Latvia
dnamiot@gmail.com, manfreds.sneps@gmail.com

Abstract

In this paper, we discuss one approach for development and deployment of web sites (web pages) devoted to the description of objects (events) with a precisely delineated geographic scope. This article describes the usage of context-aware programming models for web development. In our paper, we propose mechanisms to create mobile web applications which content links to some predefined geographic area. The accuracy of such a binding allows us to distinguish individual areas within the same indoor space. Target areas for such development are applications for Smart Cities and retail.

Keywords: Browsers, Computer networks, Context awareness, HTML5, Indoor communication

1. Introduction

Our paper deals with mobile web presentations of location-based services. How can we present some local (attached to a certain geographical location) information to mobile users? We are talking about programming (creating) mobile web sites, which content pages correspond to the current location of the mobile user. The traditional scheme is very straightforward. We have to determine the user's location and then create a dynamic web page, the issuance of which is clearly defined by specific geographical coordinates. For example, geo-location is a part of HTML5 standard [1].

As soon as web application obtained (as per user permission, of course) geo-coordinates, it can build a dynamic web page, which content depends on the current location (content is associated with obtained location). Technically, we can render our dynamic page on the client side (right in the browser), when application requests data from server via some asynchronous calls (AJAX) [2], or right on our server (in some CGI-script). In both cases obtained location info is used as a parameter either to AJAX script or to CGI script. For some of the applications (classes of applications), we may use several location-related datasets (e.g., so-called geo-fence [3]), but the common principles are similar. It is so called Location Based Services (LBS) [4].

There are different methods for obtaining location information for mobile users [5]. Not all of them use GPS (GLONAS) positioning actually. Alternative approaches use Wi-Fi, Cell ID, collaborative location, etc. [6]. The above-mentioned geo-location in HTML5 has been a wrapper (interface) for location service. For the most of LBS, their top-level architecture is standard. LBS use obtained location info as a key for

any database (data store) with location-dependent data. Location info is actually no more than a key for linking physical space (location) and virtual (e.g., coupon for the store). Only a small number of services actually use the coordinates. The typical example is indoor location based services. The paradigm *Location first* requires a digital map for an indoor space. This map should be created prior to the deployment, and it should be supported in an actual state during service's life time. On the other hand, there is a direction, called context-aware computing. In context-aware computing (ubiquitous computing) services can use other information (not related to geographic coordinates) as the "characteristics" of a user's location. Simplistically, the context is any additional information on the geographical location [7-8]. In this case, additional information (context), with the presence of certain metrics can serve as a unique (up to a certain approximation, of course) feature of a user's location. Or, in other words, we can substitute geo-location with context identification. Why might it be necessary? The typical example is indoor LBS [9]. Traditional geo-positioning can be difficult and positioning accuracy may be insufficient to distinguish the position of the mobile subscriber within the same premises. And yet, it is the distinction between positions within the same space (buildings) may be important for all kinds of services (for example, the buyer is located on the first or second floor of the hall).

Actually, it is a starting point for new approaches in LBS architecture, when the stage with obtaining (detecting) location info could be completely eliminated. Indeed, if location info is no more than a key for some database, then why do not replace geo-keys (e.g., latitude and longitude) with context-related IDs? It is sufficient to identify context and use this identification to search data.

The rest of the paper is organized as follows. In section II, we describe context identification. In section III, we describe how this identification could be used in web programming. In section IV, we discuss the generic approaches for incorporating sensing information into web pages.

2. Network Proximity

One of the widely used methods for the identification of context is the use of wireless network interfaces of mobile devices (Wi-Fi, Bluetooth). The reasons for this are straightforward. On the first hand, these interfaces are supported in all modern smart-phones. Secondly, for obvious reasons, monitoring of network interfaces is directly supported and executed by the mobile operating systems. Therefore, a survey of network interfaces on the application level can be simplified and not cause additional power consumption, as compared with, for example, a specially organized monitoring for the accelerometer. Information received through the network interface is used to estimate the proximity of the mobile user to the elements of the network infrastructure (network proximity [10). Note, that other mobile devices can act as these elements too (e.g., Wi-Fi access point, opened right onto mobile phone [11]). The classical form for collecting data about Wi-Fi devices are so-called Wi-Fi fingerprints sets [12]. Wi-Fi fingerprints are digital objects that describe availability (visibility) for network nodes. Their primary usage is navigation related tasks. The alternative approach lets users directly associate some data chunks with existing (or artificially created) net-

work nodes. In other words, it is a set of user generated links between network nodes and some content that could be used by those in proximity to networks nodes. This approach is presented in SpotEx project and associated tools [13-14]. SpotEx lets users create a set of rules (logical productions) for linking network elements and available content. A special mobile application (context-aware browser) is based on the external set of rules (productions, if-then operators). The conditional part of the each rule includes predicates with the following objects:

- identity for Wi-Fi network (name, MAC-address)
- RSSI (signal strength),
- time of the day (optionally),

In other words, it is a set of operators like this:

IF AccessPointIsVisible ('Cafe') THEN { *show content for Cafe* }

Block { *show content for Cafe* } is some data (information) snippet presented in the rule. Each snippet has got a title (text) and some HTML content (it could be simply a link to any external site for example). Snippets could present coupons/discount info for malls, news data for campuses, etc. The context-aware browser (mobile application) maps current network environment against existing database, detects relevant rules (fires them) and builds a dynamic web page. This web page is presented to a mobile user in proximity. In fact, even the name of the application (context-aware browser) suggests the movement of this functionality in a mobile browser. This would eliminate the separate rule base as well as the special (separate) application. In fact, the standard mobile browser should play a role of this application. Rules for the content (data snippets) must be specified directly on the mobile web pages. And data snippets itself are HTML code chunks anyway. As applied implementations, we can mention, for example, Internet of Things applications [15-16]. The usage is very transparent. Data snippets (data, presented to mobile users) depends on visibility for some Wi-Fi access points. It lets us specify the positions for mobile users inside of some building (campus, etc.) Mobile users will see different information for different positions. And this approach does not use geo-coordinates at all. The next interesting direction is EU project FI-WARE [17]. Integration with the FI-CONTENT platform is one of the nearest goals.

3. Information Services

Technically, for the reuse of information about network proximity, we can talk about the two approaches. At the first hand, the implementation of a mobile browser can follow the same ideology that supports geo-coding in HTML5 [18]. A function from browser's interface

```
navigator.geolocation.getCurrentPosition()
```

accepts as a parameter some user-defined callback (another function). The callback should be called as soon as geo-location is completed. Obtained data should be passed as parameters. Note, that the whole process is asynchronous. By the analogue with the above-mentioned model, a mobile browser can add a new interface function. E.g., *getNetworks()* this function will accept a user-defined callback for accumulating network information (current fingerprint). A good candidate for data model is JSON. The browser will pass fingerprint as a JSON array to a user-defined callback. Each element from this array describes one network and contains the following information:

- SSID - name for access point
- MAC - MAC-address
- RSSI - signal strength

Note, that scanning networks is an asynchronous process in mobile OS. So, callback pattern is a good fit for this. Firefox OS is closest in ideology to this approach [19]. Also, Firefox OS offers Bluetooth API [20]. It has got the similar ideology, but there is no general unifier (e.g., even fields for objects are different). It should be possible, of course, to create some unified wrapper (shell), which will give a general list of networks. The biggest problem (we are not mentioning here the own prevalence and popularity for Firefox OS) is the status for both APIs. Wi-Fi API has just been scheduled yet. At the same time, the Bluetooth API exists, but it is declared preferred (privileged). Privileged APIs can be used by the operating system only. So, it could not be used in applications. The reason for this solution is security. API combines both network scanning and network connection (data exchange). It is the wrong design by our opinion. APIs functionality should be separated. The above-mentioned SpotEx approach is not about the connectivity. Mobile OS should use two separate APIs: one for scanning (networks poll) and one for connecting. Polling for networks does not require data exchange. So, scanning API is safe, and it should not be privileged. It is simple - we should have *WiFiManager* interface (as is, and it could be privileged), and *WiFiScan* with only one function *getNetworks()*:

```
<script>
function callback_function(json_data ) { ... }
WiFiScan.getNetworks(callback_function);
</script>
```

The callback function can loop over an array of existing networks IDs and show (hide) HTML div blocks with data related (associated) to the existing (visible) networks. Actually, it is a fundamental question. Traditionally, wireless networks on mobile phones are used as networks. But they are sensors too. The fact that some network node is reachable (visible) is a separate issue. And it could be used in mobile applications even without the ability to connect to that node. It is the main idea behind SpotEx, and it is the feature (option) we suggest to embed into mobile browsers. How can we present our rules for network proximity? As per our suggestion, each

data snipped should be presented as a separate div block in HTML code. E.g., the above-mentioned example looks so:

show content for Cafe

We can use CSS styles to hide/show this block. And this CSS visibility attribute depends on the visibility of Wi-Fi (Bluetooth) nodes. Of course, CSS visibility could be changes in JavaScript. So, our rules could be implemented in JavaScript code. We can directly present the predicates in our code, or describe their parts in CSS too. HTML5 custom attributes are good candidates for new attributes [21]. It means also, that adding some set of rules to existing web page looks like as adding (including) some JavaScript code (JavaScript file).

In general, this approach could change the paradigm of designing mobile web sites. It eliminates the demand to make separate versions for local sites or events. It is enough to have one common site with local offers (events, etc.) placed in hidden blocks. Blocks will be visible to mobile users in a proximity of some network nodes. Local blocks visibility depends on the network nodes visibility and so, it depends on the current location of mobile users. E.g., for the above-mentioned example, mobile users opened Cafe site being physically present in the proximity of Cafe, will see different (additional) data compared with any regular mobile visitor.

Of course, single data source (just one web site) support simplifies (makes it cheaper) the maintenance during life time. Web Intents [22] present the next interesting model for this approach. The Web Intents formation is a client framework (everything is executed in the browser) for the monitoring (polling) and building services interaction within the application. Interactions include data exchange and transfer of control. Web Intents form the core architecture of Android OS [23], but their future status is still unknown after some initial experiments from Google. We should note in this context a similar (by its concept) initiative from Mozilla Labs - Web Activities [24]. But the further status of this initiative is also unclear.

The next possible toolbox is seriously underrated in our opinion. It is a local web server. The first implementation, as far as we know, refers to the Nokia [25]. In our opinion, this is one of the most promising areas for communicating with phone sensors. The next possible idea resembles in some ways the old projects with WAP (Wireless Access Protocol). In this case, a mobile device used some intermediate server (WAP Gateway) for access to internet resources. This intermediate server should be able to collect sensing information (including network sensors). Internet service will get sensing info from our proxy.

4. A generic approach for web sensing

In this part, we would like to discuss the more generic approach (approaches) for embedding sensing information into web pages. As a workaround and prototype for

this development, we can present a custom *WebView* for Android. On Android platform is possible to access from JavaScript to Java code for a web page, loaded into *WebView* control. Java code will provide a list of nearby network nodes (calculate the network fingerprint). The key moment here is the need for an asynchronous call from JavaScript, because scanning for wireless networks in Java is the asynchronous process. Let us describe this approach a bit more detailed. On Android side we activate JavaScript interface:

```
public void onCreate(Bundle savedInstanceState) {
super.onCreate(savedInstanceState);
WebView webView = new WebView(this);
setContentView(webView);
WebSettings settings = webView.getSettings();
settings.setJavaScriptEnabled(true);
webView.addJavascriptInterface(new MyJavascriptInterface(), ”Network”); }
```

Now we can describe our Java code for getting network fingerprint. As a parameter, we will pass a name for callback function in JavaScript.

```
@JavascriptInterface
public void getNetworks(final String callbackFunction) { }
```

We skip the code for network scanning and demonstrate the final part only. As soon as a fingerprint in obtained, we can present it as JSON array and invoke our callback:

```
webView.loadUrl(”javascript:” + callbackFunction + ”(’” + data + ”’)”);
```

And on our web page, we can describe our callback function and call Java code:

```
function f_callback(json) { }
Network.getNetworks(”f_callback”);
```

This approach lets us proceed network proximity right in JavaScript (in other words, right on the web page). Actually, by the similar manner we can work with other sensors too. It is so-called Data Program Interface [26]. We would like to see something similar as a standard feature in the upcoming versions of Android.

5. Conclusion

The paper discusses the use of information about the network environment to create dynamic web pages. We propose several approaches to the implementation of a mobile browser that can handle data on a network (network proximity) to provide users with information tied to the current context. Also, we considered possible im-

plementation details. The basic idea is to separate the functional for scanning network information and real data exchange.

REFERENCES

1. A.T.Holdener, "HTML5 Geolocation", pp. 31-35, O'Reilly Media, Inc., 2011.
2. D.Namiot and M. Sneps-Sneppe, "Where Are They Now-Safe Location Sharing," Internet of Things, Smart Spaces, and Next Generation Networking, Springer Berlin Heidelberg, pp. 63-74, 2012.
3. D.Namiot and M.Sneps-Sneppe, "Geofence and Network Proximity," Internet of Things, Smart Spaces, and Next Generation Networking, Springer Berlin Heidelberg, pp. 117-127, 2013.
4. M.Prasad, "Location based services," GIS Development, pp.3-35, 2002.
5. S.Tabbane, "Location management methods for third generation mobile systems," Communications Magazine, IEEE, vol. 35, no. 8, pp. 72-78, 1997. [Online] Available: http://dx.doi.org/ 10.1109/35.606034
6. D. Namiot, "Context-Aware Browsing – A Practical Approach", Next Generation Mobile Applications, Services and Technologies (NGMAST), 2012 6th International Conference on, pp. 18-23. [Online] Available: http://dx.doi.org/10.1109/NGMAST.2012.13
7. G. Schilit and B. Theimer, "Disseminating Active Map Information to Mobile Hosts," IEEE Network, vol. 8, no. 5, pp. 22-32, 1994. [Online] Available: http://dx.doi.org/10.1109/65.313011
8. D. Namiot and M. Sneps-Sneppe, "Context-aware data discovery", Intelligence in Next Generation Networks (ICIN), 2012 16th International Conference on, pp. 134-141, 2012. [Online] Available: http://dx.doi.org/ 10.1109/ICIN.2012.6376016.
9. K.Kolodziej and J.Danado, "In-building positioning: modeling location for indoor world," Database and Expert Systems Applications, 2004. Proceedings. 15th International Workshop on (pp. 830-834). IEEE. [Online] Available: http://dx.doi.org/ 10.1109/DEXA.2004.1333579
10. P.Sharma, Z.Xu, S.Banerjee, and S.Lee, "Estimating network proximity and latency," ACM SIGCOMM Computer Communication Review, vol.. 36, no. 3, pp. 39-50, 2006. [Online] Available: http://dx.doi.org/ 10.1145/1140086.1140092
11. D.Namiot, "Network Proximity on Practice: Context-aware Applications and Wi-Fi Proximity," International Journal of Open Information Technologies, vol. 1, no. 3, pp. 1-4, 2013.
12. Y.C.Cheng, Y.Chawathe, A.LaMarca, and J.Krumm, "Accuracy characterization for metropolitan-scale Wi-Fi localization," In Proceedings of the 3rd international conference on Mobile systems, applications, and services, ACM, pp. 233-245, 2005. [Online] Available: http://dx.doi.org/ 10.1145/1067170.1067195
13. D.Namiot and M.Schneps-Schneppe, "About Location-aware Mobile Messages: Expert System Based on WiFi Spots," In Next Generation Mobile Applications, Services and Technologies (NGMAST), 2011 5th International Conference on, IEEE, pp. 48-53, 2011. [Online] Available: http://dx.doi.org/10.1109/NGMAST.2011.19

14. D. Namiot and M. Sneps-Sneppe. ”Wi-Fi proximity as a Service,” In SMART 2012, The First International Conference on Smart Systems, Devices and Technologies, pp. 62-68, 2012.
15. M.Schneps-Schneppe and D.Namiot, ”Open API for M2M Applications: What is Next?” In AICT 2012, The Eighth Advanced International Conference on Telecommunications, pp. 18-23, 2012.
16. D.Namiot and M.Schneps-Schneppe, ”Smart cities software from the developer’s point of view,” 6 th International Conference on Applied Information and Communication Technologies (AICT2013),LUA Jelgava, Latvia, pp. 230-237, 2013.
17. M.Castrucci, M.Cecchi, F.D.Priscoli, L.Fogliati, P.Garino, and V.Suraci, ”Key concepts for the Future Internet architecture,” In Future Network & Mobile Summit (FutureNetw), IEEE, pp.1-10, 2011. [Online] Available: http://dx.doi.org/10.1145/1067170.1067195
18. S.Aghaee and P.Cesare, ”Mashup development with HTML5,” Proceedings of the 3rd and 4th International Workshop on Web APIs and Services Mashups. ACM, p. 10, 2010. [Online] Available: http://dx.doi.org/10.1145/1944999.1945009
19. S.Amatya and A.Kurti, ”Cross-Platform Mobile Development: Challenges and Opportunities,” In ICT Innovations 2013, Springer International Publishing, pp. 219-229, 2013.
20. A.Paul and S.Steglich, ”Virtualizing Devices. In Evolution of Telecommunication Services,” Springer Berlin Heidelberg, pp. 182-202, 2013.
21. F.Cazenave, V.Quint, and C.Roisin, ”Timesheets. js: When SMIL meets HTML5 and CSS3,” In Proceedings of the 11th ACM symposium on Document engineering, ACM, pp. 43-52, 2011. [Online] Available: http://dx.doi.org/10.1145/2034691.2034700
22. C.Zheng, W.Shen, and H.H.Ghenniwa, ”An intents-based approach for service discovery and integration,” In Computer Supported Cooperative Work in Design (CSCWD), 2013 IEEE 17th International Conference on, pp. 207-212, 2013 [Online] Available: http://dx.doi.org/10.1109/CSCWD.2013.6580964
23. E.Chin, A.P.Felt, K.Greenwood, and D.Wagner, ”Analyzing inter-application communication in Android,” In Proceedings of the 9th international conference on Mobile systems, applications, and services, ACM, pp. 239-252, 2011 [Online] Available: http://dx.doi.org/10.1145/1999995.2000018
24. M.Firtman, ”Programming the mobile web”, pp. 619-624, O’Reilly Media, Inc., 2013.
25. L.Oliveira, N.R.Antonio, and C.Jose, ”The Mobile Context Framework: providing context to mobile applications.” Distributed, Ambient, and Pervasive Interactions. Springer Berlin Heidelberg, pp. 144-153, 2013.
26. Namiot, D., Sneps-Sneppe, M. (2014, June). On software standards for smart cities: API or DPI. In ITU Kaleidoscope Academic Conference: Living in a converged world-Impossible without standards?, Proceedings of the 2014 (pp. 169-174). IEEE.

OPTIMIZATION OF TOPOLOGICAL STRUCTURE OF BROADBAND WIRELESS NETWORKS ALONG THE LONG TRAFFIC ROUTES[1]

V.M. Vishnevsky[1], *A.A. Larionov*[2], *R.V. Smolnikov*[3]

[1,2,3]V. A. Trapeznikov Institute of Control Sciences of Russian Academy of Sciences, Moscow, Russia

[1]vishn@inbox.ru, [2]Larioandr@gmail.com, [3]rodion.smolnikov@gmail.com

Abstract

The paper deals with a relevant problem of optimal placement of base stations in broadband wireless networks along the long highways. This problem is formulated in terms of integer linear programming. The exact solution of the problem as well as the high-speed heuristic algorithm are presented.

Keywords: topology structure synthesis, broadband wireless networks, linear and nonlinear programming, heuristic algorithm, network protocols

1. Introduction

Broadband wireless networks and communication channels have become recently one of the main directions of development of the telecommunications industry [1, 2]. The development of networks of this class for creation of a modern infrastructure for multimedia data transmission along the long transport routes is one of the major problems of implementation of new transport routes and pipelines and exploiting the existing ones.

Creation of such communication infrastructure allows to: provide the operating control over the technical parameters of a route by the means of high-speed data transfer from sensors and data units to the Control Center; provide the security control over the route sections and strategically important objects using data from the video surveillance systems; provide the voice communication (IP-telephony) and transmission of multimedia information between the stationary and mobile objects on long highways as well as communication with the Control Center, etc.

Due to the high practical demands for high-performance communication networks along the long transport routes based on IEEE 802.11-2012 [3] hardware and software, in recent years a considerable number of studies on this class of networks have appeared [4–11]. Particularly, in [4] the problem of wireless network base stations

[1]This research was financially supported by the Ministry of Education and Science of the Russian Federation in the framework of the applied research project №14.613.21.0020 of 22.10.2014 (RFMEFI61314X0020).

deployment, maximizing the network coverage is investigated under constraints on the total network cost. The initial data for solution of the problem are the potential locations for deployment of the base stations, and the beforehand-collected statistics of traffic from fixed and mobile users. The close problem of the deployment of base stations, maximizing the coverage, is formulated in [5]. For the analytical description of the problem, it is modeled as a Maximum Coverage with Time Threshold Problem (MCTTP), and a genetic algorithm is used to solve it. Strategies for roadside units (RSUs) placement based on the road traffic characteristics, aiming at improving connectivity in vehicular ad hoc networks, are presented in [6]. To divide the coverage area of each RSU, the authors propose an Expansion and Coloration Algorithm (ECA). The average connectivity model for all vehicles in the network is established based on the results obtained from ECA. The RSUs placement problem is formulated as a combinatorial optimization problem, of which the objective is to maximize the average connectivity probability by searching for an optimal position combination of the given RSUs. Taking part of an actual urban road network as an example, the RSUs placement problem is calculated and the optimal placement scheme is evaluated. Simulation results show that the optimal placement scheme obtained from the proposed strategy leads to the best connectivity compared to uniform placement and hot-spot placement.

The problem of roadside units placement in IEEE 802.11p / WAVE (wireless access in vehicular environment) networks is studied in [7]. An analytical model is presented that allows to analyze the delay of data transmission in communication networks along highways. The cases of bound and unbound base stations are investigated. It is shown that only those deployment strategies are efficient in which roadside units are connected to each other within the line-of- sight range. Paper [8] designs a connectivity analysis scheme for the roadside-to- vehicle telematics network based on the real movement history of vehicle objects collected from taxi telematics system currently in operation, aiming at providing a useful guideline and information to build a telematics network. The implemented analyzer can locate the current and previous positions of all vehicles, decide whether it can be connected to an RSU, and calculate the duration of disconnected state, taking into account the transmission range, the number of RSUs, and RSU deployment. The RSU placement scheme can improve the network coverage exploiting the analysis result.

The results of simulation and measurement of the performance of a roadside unit placement scheme for the vehicular telematics network on the road network of Jeju city (South Korea) are presented in [9]. The calculated optimal topology of the backbone wireless network provides improvement of connectivity and reduction of the disconnection interval for the given number of roadside units, the transmission range, and the overlap ratio. A research problem of finding the optimal locations to place dissemination points (i.e. roadside infrastructure nodes for information dissemination) is considered in [10]. In this paper a novel approach is proposed for dissemination points placement in grid road networks without knowing trajectories. Based on the analysis of path number between two intersections, a probabilistic model is proposed to get the trajectories estimation of vehicles. The problem of the roadside unit placement

in vehicular networks is studied in [11], where the authors focus on the highway-like scenario in which there may be multiple lanes with exits or intersections along the road. In the proposed model, each vehicle can access RSUs in two ways: 1) direct delivery, which occurs when the vehicle is in the transmission range of the RSUs, and 2) multi-hop relaying, which takes place when the vehicle is out of RSU transmission range. Both access patterns in this placement strategy are worked out and this placement problem is formulated via an integer linear programming model such that the aggregate throughput in the network can be maximized. The impact of wireless interference, vehicle population distribution, and vehicle speeds are also taken into account in the formulation. The performance of the proposed placement strategy is evaluated via NS-2 simulations to generate vehicle mobility patterns.

In this article is considered RSU optimal placement problem (discrete Case) with number of practical limitations. This limitations was founded while development broadband wireless network for Ring Road of Kazan City (M7 "Volga").

2. Formulation of the problem

The problem is to choose placements of Road Side Units (RSU), where every RSU has its own connection coverage radius, radio transmission range and cost. Besides this set of RSU must has total cost less than a budget and the arrangement must cover as big area as possible. Consider a one-dimensional road, which is a straight line $\alpha[A, B]$ of length a. Consider set of points $X \subset \alpha$, and these points are possible placements for considered RSU. Consider C - a budget to be spent for net creation. T=$\{t_1 \dots t_n\}$ - set of station types, where every type is characterized by three parameters:

- r - coverage radius. It is a half of length of such segment *coverage segment*, containing a RSU in the middle of one, as every client inside the segment can communicate to the RSU and every other one can't.
- R - transmission range. It is a half of length of such segment *connection segment*, containing a RSU in the middle of one, as every another RSU inside the segment can establish a connection with the RSU and every other one can't.
- c - station cost.

Let give some definitions. Let the section α has orientation and $x_A < x_B$. Designate r(t), R(t), c(t) as coverage radius, transmission range and station cost of station type t respectively.

Definition 1. *Internal arrangement* or just arrangement of RSU is called ending set of pairs P=$\{(x_1, t_1), (x_2, t_2) \dots (x_N, t_N)\}$, $x_1 < x_2 < \dots < x_N$, where $x_i \in X$ - possible placement, $t_i \in T$ - station type.

Definition 2. *Correct arrangement* of RSU is called an arrangement $P = \{(x_1, t_1), (x_2, t_2) \dots (x_N, t_N)\}$, which:

$$\forall i = \overline{1, N-1} : x_{i+1} - x_i \leq \min\{R(t_i), R_(t_{i+1})\}, \tag{1}$$

Ie every neighbour stations can establish a connection with each other.

Designate $x(p)$ $t(p)$ as coordinate and station type respectively in pair p.

Definition 3. *Extended arrangement* of RSU is called ending set of such pairs $\hat{P} = \{(x_A, t_0), (x_1, t_1) \dots (x_N, t_N)(x_B, t_0)\}$, where $\{(x_1, t_1), (x_2, t_2) \dots (x_N, t_N)\}$ is an internal arrangement, t_0 - is special station type, which:

- $c(t_0) = 0$, RSU of this type has zero cost
- $r(t_0) = 0$, RSU of this type doesn't serve clients
- $R(t_0) = |\alpha|$, Ie if RSU in x_A and in x_B of type t_0 are situated inside transmission ranges of RSU in x_1 and x_n respectively than connections is established. (Ends of α achievable by internal arrangement)

Type t_0 - is an device for connection with gateways.

Obviously, an extended arrangement is correct if and only if its internal arrangement is correct.

Definition 4. *Coverage* $\Delta(P)$ is called set of points, which consists of every points of α being situated inside coverage radius at least one of placed RSU:

$$\Delta(P) = \{x \in \alpha, \quad \exists p \in P : |x(p) - x| \leq r(t(p))\}, \tag{2}$$

Obviously, $\Delta(P)$ is a set of segments.

Definition 5. Let P - an arrangement. Coverage area $\mathfrak{M}(\Delta(P))$ is called summary length of set of segments defined in .

Definition 6. *Coverage percentage* is called ratio:

$$\delta(P) = \frac{\mathfrak{M}(\Delta(P))}{\mathfrak{M}(\alpha)}, \tag{3}$$

Considering all definitions above, the general formulation of the problem is:

Problem 1. Find such extended correct arrangement $\hat{P}_0$ as:

$$\mathfrak{M}(\Delta(\hat{P}_0)) = \max(\mathfrak{M}(\Delta(\hat{P}))), \tag{4}$$

with limitations

$$\begin{cases} \hat{P} \text{ - an extended correct arrangement} \\ \sum\limits_{i=1}^{N} c(t(p_i)) \leq C, \text{ - total system cost is less than budget C} \end{cases}$$

Other words, telecommunications coverage area along segment of highway α is needed to be maximized considering limitations of total cost and connection availability as described in 2. That problem is too difficult to solve, so *discrete case* will be considered as being of the most practical interest. The discrete case has additional limitations for set X:

Let $X = \{x_1 \dots x_m\}$, $\forall i = \overline{0 \dots m+1} : x_i \in \alpha$, where m - ending number and every $x_i \in \alpha$, so X is a discrete set of points.

The formulation with last limitation may be used for modeling of real problem, for example, when there are number of power transmission towers along highway.

3. Formulation discrete problem in terms of linear programming

Input parameters:

X=$\{x_0, x_1 \dots x_m, x_{m+1}\}$, - set of possible placement points and $x_0 = A$, $x_{m+1} = B$.

T=$\{t_0 \dots t_n\}$, - set of available station types and t_0 is the special type, described in definition 3.

C - budget size for purchase RSUs.

Boolean variables are introduced:

$$\mathbf{e_i} = \begin{cases} 1, & \text{If in point } i \text{ is situated an RSU} \\ 0, & \text{If point } i \text{ is empty} \end{cases}, \forall i = \overline{0 \dots m+1}, \tag{5}$$

$$\mathbf{v_{i,j}} = \begin{cases} 1, & \text{If in point } i \text{ is situated an RSU of type } j \\ 0, & \text{If point } i \text{ is empty} \\ & \text{or type installed RSU different.} \end{cases}, \forall i = \overline{0 \dots m+1}, \tag{6}$$

$$\mathbf{u_{i,j}} = \begin{cases} 1, & \text{If stations are situated in points} \\ & i \text{ and } j \text{ respectively and} \\ & \text{establish a connection between each other.} \\ 0, & \text{in other case.} \\ & \text{In addtion, the station is considered not} \\ & \text{to connect itself. } u_{i,j} = 0 \text{ if } i = j. \end{cases}, \begin{cases} \forall i = \overline{0 \dots m+1}, \\ \forall j = \overline{0 \dots m+1}, \end{cases} \tag{7}$$

$$\mathbf{z_{i,j,k}} = \begin{cases} 1, & \text{If a station of type } j \text{ are situated in point } i \\ & \text{and establish a connection with} \\ & \text{station in point } k \\ 0, & \text{In other case} \end{cases}, \begin{cases} \forall i = \overline{0 \dots m+1}, \\ \forall j = \overline{1 \dots n}, \\ \forall k = \overline{0 \dots m+1}, \end{cases} \tag{8}$$

Real variables are introduced:

ρ_i^+ – segment, *left* end of one is in point i

ρ_i^- – segment, *left* end of one is in point i

This sets of segments are real coverage segment of an RSU:
ρ_i^+ and ρ_i^- are covered only by RSU in point i.
if $e_i = 0$ (i.e. point i is empty) than $\rho_i^{+-} = 0$.

θ_i – cost of station being situated in point. i $e_i = 0 \Rightarrow \theta_i = 0$.

The objective function:

$$\Phi = \sum_{i=0}^{m+1} (\rho_i^+ + \rho_i^-), \tag{9}$$

Maximum of Φ is sought. It ensures the widest area of highway coverage.

Limitations of correctness of arrangement:

Condition 1. There isn't more than one RSU in a point:

$$\forall i = \overline{0 \dots m+1} : \sum_{j=1}^{n} v_{i,j} = y_i, \tag{10}$$

Condition 2. Points A and B has RSUs of special type t_0

$$\begin{aligned} v_{0,0} &= 1, \\ v_{m+1,0} &= 1, \end{aligned} \tag{11}$$

This limitation has come from 3.

Condition 3. There aren't another points except A and B where is installed RSU of type t_0:

$$\forall i = \overline{1 \dots m}, v_{i,0} = 0, \tag{12}$$

Cost limitations:

Condition 4. Cost of RSU in point i:

$$\forall i = \overline{0 \dots m+1}, \sum_{j=1}^{n} c(t(j)) v_{i,j} = \tilde{c}_i, \tag{13}$$

Condition 5. Total cost of the system is less than budget size:

$$\sum_{i=0}^{m+1} \theta_i \le C, \tag{14}$$

As it follows from *4* the sum $\sum\limits_{i=0}^{m+1} \theta_i$ is exactly total cost of set all installed RSU.

Limitations of coverage:

Condition 6. Coverage segments ρ_i^+ ρ_i^- must be less than coverage radius of RSU in point i:

$$\begin{aligned} \forall i = \overline{0 \dots m+1}, \sum_{j=1}^{n} r(t(j)) v_{i,j} \ge \rho_i^+, \\ \forall i = \overline{0 \dots m+1}, \sum_{j=1}^{n} r(t(j)) v_{i,j} \ge \rho_i^-, \end{aligned} \tag{15}$$

Ie real coverage segment is everywhere equal ar less than declared by RSU type.

Condition 7. Every segments ρ_i^+ ρ_j^- don't have more than one mutual points:

$$\forall i = \overline{0 \dots m+1}, \forall j = \overline{0 \dots m+1}, x_i + \rho_i^+ \le x_j - \rho_j^-, \tag{16}$$

In terms of coverage area it means that the segments not intersect each other.

Limitations RSUs links :

Condition 8. Every station except RSU in point B has only one RSU with bigger coordinate value which establish a connection to:

$$\forall i = \overline{0 \dots m+1}, \sum_{j=i+1}^{k+1} u_{i,j} = y_i, \tag{17}$$

Condition 9. Every station except RSU in point A has only one RSU with less coordinate value which establish a connection to:

$$\forall i = \overline{0 \dots m+1}, \sum_{j=0}^{i-1} u_{i,j} = y_i, \tag{18}$$

Condition 10. Every connection is reciprocal:

$$\forall i = \overline{0 \dots m+1}, j = \overline{0 \dots m+1}, u_{i,j} = u_{j,i}, \tag{19}$$

That limitations means every RSU except of RSUs in points A and B has exactly two neighbours one - on the left and one on the right.

Condition 11. Relation between variables $z_{i,j,k}$ and $u_{i,k}$, $v_{i,j}$:

$$\forall i = \overline{0 \dots m+1}, \forall j = \overline{1 \dots n}, \forall k = \overline{0 \dots m+1}, z_{i,j,k} = u_{i,k} v_{i,j}, \tag{20}$$

This equation matches with $z_{i,j,k}$ description.

Condition 12. RSU of type j installed in point i establish a connection with RSU in point k only if the last is inside of transmission range of first:

$$\forall i = \overline{0 \dots m+1}, \forall j = \overline{1 \dots n}, \forall k = \overline{0 \dots m+1}, z_{i,j,k}(R(t_k) - |x_i - x_k|) \geq 0, \tag{21}$$

Limitation *11* is not linear. But according to [12] it may be replaced by two linear ones:

Lemma 1. *Let x_j and z - boolean variables. Than equality:*

$$\begin{cases} \sum\limits_{i=1}^{J} x_i - z \leq J - 1 \\ -\sum\limits_{i=1}^{J} x_i + J \cdot z \leq 0 \end{cases} \quad \Leftrightarrow \quad \prod_{i=1}^{J} x_i = z, \tag{22}$$

Using the lemma, *Condition 11* becomes identically equal to:

Condition 13.

$$\begin{cases} u_{i,k} + s_{i,j} - z_{i,j,k} \leq 1, \\ 2z_{i,j,k} - u_{i,k} - s_{i,j} \leq 0, \end{cases} \quad \begin{cases} \forall i = \overline{0 \dots m+1}, \\ \forall j = \overline{1 \dots n}, \\ \forall k = \overline{0 \dots m+1}, \end{cases} \tag{23}$$

Formulation of discrete problem of optimization of topological structure of broadband wireless networks along the long traffic routes in terms of linear programming:

Problem 2. Find set $\rho_i^+(0), \rho_i^-(0), i = \overline{0 \dots m+1}$ which:

$$\Phi(\rho_i^+(0), \rho_i^-(0)) = \max(\Phi(\rho_i^+, \rho_i^-)), \tag{24}$$

with number of limitations *1-13* (except *12*).

4. Solution of discrete problem in terms of linear programming.

To solve the problem of Integer linear programming 2 can be used a variety of methods including: Branch and bound method, implemented by, inter alia, in the form of the application package GLPK. For the exact solution of the problem of low dimension can be used restricted brute-force algorithm. The definition the problem of Low dimension refers to input parameters, in which the brute-force method works within a reasonable time. Branch and bound method and other popular methods for solving the mixed integer linear programming tasks are now implemented in such packages of applied programs as FortMP, Gurobi, Parma Polyhedra Library, GLPK, etc. This problem of Integer linear programming (discrete task) solved by using package GLPK (GNU Linear Programming Kit). GLPK prefers because it is a popular cross- platform system for solution of linear programming problems using the full power of advanced algorithms. The GLPK has the ability to solve problems in many different methods, such as simplex method, branch and bound method and many others. In this case- the case of the mixed problem of linear programming (Mixed Integer Problem)-branch and bound method is used. For each ongoing experiment GLPK-project consists of two text files: file with a description of the model (common for all experiments) and model data file generated by Java application from user input. Both files are written in the programming language MathProg (GNU MathProg Language)-main language for modeling of linear programming solvers (for example, glpsol).

Results of the problem solution by Branch and bound method are presented in table 1.Time characteristics expressed in seconds. Dashes means that too much time took calculation of problem with appropriate dimension (reasonable time is 10 min).

	Number of RSU types		
Number of possible placements	**T = 4**	**T = 5**	**T = 6**
X = 18	4,5	9,2	16,9
X = 28	86,4	614,2	-
X = 54	241,5	-	-

Table 1: Calculation of problem 2 by Branch and bound method

GLPK successfully copes with 2 of dimensions $0 < |P| < 50$, $0 < |T| < 5$. It needs to note, that parameter $|T|$ has more influence to calculation time than parameter $|P|$. Heuristic algorithm based on gradient descent is developed for solving bigger dimension problems.

5. Conclusion

Practically important and theoretically interesting problem has been formulated in this article. General results are:

The problem of optimization of topological structure of broadband wireless networks along the long traffic routes has been formulated. Exact solution has been got using GLPK packages. Heuristic algorithm for high dimension cases has been researched. Calculation utilities have been developed.

REFERENCES

1. V.M. Vishnevsky, S.L. Portnoi, I.V. Shakhnovich. WiMAX Encyclopaedia. Way to 4G. // Tekhnosfera, Moscow, 2010. - 470 pp.
2. V.M. Vishnevsky, O.V. Semenova. Polling Systems: Theory and Applications for Broadband Wireless Networks. - London: Academic Publishing, 2012. - 317 pp.
3. IEEE Std. - IEEE Standard for Information technology. Telecommunications and information exchange between Local and metropolitan area networks. Specific requirements Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications., 3. 2012. - March.
4. M.B. Brahim, W. Drira, F. Filali. Roadside units placement within city-scaled area in vehicular ad-hoc networks // 3rd International Conference on Connected Vehicles and Expo (ICCVE 2014). - Vienna, Austria: 3-7 Nov 2014.
5. E.S. Cavalcante, A.L. Aquino, G.L. Pappa, A.A. Loureiro. Roadside unit deployment for information dissemination in a VANET: an evolutionary approach // 14th Annual Conference Companion on Genetic and Evolutionary Computation (GECCO '12)/ (May 2012). NY, USA: pp. 27-34
6. Liu H., Ding S., Yang L., Yang T. A. Connectivity-based Strategy for Roadside Units Placement in Vehicular Ad Hoc Networks // 2014. - Vol. 7. - pp. 91
7. Reis A.B., Sargento S., Neves F., Tonguz O.K. Deploying Roadside Units in Sparse Vehicular Networks: What Really Works and What Does Not / 2014. - Vol. 63. - pp. 2794-2806
8. Lee J. Design of a Network Coverage Analyzer for Roadside-to-Vehicle Telematics Networks // Ninth ACIS International Conference on Software Engineering, Artificial Intelligence, Networking, and Parallel / Distributed Computing,. - 6-8 Aug 2008.
9. J.Lee, C.M. Kim. A Roadside Unit Placement Scheme for Vehicular Telematics Networks./ Springer Lecture Notes in Computer Science, 2010. - Vol. 6059. - pp.196-202
10. Xie B., Xia G., Chen Y., Xu M. Roadside Infrastructure Placement for Information Dissemination in Urban ITS Based on a Probabilistic Model / Springer Lecture Notes in Computer Science / 2013. - Vol. 8147. - pp. 322-331
11. T.-J. Wu, W. Liao, C.-J. Chang. A Cost-Effective Strategy for Road-Side Unit Placement in Vehicular Networks. / IEEE Transactions on Communications, 2012. - Vol. 60. - pp. 2295 - 2303
12. Vishnevsky V.M., Lyakhov A.I., Portnoy S.L., Shakhnovich I.V. Broadband wireless networks of information transfer. [Shirikopolosnye besprovodnye seti peredachi informacii]. // Technosfera, Moscow, 2005. - pp 435-436.

INTEGRAL EVALUATION OF PERFORMANCE AND QUALITY OF INFORMATION SERVICES IN A CLOUD INFRASTRUCTURE

V. Efimov[1], *S. Mescheryakov*[2], *D.Sc., D. Shchemelinin*[1], *Ph.D.*
[1] RingCentral Inc., San Mateo, CA, USA
[2] St. Petersburg Polytechnic University, St. Petersburg, Russia
2vadim@inbox.ru, serg-phd@mail.ru, dshchmel@gmail.com

Abstract

Analysis of the quality of information services, which are provided over the Internet cloud distributed infrastructure, is given. For integral evaluation of services, the key performance indicators of IT industry level are proposed. Real world examples of change management control at RingCentral Telecommunication Company are presented that allowed reaching the service availability level of 99.998%.

Keywords: cloud services, service availability, quality of service, key performance indicators.

ИНТЕГРАЛЬНАЯ ОЦЕНКА ПРОИЗВОДИТЕЛЬНОСТИ И КАЧЕСТВА ИНФОРМАЦИОННЫХ УСЛУГ В ОБЛАЧНОЙ ИНФРАСТРУКТУРЕ

В. Ефимов[1], *С. Мещеряков*[2], *д.т.н., Д. Щемелинин*[1], *к.т.н.*
[1] Компания RingCentral Inc., Сан Матео, США
[2] Санкт-Петербургский политехнический университет, Санкт-Петербург, Россия
2vadim@inbox.ru, serg-phd@mail.ru, dshchmel@gmail.com

Аннотация

В работе дан анализ качества информационных сервисов, предоставляемых в распределенной облачной инфраструктуре Интернет. Для интегральной количественной оценки сервисов предложено использовать ключевые характеристики производительности, общепринятые в сфере информационных технологий. Приведены примеры управления качеством информационных услуг в телекоммуникационной компании RingCentral, что позволило достичь уровня доступности сервисов 99,998%.

Ключевые слова: облачные сервисы, уровень доступности услуг, качество сервиса, ключевые показатели производительности.

1. Введение

На международной конференции ICUMT-2014 [1] авторами был представлен адаптивный подход к управлению облачными сервисами, а также использование мониторинговой системы с открытым кодом Zabbix [2] для оперативного контроля функционирования распределенной инфраструктуры в телекоммуникационной компании RingCentral (RC), предоставляющей до 150 облачных сервисов в США и Канаде [3]. Учитывая огромный парк из 10 тыс виртуальных машин (Virtual Machines, VMs), а также устойчивый годовой рост 40% и расширение глобального рынка информационных услуг RC в Западной Европе и Юго-Восточной Азии,

возникла объективная потребность в сравнительной оценке качества сервиса (Quality of Service, QoS) на основе общепринятых во всем мире ключевых показателей производительности (Key Performance Indicators, KPIs).

В данной работе описаны метрики KPI, которые вычисляются автоматически на базе модели интеграции различных источников данных, предложенной в работе [4], и реализованы в единой информационно-управляющей системе (ИУС) масштаба крупной компании RingCentral.

2. Процесс обнаружения и устранения аномалий в RC сервисах

Если раньше Интернет провайдеры обеспечивали пользователям каналы связи и передачу по ним потоков данных, то в настоящее время все больше телекоммуникационных компаний предоставляют доступ к внутренним ресурсам (Infrastructure as a Service, IaaS) и/или информационным сервисам (Software as a Service, SaaS), которые реализуются по облачным технологиям виртуализации.

В компании RC функционирование всей глобальной облачной инфраструктуры отслеживается специальным оперативным центром (Global Network Operating Center, GNOC), который работает 24 часа в сутки 7 дней в неделю и несет ответственность за бесперебойность облачных сервисов в соответствии с принятым соглашением об уровне обслуживания (Service Level Agreement, SLA) более 300 тыс корпоративных клиентов. Все данные в реальном времени и критические события на многочисленных удаленных серверах отслеживаются мониторинговой системой Zabbix в единой базе данных (БД) и автоматически отображаются вэб приложением на централизованном табло (Network Monitoring Console, NMC) с целью быстрого обнаружения и реагирования на возможные аномалии. Анализ оперативных и исторических данных в БД Zabbix позволяет оценить динамику, определить повторяющиеся аномалии, предсказать и заранее предотвратить многие проблемы не допуская полного отказа распределенных сервисов в обслуживании (Distributed Denial of Service, DDoS).

В таблице 1 приведены источники информации и методы отслеживания критических событий в RC сервисах. В случае обнаружения аномалии действия GNOC, как и в любой другой компании в сфере предоставления SaaS/IaaS, включают следующие главные этапы:

Источники и способы обнаружения аномалий	Средняя статистика RC
Автоматические оповещения в мониторинговой системе Zabbix	90%
Сообщения от внешних партнеров и поставщиков Интернет услуг	4%
Функциональные тесты, выполняемые вручную с заданной периодичностью	3%
Уведомления от службы поддержки корпоративных клиентов	2%
Прочие источники	1%

Таблица 1: Методы обнаружения аномалий в компании RingCentral.

1. Обнаружение и подтверждение аномалии посредством NMC либо одним из способов, указанных в таблице 1.

2. Уведомление других членов команды и протоколирование сообщений об ошибке в одной из систем документирования. В компании RC для этих целей используется популярная в мире информационная система JIRA [5], а также инструментальные средства собственной разработки с отдельной базой данных инцидентов Incident Management Portal (IMP).

3. Восстановление SaaS имеющимися автоматизированными средствами GNOC, если проблема хорошо известна и для нее предусмотрено оперативное решение, например перезагрузка VM или перенаправление рабочей нагрузки на запасной сетевой ресурс.

4. Когда SaaS не может быть восстановлен средствами GNOC, инцидент эскалируется инженерам службы технической поддержки. Если в процессе решения проблемы модифицируется программное обеспечение или инфраструктура, то все изменения отражаются в отдельной информационной системе собственной разработки Change Management Portal (CMP), связанной с системой планирования проектов Roadmap.

Для ряда хронических проблем, относящихся к внешним поставщикам и не зависящих от внутренних ресурсов RC, предусмотрены автоматические процедуры восстановления SaaS. В таких случаях участие GNOC на этапах 1 и 2 не требуется, и документ JIRA также создается автоматически. Примером такой реализации является автоматическое восстановление Java сервисов в результате общеизвестной проблемы утечки памяти при пиковой пользовательской нагрузке на приложения с целью предотвращения DDoS, подробно рассмотренные в работе [6].

3. Ключевые показатели эффективности облачных сервисов

При оценке эффективности процессов в IMP/CMP/JIRA/Zabbix/Roadmap с помощью KPI главная проблема заключается в разрозненности перечисленных источников данных, хранящихся в отдельных БД реляционного типа. Поскольку в БД мониторинговой системы Zabbix сконцентрирована вся информация об удаленных серверах, то было решено использовать встроенные средства Zabbix для интеграции всех БД и автоматизации вычислений KPI. С этой целью разработаны гетерогенные SQL запросы, которые направляются по заданному расписанию с прокси-сервера Zabbix на распределенные серверы БД, и результат их выполнения возвращается в Zabbix как отдельные значения метрик по каждой из систем IMP/CMP/JIRA/Roadmap.

В таблице 2 в качестве примера приведены некоторые метрики и их ежедневные значения для системы IMP. Интеграция данных в Zabbix в последствии позволяет создавать новые метрики KPI на основе агрегированных выражений любой степени сложности, которые уже вычисляются локально на стороне сервера Zabbix и сохраняются в его БД. Для интегральной оценки QoS в RC реализованы следующие метрики KPI, показанные на схеме рис. 1 [1]:

1. Время обнаружения аномалии (Time To Detect, TTD). Традиционно TTD определяется как разница между фактическим событием на удаленном сервере и временем его фиксации в мониторинговой системе [7]. В реальности существует незначительная задержка, вызванная объективной потребностью передачи самых последних данных с удаленного сервера через прокси-сервер в БД мониторинговой системы Zabbix. Поэтому в компании RC принято оценивать TTD как время реакции GNOC на критическое событие в NMC с максимальным SLA 2 минуты.

Наименование метрики IMP в Zabbix	Значение за последний день
% инцидентов, автоматически обнаруженных в Zabbix	84
% инцидентов, эскалированных в службу поддержки	0
% инцидентов по результатам часовых ручных тестов	11
% инцидентов, выявленных корпоративными клиентами	0
% инцидентов, обнаруженных другими источниками	5
Среднее время устранения инцидента, минут	6
Общее количество инцидентов	19
Максимальная продолжительность инцидентов, минут	40
Максимальное число затронутых клиентов	0
Максимальное число оборванных соединений	500

Таблица 2: Пример ежедневного отчета по инцидентам в Zabbix.

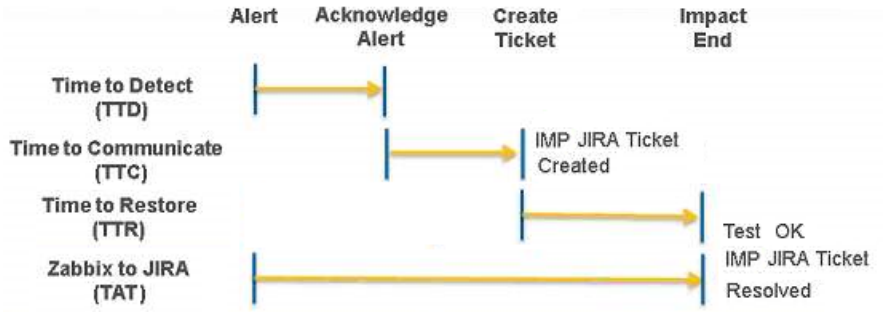


Рис. 1: Схема показателей интегральной оценки качества сервисов.

2. Время уведомления об аномалии (Time To Communicate, TTC). Под TTC понимается именно письменное уведомление заинтересованных лиц [8]. В компании RC для этих целей эффективно используются ИУС IMP и JIRA. Для документирования инцидента в GNOC установлен SLA 2 минуты, но на практике это занимает не более 1 минуты, поскольку многие процессы передачи информации в цепочке Zabbix- JIRA-IMP автоматизированы.

3. Время восстановления SaaS (Time To Restore/Repair, TTR). TTR – это базовая мера оценки периода недоступности SaaS для пользователей от момента времени, когда об этом стало известно, независимо от способа его восстановления и степени автоматизации [9]. Некоторые провайдеры ошибочно интерпретируют TTR как время реакции на инцидент (time to react), однако это некорректно, поскольку SaaS недоступен для пользователей до тех пор, пока процедура восстановления не закончена и QoS не протестировано.

На практике часто оперируют усредненными понятиями MTTD, MTTC, MTTR, которые являются частью SLA [10] и в компании RC вычисляются согласно схемы рис. 1 по формулам:

$$MTTD = \sum (T_2 - T_1) / N \quad (1)$$

$$MTTC = \sum (T_3 - T_2) / N \quad (2)$$

$$MTTR = \sum (T_4 - T_3) / N \quad (3)$$

где *MTTD* – среднее время обнаружения аномалии (Mean Time To Detect, MTTD); *MTTC* – среднее время уведомления об аномалии (Mean Time To Communicate, MTTC); *MTTR* – среднее время восстановления SaaS (Mean Time To Restore/Repair, MTTR); T_1 – время фактического события на удаленном сервере; T_2 – момент времени обнаружения аномалии в GNOC; T_3 – время создания документа JIRA/IMP; T_4 – время завершения процедуры восстановления SaaS и перевода инцидента JIRA/IMP в статус “Resolved/Closed”; N – количество инцидентов.

4. Время полного цикла изменения/восстановления SaaS (Turnaround Time, TAT). TAT – это категория мирового класса, используемая для общей оценки эффективности мероприятий по техническому обслуживанию и эксплуатации производственных систем независимо от отрасли экономики [11]. В IT сфере TAT означает общее время обработки заявки на сервисное обслуживание [12]. В компании RC согласно политики CMP в полный цикл TAT включен весь процесс создания, планирования, согласования изменений и их внедрение в работающую систему (рис. 2). Применительно к IMP эскалациям TAT определяется как длительность полного цикла обработки инцидента от момента его возникновения в работающей системе до окончательного восстановления SaaS для пользователей.

Таким образом, справедливо одно из следующих выражений в зависимости от процедуры восстановления SaaS:

$$TAT = MTTD + MTTC + MTTR \quad (4)$$

$$TAT = MTTD + MTTR \quad (5)$$

Выражение (5) применимо при наличии полностью автоматических процедур обнаружения аномалии и восстановления SaaS, когда MTTC и MTTR происходят параллельно и ручного вмешательства в процесс IMP эскалации не требуется.

4. Оценка эффективности управления изменениями в RC сервисах

Анализ статистики инцидентов в IMP показывает, что они во многих случаях являются следствием изменений инфраструктуры, поэтому их контроль в CMP с оценкой KPI имеет важное значение для достижения требуемого уровня SLA. Любое изменение всегда выполняется как часть одного из проектов, планируемых в Roadmap, с установленным рабочим циклом (рис. 2), который включает разработку SaaS (Development, DEV), аннонсирование новых функциональных возможностей (Release Notes Meeting, RNM), контроль развертывания SaaS (Deployment Tracking, DT) сначала на тестовой и затем на реальной системе (Production, PRO), управление внедрением (Release Management, RM), тестирование и контроль QoS.

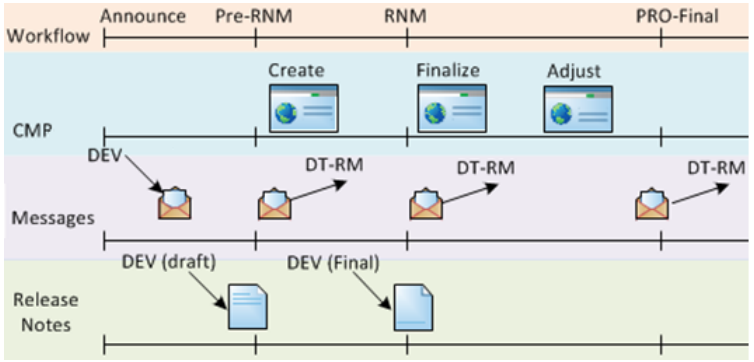


Рис. 2: Рабочий цикл выполнения заявок на изменения в RingCentral.

Заявка с детальным описанием изменений, включая плановую дату и время реализации, пошаговый план работ, целевые VM, план тестирования и отката в случае ошибки и т. п., подается на рассмотрение через CMP. В RC как крупной компании этот цикл согласования заявки в CMP составляет 8 дней, что позволяет провести техническую экспертизу, проверку плана реализации, анализ возможных последствий, подтверждение на уровне вовлеченных менеджеров отделов и совета директоров (Change Management Board, CMB). Статус заявки, CMB и весь процесс согласования отслеживается в CMP в реальном времени и фиксируется в БД CMP для последующего анализа и объективной оценки эффективности CMP.

С этой целью в мониторинговой системе Zabbix созданы метрики CMP KPI аналогично метрикам IMP, рассмотренным в разделе 3. Пример ежедневного отчета CMP KPI в Zabbix приведен в таблице 3, а на графике рис. 3 показаны исторические данные за месяц по одной метрике, вычисляющей длительность обслуживания заявок CMP. Данная метрика полезна для анализа загруженности ежедневной плановой профилактики. На графике видно, что 4-часовое окно используется в среднем менее чем на 50%, а выполнение одной заявки превысило максимальный предел 4 ч, но поскольку это произошло в выходной день в часы низкой пользовательской нагрузки, то это не считается инцидентом.

Наименование метрики CMP в Zabbix	Значение за последний день
% изменений, вызванных инцидентами	4
% изменений, выполненных в запланированное время	84
% срочных внеплановых изменений	12
Среднее время выполнения заявки, минут	67
Всего подано заявок	25
Всего отказано в обработке заявок	1
Всего обработано заявок	24

Таблица 3: Пример ежедневного отчета Zabbix по изменениям в RingCentral.

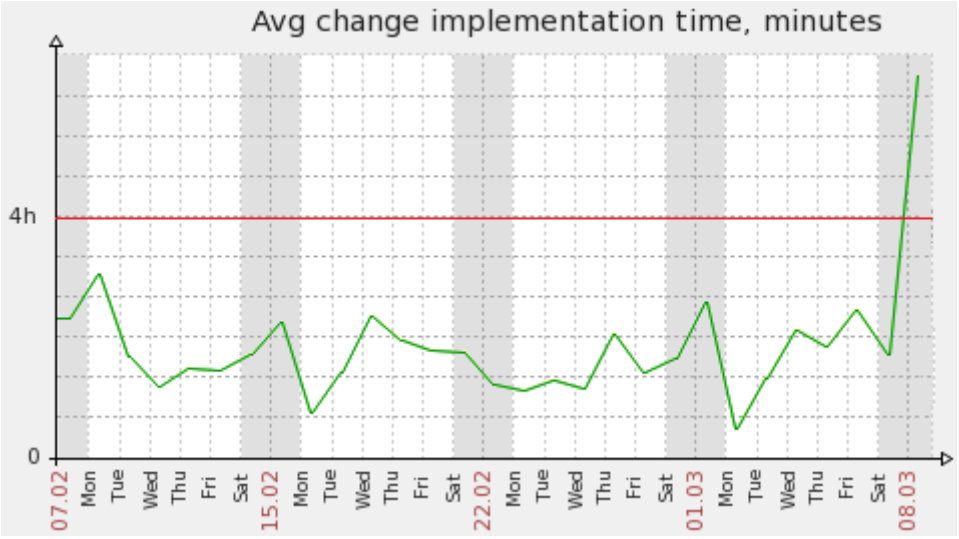


Рис. 3: Исторические данные Zabbix о времени выполнения заявок в RingCentral.

В работе [13] представлены другие примеры успешной реализации KPI с целью повышения эффективности IMP/CMP в условиях непрерывности процесса обновления SaaS/IaaS в динамично развивающейся многофункциональной облачной инфраструктуре RC.

5. Заключение

Главная проблема в повышении эффективности управления большой IT компанией, предоставляющей инфраструктурные и информационные сервисы в глобальной сети Интернет, состоит в разрозненности корпоративных информационных систем различного назначения. С целью интеграции и создания единого информационного пространства в компании RingCentral предложено использовать мониторинговую систему Zabbix масштаба предприятия, поскольку в ее централизованной базе данных в реальном режиме времени собираются данные о производительности, загруженности ресурсов, пользовательской активности и других системных событиях на многочисленных удаленных серверах и сетевых каналах связи в облачной инфраструктуре. Исторические данные БД Zabbix используются не только при решении текущих задач мониторинга и восстановления сервисов после инцидентов, но и для анализа тенденций дальнейшего развития, прогнозирования темпов масштабирования и потребности в системных ресурсах.

В данной работе преимущества автоматизированных средств мониторинга Zabbix использованы для интегральной оценки эффективности корпоративных систем управления на основе общепринятых в мире ключевых показателей производительности, таких как TTD, TTC, TTR, TAT, а также их усредненных значений. Внедрение KPI и анализ полученных результатов позволили улучшить качество и доступность бизнес сервисов компании RingCentral до уровня 99.998%, что является относительно высоким показателем в сфере телекоммуникаций.

ЛИТЕРАТУРА

1. Mescheryakov S., Shchemelinin D., Efimov V. Adaptive Control of Cloud Computing Resources in the Internet Telecommunication Multiservice System. The 6th International Congress on Ultra Modern Telecommunications and Control Systems, St. Petersburg, Russia, 2014, http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=7002117
2. Zabbix Enterprise-class Monitoring System, http://www.zabbix.com
3. RingCentral Cloud Services, http://www.ringcentral.com
4. Volkov A., Efimov V., Mescheryakov S., Shchemelinin D. Integrated Data Model for Managing a Multi-service Dynamic Infrastructure. Proceedings of the 3rd International Conference on Computer Modeling and Simulation, St. Petersburg, Polytechnic University Publishing House, 2014, http://dcn.icc.spbstu.ru/index.php?id=344&L=2
5. Atlassian JIRA, https://www.atlassian.com/software/jira
6. Mescheryakov S., Shchemelinin D. Capacity Management of Java-based Business Applications Running on Virtualized Environment. Proceedings of the 39th International Conference for the Performance and Capacity by CMG. La Jolla, CA, USA, 2013, http://www.cmg.org/conference/cmg2013/
7. Fault Detection and Isolation. Wikipedia, Free Encyclopedia, http://en.wikipedia.org/wiki/Fault_detection_and_isolation
8. Nonverbal Communication. Wikipedia, Free Encyclopedia, http://en.wikipedia.org/wiki/Nonverbal_communication
9. Mean Time to Repair. Wikipedia, Free Encyclopedia, http://en.wikipedia.org/wiki/Mean_time_to_repair
10. Service Level Agreement. Wikipedia, Free Encyclopedia, http://en.wikipedia.org/wiki/Service-level_agreement
11. Performance Engineering. Wikipedia, Free Encyclopedia, http://en.wikipedia.org/wiki/Performance_engineering
12. Turnaround Time. Performance Management. Wikipedia, Free Encyclopedia, http://en.wikipedia.org/wiki/Turnaround_time
13. Mescheryakov S., Shchemelinin D. Big Software Deployments in a Big Enterprise Environment. Proceedings of the 2nd ASE International Conference on Big Data Science and Computing, Stanford University, CA, USA, 2014, http://www.ase360.org/handle/123456789/135/

PERFORMANCE EVALUATION OF A TANDEM QUEUE WITH COMMON FOR PHASES SERVERS

T. Efimushkina[1,2]
[1] Peoples Friendship University of Russia, Moscow, Russia,
[2] Fraunhofer Heinrich Hertz Institute, Berlin, Germany

The fifth generation of mobile technology (5G) is positioned to enable a fully mobile and connected society by enhancing considerably the performance in terms of throughput, latency and reliability and controlling a highly heterogeneous environment characterized by the existence of multi-layer networks, multiple types of devices, etc. In such an environment, there is a fundamental need for enablers to achieve a consistent and seamless user experience, and therefore, the selection of algorithm that manages scarce resources in the network is of great importance. In this paper a tandem queue in discrete time is proposed and investigated with servers distributed among the phases. Various resources allocation schemes, including deterministic algorithms and proportional ones that adaptively distribute resources between the phases are compared by means of extensive simulation analysis.

ИССЛЕДОВАНИЕ МНОГОФАЗНОЙ СМО В ДИСКРЕТНОМ ВРЕМЕНИ С РАСПРЕДЕЛЯЕМЫМ МЕЖДУ ФАЗАМИ МНОЖЕСТВОМ ПРИБОРОВ

Т. Ефимушкина[1,2]
[1] Российский университет дружбы народов, Москва, Россия,
[2] Fraunhofer Heinrich Hertz Institute, Берлин, Германия,
tatiana.efimushkina@hhi.fraunhofer.de

Аннотация

Одной из задач технологий мобильной связи пятого поколения (5G) является создание полностью подключенного и мобильного общества за счет улучшения производительности с точки зрения пропускной способности, задержки, и обеспечения надежности, а также эффективного управления гетерогенными сетями связи, характеризующихся высокой плотностью пользователей и разнообразием пользовательских терминалов. При этом для обеспечения стабильного и приемлемого уровня обслуживания распределение ресурсов в гетерогенных сетях остается важным исследовательским вопросом. В статье предлагается и исследуется многофазная СМО с распределяемым между фазами общим множеством приборов. Разработана

имитационная модель, на базе которой проведен численный анализ различных схем распределения ресурсов на базе детерминированного и пропорциональных алгоритмов.

Ключевые слова: гетерогенная сеть, нисходящий канал, многофазная СМО, схема распределения приборов

1. Введение

Введем основные структурные параметры рассматриваемой СМО: пусть M - число фаз в многофазной системе, $1 \leq M \leq \infty$. Рассмотрим СМО с буферным накопителем (БН) фазы m (БН_m) емкости r_m, $0 \leq r_m \leq \infty$, $\mathbf{r}^T := (r_1, r_2, ..., r_M)$. Здесь и далее в статье, если не определяется особо, $m = \overline{1, M}$ и m соответствует номеру $m-$й фазы. На БН_m поступает неординарный поток однородных заявок. После обслуживания прибором фазы m заявка либо с заданной вероятностью покидает СМО, либо с дополнительной поступает на БН_{m+1} для обслуживания на фазе $m+1$, $m = \overline{1, M-1}$; при $m = M$ все обслуженные заявки покидают СМО. Структура СМО иллюстрируется рис.1. Представленные на нем, но не введенные выше параметры, определяются далее. Пусть далее точка вместо индекса означает полную сумму переменной по этому индексу.

Будем считать, что из совокупного числа c, $c < \infty$, приборов в многофазной СМО за фазой m постоянно закреплены c_m приборов, $\mathbf{c}^T := (c_1, c_2, ..., c_M)$, а оставшиеся $c - c_\bullet$ приборов распределяются между всеми фазами случайным образом в соответствии с заданной процедурой; $0 \leq c_m < c, 0 \leq c_\bullet \leq c$, при условии, что $c \neq 0$, если $c = c_\bullet$. Случай $c_m = 0$ означает отсутствие закрепленных за фазой m приборов; если $c_\bullet = 0$, то все множество приборов распределяется между фазами; при $c_\bullet = c$ приходим к варианту традиционной многофазной СМО с фиксированным числом c_m приборов на фазе m без распределяемых между фазами приборов.

Будем рассматривать функционирование СМО в дискретном времени с тактом постоянной длины $h, h > 0$. Разделим ось времени на такты и примем, что все изменения в системе происходят лишь в моменты $nh, n = 0, 1, ...$ Для определенности будем считать, что такт n есть полуинтервал $[nh, (n+1)h)$. Будем предполагать, что события в СМО в n-м такте (в момент nh) происходят в следующей последовательности:

- распределение $c - c_\bullet$ приборов между фазами;
- окончание обслуживания заявок на фазах, начиная с фазы M, затем на предыдущей фазе и т.д. до фазы 1; последовательно с обслуживанием освобождение мест, занимаемых обслуженными заявками в БН фазы; уход из СМО или поступление этих заявок на следующую фазу и буферизация на свободные места ее БН (только уход из СМО, если это фаза M);
- поступление новых заявок на фазы СМО;
- фиксация состояния СМО.

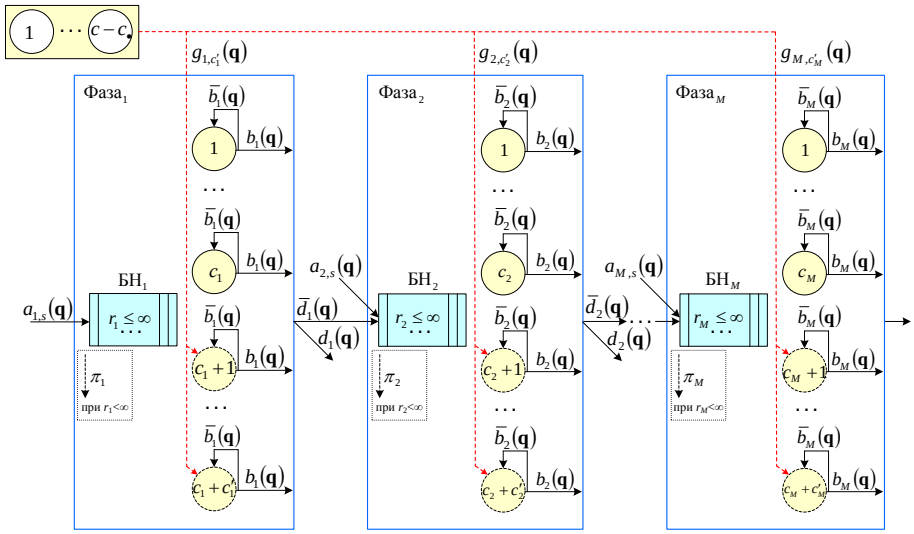


Рис. 1: Многофазная СМО в дискретном времени с распределяемым между фазами общим множеством приборов.

1.1. Описание процессов распределения приборов, обслуживания, буферизации и поступления. Далее будем рассматривать лишь неотрицательные целочисленные случайные величины (СВ). Пусть $\xi_{m,n}, \xi_{m,n} \in \{0,1,...,r_m\}$ - СВ числа заявок в БН$_m$ на такте n, СВ $\boldsymbol{\xi}_n^T := (\xi_{1,n}, \xi_{2,n}, ..., \xi_{M,n})$ будем называть состоянием СМО на такте $n, n \geq 0$. Рассмотрим подробнее в соответствии с введенной последовательностью событий в СМО процессы распределения приборов между фазами, обслуживания на фазах, буферизации и поступления заявок на систему.

Обозначим $\gamma_{m,n}, \gamma_{m,n} \in \{0,1,...,c-c_\bullet\}$, - СВ числа приборов, выделяемых фазе m в такте n. Пусть при $\boldsymbol{\xi}_{n-1} = \mathbf{q}$, $\mathbf{q}^T := (q_1, q_2, ..., q_M), q_m = \overline{0, r_m}$, приборы числом $\boldsymbol{\gamma}_n^T := (\gamma_{1,n}, \gamma_{2,n}, ..., \gamma_{M,n}) = (c'_1, c'_2, ..., c'_M) = \mathbf{c'}^T$ из множества общих приборов выделяются фазам с вероятностями
$\mathbf{g}_{\mathbf{c}'}^T(\mathbf{q}) := (g_{1,c'_1}, g_{2,c'_2}, ..., g_{M,c'_M})(\mathbf{q}) = P\{\boldsymbol{\gamma}_n = \mathbf{c}' \mid \boldsymbol{\xi}_{n-1} = \mathbf{q}\}, n \geq 1$, где
$\mathbf{g}_{\mathbf{c}'}^T(\mathbf{q})\mathbf{1} = 1, c'_m = \overline{0, c - c_\bullet}, \mathbf{c}' \in C'$
$C' := \{\mathbf{c}' : c'_\bullet = c - c_\bullet\}, \mid C' \mid = \binom{M+c-c_\bullet-1}{c-c_\bullet} = \frac{(M+c-c_\bullet-1)!}{(c-c_\bullet)!(M-1)!}$,
при условии

$$\exists m \forall \mathbf{q} := c_m = 0 \wedge g_{m,0}(\mathbf{q} = 1). \tag{1}$$

Условие (1) обеспечивает возможность выделения фазе m приборов из множества распределяемых при отсутствии на ней постоянно закрепленных приборов. Чтобы подчеркнуть зависимость в общем случае выбора c'_m от состояния СМО на предыдущем такте в ряде мест далее по тексту используется мнемоника $c'_m(\mathbf{q})$. Отметим, что распределение приборов

может осуществляться и в качестве последнего действия в последовательности событий в рассматриваемой СМО. Если приведенный выше вариант последовательности событий позволяет учесть сложившиеся на предыдущем такте потребности фаз в обслуживании заявок, то при распределении приборов в конце такта можно учесть потребности фаз, сформировавшиеся за текущий такт.

Пусть $\beta_{m,n}, \beta_{m,n} \in \{0,1\}$- СВ числа окончаний обслуживания за такт n на любом из приборов, занятом обслуживанием заявки на фазе m, причем вероятность завершения обслуживания заявки на приборе в такте n зависит в общем случае от ξ_{n-1}:
$b_m(\mathbf{q} := P\{\beta_{m,n} = 1 \mid \xi_{n-1} = \mathbf{q}\}, 0 < b_m(\mathbf{q}) \leq 1, n \geq 1,$
$\mathbf{b}^T(\mathbf{q}) := (b_1, b_2, ..., b_M)(\mathbf{q}).$

С дополнительной вероятностью $\overline{b}_m(\mathbf{q})$ обслуживание заявки будет продолжено, в т.ч. в следующем такте при наличии прибора. Для удобства здесь и далее черта над обозначением вероятности будет означать ее дополнение до единицы. Таким образом, процесс обслуживания прибором в текущем такте определяется распределением Бернулли с параметром, зависящим от состояния СМО на предыдущем такте. Введем обозначение $c''_m(\mathbf{q}) := c_m + c'_m(\mathbf{q})$ для общего числа приборов на фазе m после распределения приборов, $c''^T(\mathbf{q}) := (c''_1, c''_2, ..., c''_M)(\mathbf{q}) = \mathbf{c}^T + \mathbf{c}'^T(\mathbf{q})$.

Пусть $\eta_{m,n}, \eta_{m,n} \in \{0, 1, ..., c''_m(\mathbf{q}) \circ q_m\}$ - СВ числа окончаний обслуживания за такт n на приборах, занятых обслуживанием заявок на фазе m. Здесь и далее используется обозначение $x \circ y := \min(x, y)$, где x и y могут быть выражениями, и знак $\circ$ действует на выражение x слева и выражение y справа от него; обозначение $x \circ y$ есть минимум между x и y без учета других членов выражения, в котором стоит $x \circ y$. Вследствие предположения о распределении Бернулли окончания обслуживания на приборе, распределение числа обслужившихся за такт на фазе m заявок имеет биномиальное распределение: $\eta_{m,n} \sim \mathrm{Bin}(c''_m(\mathbf{q}) \circ q_m, b_m(\mathbf{q}))$. В случае ухода с фазы m прибора, обслуживавшего на предыдущем такте заявку, не завершившую за этот такт обслуживание (частично обслуженную), она ожидает в $\mathrm{БН}_m$ его продолжения при поступлении прибора в будущем.

Пусть заявка, завершившая обслуживание на фазе m, с вероятностью $d_m(\mathbf{q})$ уходит из СМО, и с вероятностью $\overline{d}_m(\mathbf{q})$ поступает на фазу $m+1$,$0 \leq d_m(\mathbf{q}) < 1, m = \overline{1, M-1}, d_M(\mathbf{q}) = 1, \mathbf{d}^T(\mathbf{q}) := (d_1, d_2, ..., d_M)(\mathbf{q})$. Отметим, что распределение числа покидающих СМО заявок из закончивших обслуживание за такт n на фазе m также имеет биномиальное распределение.

Будем считать, что заявка, которой не хватило места для буферизации в $\mathrm{БН}_m$ с $r_m < \infty$, теряется, не возобновляется и не оказывает влияния на дальнейшее функционирование системы. Как указывалось ранее на рис.1 и в описании к нему, обслуживаемая заявка занимает одно место в БН, которое освобождается по окончании обслуживания.

Пусть $\alpha_{m,n}, \alpha_{m,n} \in \{0,1\}$ - СВ числа поступлений извне групп заявок на фазу m за такт n, причем вероятность поступления группы в такте n зависит от состояния СМО $\boldsymbol{\xi}_{n-1}$ на такте $n-1$ и равна
$a_m(\mathbf{q}) := P\{\alpha_{m,n} = 1 \mid \boldsymbol{\xi}_{n-1} = \mathbf{q}\}, m = \overline{1,M}, 0 < a_l(\mathbf{q}) < 1, l = \overline{2,M}, n \geq 1.$

Хотя $a_l(\mathbf{q})$ в общем случае может принимать нулевые значения для некоторых $\mathbf{q}$ при условии не блокирования полностью входящего потока, без ограничения общности будем полагать, что $\forall \mathbf{q} : a_1(\mathbf{q}) \neq 0$. Вероятность поступления извне на фазу m за такт группы заявок объема k обозначим

$$a_{m,k}(\mathbf{q}) = \begin{cases} \overline{a}_m(\mathbf{q}) & ,k = 0 \\ a_m(\mathbf{q}) t_k & ,k = \overline{1,K_m} \end{cases},$$

где K_m - максимальное значение объема поступившей группы заявок на фазу m за такт n и t_k - вероятность группы объема k.

Таким образом, поступающие на фазы СМО потоки заявок зависят от состояния системы, являются неординарными геометрическими потоками второго рода и обозначаются $Geom^G(\mathbf{q})$.

Предложенная выше дисциплина распределения приборов между фазами обладает абсолютным приоритетом по отношению к обслуживаемым заявкам: прибор может прервать обслуживание заявки и перейти на другую фазу.

1.2. Мнемоническое обозначение. Будем обозначать аббревиатурой SAT наличие процедуры распределения приборов между фазами многофазной СМО (от английского выражения Servers Allocation in Tandem Queues).

В соответствии с классификацией Кендалла-Башарина [1], установившейся для СМО в дискретном времени мнемоникой [2] и предложенными обозначениями условимся обозначать введенную многофазную СМО как $(Geom^G(\mathbf{q}) \mid Bin(\mathbf{q}) \mid c''_m(\mathbf{q}) \mid r_m \leq \infty \mid B(\mathbf{q}))^M_{m=1}/SAT$.

1.3. Пространство состояний. Функционирование СМО в ее общем виде при $r_m < \infty, m = \overline{1,M}$, можно описать однородным процессом $\xi_n^T := (\xi_{1,n}, \xi_{2,n}, ..., \xi_{M,n})$ по моментам $nh, n \geq 0$, над множеством состояний

$$Q = \bigcup_{q_1=0}^{r_1} \bigcup_{q_2=0}^{r_2} ... \bigcup_{q_M=0}^{r_M} (q_1, q_2, ..., q_M), \mid Q \mid = \prod_{m=1}^{M} R_m, R_m := r_m + 1. \quad (2)$$

Из сделанных выше предположений следует, что процесс $\xi_n, n \geq 0$ является однородной цепью Маркова, причем все состояния цепи $\xi_n, n \geq 0$ сообщаются и образуют один эргодический класс без подклассов. При этом предполагается, что введение зависимости нагрузочных параметров от состояния цепи Маркова на предыдущем такте не приводит к появлению несообщающихся состояний.

Таким образом, в данном разделе введена и описана новая многофазная СМО в дискретном времени с распределяемым между фазами множеством

приборов в самом общем виде. Это вызвано практическими задачами, в том числе задачами расчета гетерогенных сетей, в которых общие ресурсы передачи распределяются на каждом такте между базовой станцией и ретрансляционными станциями. Предложенный в статье подход с распределением множества общих приборов в многофазной СМО можно использовать также в сетях массового обслуживания и системах циклического обслуживания в дискретном времени. Для случая непрерывного времени как для многофазной СМО, так и для сетей массового обслуживания или систем циклического обслуживания временные интервалы между моментами перераспределения общих приборов, соответственно, между фазами, узлами, подсистемами должны задаваться вероятностным распределением.

2. Анализ различных схем распределения ресурсов на базе имитационной модели

2.1. Схемы распределения ресурсов. Для введенной в предыдущем разделе модели был разработан программный комплекс на языке программирования C++, реализующий имитационную модель, позволяющую проанализировать различные схемы распределения ресурсов между фазами. Следует отметить, что при численном анализе могут быть рассмотрены любые схемы, включая детерминированные и пропорциональные алгоритмы, однако, в данной статье приведено сравнение трех схем распределения ресурсов, предложенных в [3]: детерминированного, пропорционального и пропорционального с ограничениями (далее D, P, P-c соответственно).

В соответствии с алгоритмом D ресурсы выделяются фазам фиксированным образом, и не перераспределяются в ходе имитации. Пропорциональные методы (P, P-c), с зависимостью от состояния системы, нацелены на повышение общей пропускной способности соты за счет адаптированного к нагрузке распределения приборов между фазами, поэтому они могут быть отнесены к классу планировщиков Proportional Fair [4]. Данный класс планировщиков характеризуется выделением ресурсов пользователям с наилучшим качеством канала в целях повышения пропускной способности сети, учитывая при этом среднее число переданных бит всех активных пользователей в предыдущих тактах, что позволяет также обеспечить минимальным числом ресурсов пользователей, находящихся в плохих канальных условиях.

Особенностью алгоритма P-c является выделение ресурсов на обслуживание заявок группы объема, не приводящего к потерям на буферных накопителях следующих фаз. При этом, рассматривается сеть с централизованной архитектурой, при которой базовая станция обладает полной информацией о состоянии БН, длительности обслуживания заявок, и т. д.

2.2. Численный анализ. Моделирование проводится для трехфазной системы с числом приборов, равным 25. При этом предполагается, что емкости БН трех фаз имеют одинаковые значения, равные 15. Объем посту-

пающей группы заявок имеет распределение Пуассона со средним, равным числу закрепленных за фазами приборов, т. е. 5. Вероятности завершения обслуживания и ухода из системы приняты равными 0.5. Вероятности поступления групп заявок на вторую и третью фазы принимают значение 0.25. Выбор исходных данных имеет целью выявить тенденции поведения системы, но не исследовать ее в реальных условиях. На рис.2-3 представлены графики для некоторых показателей исследуемых алгоритмов распределения ресурсов, указанных в подрисуночных подписях, при изменении вероятности поступления a на первую фазу за такт.

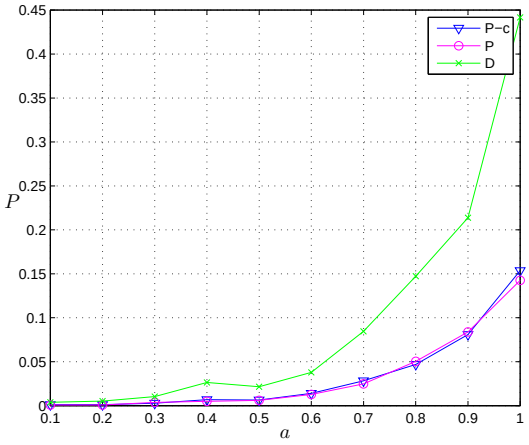


Рис. 2: Графики зависимости вероятности потерь в СМО от вероятности поступления заявки на СМО.

Как видно из графиков, с увеличением нагрузки детерминированный алгоритм D характеризуется наибольшей вероятностью потерь в СМО(рис.2). Отметим, что при исследовании гетерогенной сети LTE в виде двухфазной СМО с неординарным потоком неоднородных заявок и второй фазой сложной структуры, состоящей из параллельных СМО конечной емкости [3], пропорциональный метод P-c показал наилучший результат для различных вероятностно-временных характеристик. В данном численном эксперименте два алгоритма показывают сравнительно схожее поведение, что может быть объяснено более высоким числом случайных процессов, характерных для рассматриваемой СМО: введением вероятности ухода из системы, вероятности поступления группы заявок на каждую из фаз, вероятности обслуживания $b_m \neq 1$. На рис.3 показана зависимость среднего числа выделенных ресурсов на фазу 1 и фазу 2 в зависимости от изменения нагрузки на первую фазу слева направо, соотвественно. Отметим, что график для фазы 3 имеет схожее с фазой 2 поведение.

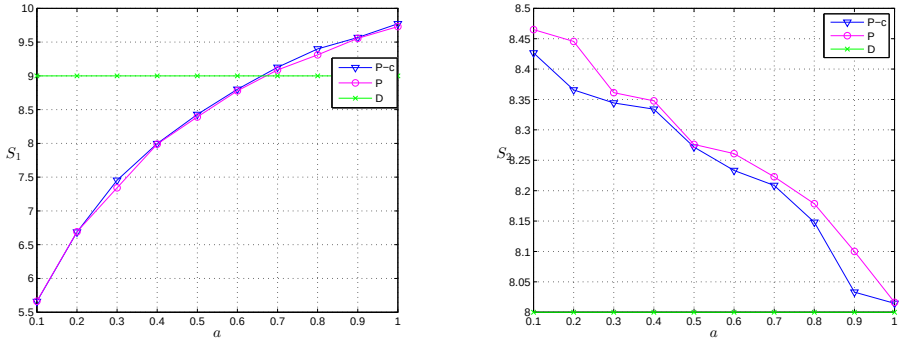


Рис. 3: Графики зависимости среднего числа выделенных ресурсов от вероятности поступления заявки на СМО для фаз 1 и 2, соответственно.

3. Заключение

В статье предлагается и исследуется многофазная СМО с распределяемым между фазами общим множеством приборов. Разработана имитационная модель, на базе которой приведен численный анализ различных схем распределния ресурсов на базе детерминированного и пропорциональных алгоритмов. Предложенный в статье подход с распределением множества общих приборов в многофазной СМО можно использовать также в сетях массового обслуживания и системах циклического обслуживания в дискретном времени.

ЛИТЕРАТУРА

1. Башарин Г.П., Харкевич А.Д., Шнепс М.А. Массовое обслуживание в телефонии // М.: Наука, 1968. - 247 с.
2. Башарин Г.П., Ефимушкин В.А. Исследование однолинейной системы с заявками нескольких типов в дискретном времени // Проблемы передачи информации. - 1984. - С. 95-104.
3. Ефимушкина Т.В. Исследование вероятностно-временных характеристик для усовершенствованной схемы распределения ресурсов в гетерогенной сети LTE // T-Comm -7 - Телекоммуникации и Транспорт. - 2013. - С. 58-65.
4. Кучерявый Е.А. Управление трафиком и качество обслуживания в сети Интернет // СПб.: Наука и техника, 2004. - 336 с.

MODELING AND PERFORMANCE ANALYSIS OF INTERCONNECTED SERVERS IN A CLOUD COMPUTING SYSTEM WITH DYNAMIC LOAD BALANCING

Udo R. Krieger
Otto-Friedrich-Universität, Fakultät WIAI
An der Weberei 5, D-96047 Bamberg, Germany
udo.krieger@ieee.org

Abstract

We consider the efficiency of dynamic resource pooling and allocation in a cloud computing system offering infrastructure-as-a-service (IaaS). We assume that the demand for service computing by virtual machines (VMs) follows a Poisson load pattern and that the response times of the provided computing services can be classified into several service categories that are governed by exponential service time patterns. A hierarchical, dynamic, class-dependent balancing policy based on a least-loading scheme is applied to provide a uniform utilization among the servers. It is derived from cascaded mutual overflow routing using information on the utilization of VM clusters of similar type on adjacent servers within this resource pool. Regarding the allocation of virtual machines of these different service types to the user demand by a pool of physical servers, we derive a Markovian loss model with adaptive routing induced by cascaded mutual overflow as effective, state-dependent load balancing policy. We determine its basic performance characteristics applying a Markovian fixed-point model. Based on the latter we gain insight on the power of the proposed dynamic load balancing policy among service classes.

Keywords: Cloud computing, IaaS, performance analysis, randomized load balancing, mutual overflow system

1. Introduction

In recent years modern Web services have been provided by cloud computing systems that are hosted by powerful infrastructures in modern data centers like Microsoft Windows Azure or Amazon Web Services. The latter environment constitutes an example of an infrastructure-as-a-service (IaaS) system where users can deploy their virtualized service computing systems on the physical resources of the infrastructure provider. Pooling the virtualized resources offers the chance to follow efficiently the dynamically changing demand curves of the Web services advertised by a service provider, and to satisfy the scalability, elasticity, and resilience requirements of service-oriented computing.
Effective load balancing and resource allocation schemes are an important ingredient

of IaaS systems. Recently, new randomized resource assignment policies based on sampling the utilization of physical servers such as the power-d scheme have been studied (cf. [2], [3], [4]). Following Mukhopadhyay et al. [3], we can argue that the resource allocation of virtual machines (VMs) on an interconnected cluster of physical servers may be considered as a randomized routing among loss systems hosting the VMs as service units. Then a utilization-oriented resource allocation policy that first senses the status of the physical servers, allocates VMs and tries to optimize the load assignment subject to loss minimization and uniformization constraints can be translated into a state-dependent routing policy of VM requests to the coupled loss systems (cf. [2], [3], [5]).
Related research [3] has revealed that randomized power-d load balancing schemes for interconnected loss systems are very powerful mechanisms, but their technical realization requires some overhead. Therefore, we propose a much simpler load balancing mechanism that is derived from classical mutual overflow routing (MOR) with state-dependent load splitting (cf. [1]). Its superior performance has already been revealed in the context of circuit-switched networking and guided a worldwide deployment in the switching equipment of two big European manufacturers.
We suppose that a cluster of interconnected physical machines is given which hosts groups of VMs as virtualized computing resources. The latter are divided into different service classes. VM groups of the same class on two neighboring servers are coupled by a MOR scheme. In this way we construct a hierarchical binary load balancing tree among groups of VMs of the same class. Considering a binary tree component, the load is balanced is such a way that a VM request is first offered to the least loaded component with the option to overflow to the other one in case of a fully loaded structure. In this way an effective adaptive local load balancing scheme can be extended to a tree structure.
In this paper we first analyze the derived performance model of a basic component of this IaaS system and describe a fixed-point model of the underlying coupled loss systems with dynamic load balancing in Section 2. Then we investigate its basic performance metrics in Section 3. Finally, some general conclusions are drawn.

2. Performance Modeling of Dynamic Load Balancing Among Virtualized Server Units in Cloud Computing

We consider a dynamic resource allocation of virtual machines hosted on the interconnected physical servers that is governed by a dynamic load balancing scheme. The latter uses both information about the utilization of the C virtual machines of M different service types and the number of different machines on the interconnected cluster of the $N = 2^k, k \in \mathbb{N}$, physical servers. Motivated by the approach of Mazumdar et al. [3] and references therein, we assume that the server $i \in \mathbb{S} = \{1, \ldots, N\}$ accomodates C_{ij} virtual machines as service units of M different service types $j \in \mathbb{T} = \{1, \ldots, M\}$. Then we suppose that the resulting number $C_j = \sum_{i \in \mathbb{S}} C_{ij}$ of virtualized servers of a certain type $j \in \mathbb{T}$ can be arranged such that $C_1 \leq C_2 \leq \ldots \leq C_M$ holds.

2.1. Dynamic Load Balancing by Mutual Overflow Routing. We propose to apply a dynamic load balancing scheme among two adjacent virtualized clusters of the same type $j \in \mathbb{T}$ on pairs $(i, i+1), i = 2l-1 \in \{1, \ldots, N-1\}, l = 1, \ldots N/2$ of adjacent servers that is derived from a mutual overflow scheme (see figure 1, cf. [1]). Then the scheme can be applied in a hierarchical way to the compound of two clustered servers $(i, i+1)$ and $(i+2, i+3)$ as single load balancing block of the mutual overflow scheme and so on. In this way a hierarchical, binary load balancing tree is formed among the service units of each service type.
The offered traffic to service category $j \in \mathbb{T}$ is resulting from a splitting of the overall load to all interconnected servers with rate λ_S by a splitting ratio $p_j^{(T)}$. Each server $i \in \mathbb{S}$ gets a conditional splitting ratio $p_i^{(S|T=j)}$ of the overall traffic of type j. We assume that the offered load is determined by a Poisson stream, hence, all traffic of service demand for virtual machines of a certain class $j \in \mathbb{T}$ at different servers $i \in \mathbb{S}$ is governed by Poisson processes with rates $\lambda_{ij} = \lambda_S \cdot p_j^{(T)} \cdot p_i^{(S|T=j)}$.

2.2. Performance Model of a Two-Server System. Let us now consider two adjacent servers $(i, i+1), i \in \{1, \ldots, N-1\}$ within the server farm as building block of the service infrastructure with hierarchical, dynamic load balancing among a certain service class. For simplicity we assume $i = 1$. We call the latter service computing system 1 and 2, respectively (see figure 1). Then we look at a fixed service category $j \in \mathbb{T}$, e.g. $j = 1$, and apply the mutual overflow scheme as basic load balancer among the $N_1 = C_{1j}$ and $N_2 = C_{2j}$ virtual machines on system 1 and 2, respectively, that are serving the incoming service requests of type j. In this case we interpret the system of parallel virtual machines on each server as fully available trunk groups 1 and 2 with Poisson arrival streams and rates λ_1, λ_2 as offered traffic 1 and 2, and exponential service times. Without loss of generality, we assume a common service time with rate $\mu = 1$. If all virtual machines in both service groups are busy, an arriving VM service request is lost in this combined server system. In the binary tree structure this portion of the traffic will overflow to the neighboring block of the server cascade. Thus, a coupled system of two isolated Erlang loss models can be used to describe the basic performance behavior of the coupled virtualized server cluster of fixed type j (cf. [1]). We propose to use both the information on the relative server capacities N_1, N_2 and the VM utilizations ρ_1, ρ_2 on the different servers 1 and 2 to allocate the incoming requests to the least loaded server. This state-dependent dynamic routing policy within the mutual overflow system between system 1 and 2 is modelled by an adaptive routing with a random splitting of the offered Poisson traffic of class j with common rate:

$$\lambda = \lambda_{ij} + \lambda_{i+1\,j} = \lambda_S \cdot p_j^{(T)} \cdot (p_i^{(S|T=j)} + p_{i+1}^{(S|T=j)}) \quad i = 1, j \in \mathbb{T}.$$

Due to the dynamic MOR-type load balancing subject to the server capacities and utilization states, we assume that this offered traffic with rate λ is randomly split up into two portions for system 1 and 2 with rates λ_1 and λ_2. Then the described policy

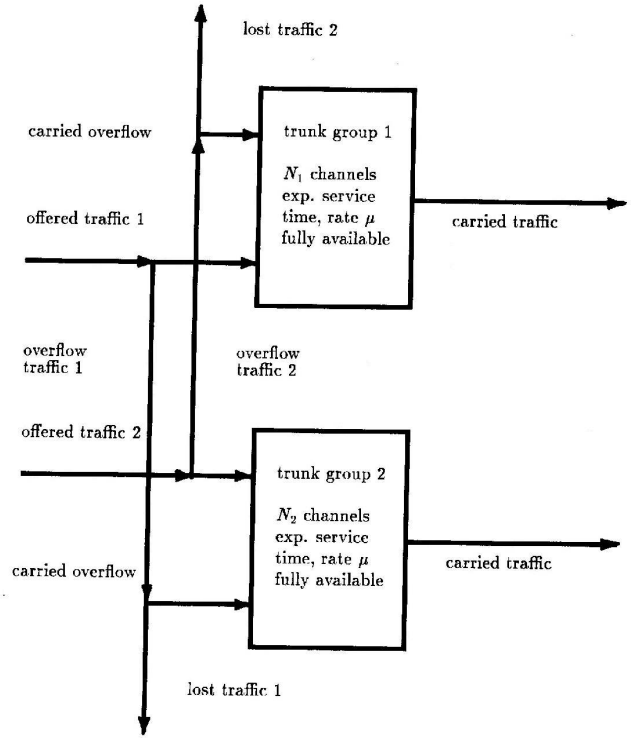


Figure 1: Principle of mutual overflow routing between two coupled loss systems of virtualized machines that are interpreted as trunk group 1 and 2.

can be modelled by the following splitting probabilities

$$p = \frac{\frac{N_1}{N_1+N_2}(1-\rho_1(N_1))}{1-\rho(N_1+N_2)}, \quad 1-p = \frac{\frac{N_2}{N_1+N_2}(1-\rho_2(N_2))}{1-\rho(N_1+N_2)} \tag{1}$$

of the Poisson stream with rate λ offered to the consecutive virtualized server groups on system 1 and 2. Here $\rho_1(N_1) = E(X_1)/N_1, \quad \rho_2(N_2) = E(X_2)/N_2$ denote the utilization of a single virtual machine on server 1 and 2, respectively, where the state variable X_k records the number of active virtual machines on server $k \in \{1,2\}$.

$$\rho(N_1+N_2) = E(X_1+X_2)/(N_1+N_2) = \frac{N_1}{N_1+N_2}\cdot\rho_1(N_1) + \frac{N_2}{N_1+N_2}\cdot\rho_2(N_2)$$

is the utilization of a single machine of the same type on two adjacent servers that are coupled by mutual overflow load balancing. Here we assume that the systems 1 and 2

will not be fully utilized. The latter case may be simply incorporated by adopting the modified splitting probability

$$p = \frac{\frac{N_1}{N_1+N_2}(1-\rho_1(N_1)+\varepsilon)}{1-\rho(N_1+N_2)+\varepsilon}, \quad 1-p = \frac{\frac{N_2}{N_1+N_2}(1-\rho_2(N_2)+\varepsilon)}{1-\rho(N_1+N_2)+\varepsilon}$$

for sufficiently small $\varepsilon > 0$. It yields a static splitting according to the relative number of VMs for fully utilized service blocks in the coupled VM clusters.
Then the resulting fresh Poisson stream of type j to system 1 has the rate

$$\lambda_1 = \lambda \cdot p = \lambda \cdot \frac{\frac{N_1}{N_1+N_2}(1-\rho_1(N_1))}{1-\rho(N_1+N_2)} \tag{2}$$

and that one offered to system 2 has the rate

$$\lambda_2 = \lambda \cdot (1-p) = \lambda \cdot \frac{\frac{N_2}{N_1+N_2}(1-\rho_2(N_2))}{1-\rho(N_1+N_2)} \tag{3}$$

The flow analysis of the interacting server systems coupled by the adaptive routing which is induced by the mutual overflow load balancing now yields the following offered traffic rates L_1 and L_2 on system 1 and 2, respectively:

$$L_1 = L_1(p) = \lambda_1 + \lambda_2 \cdot B_2 = \lambda \cdot p + \lambda \cdot (1-p) \cdot B_2 = \lambda[1-(1-p)\cdot(1-B_2)] \tag{4}$$

$$L_2 = L_2(p) = \lambda_2 + \lambda_1 \cdot B_1 = \lambda \cdot (1-p) + \lambda \cdot p \cdot B_1 = \lambda[1-p\cdot(1-B_1)] \tag{5}$$

Here the terms B_1 and B_2 denote the arrival-stationary blocking probabilities of system 1 and 2, respectively. If we make the simplifying assumption that the Markov-modulated overflow traffic is again approximated by Poisson streams, the latter coincide with the time-stationary blocking probabilities due to the PASTA property.

2.3. Fixed-Point Approximation of the Blocking Probabilities. The time-stationary blocking probabilities B_1, B_2 are determined by Erlang's formula $B = E(N, A)$ for a pure loss system with offered load A and N service units if we suppose that the overflow streams are Poisson flows. Then they are given by the following quantities

$$B_1 = f_1(N_1, L_1, B_2, p) = {L_1}^{N_1}/[N_1!(\sum_{i=0}^{N_1} {L_1}^i/i!)] = E(N_1, \lambda\cdot[1-(1-p)\cdot(1-B_2)])$$

$$B_2 = f_2(N_2, L_2, B_1, p) = {L_2}^{N_2}/[N_2!(\sum_{j=0}^{N_2} {L_2}^j/j!)] = E(N_2, \lambda\cdot[1-p\cdot(1-B_1)])$$

if we assume without loss of generality that the service rates of all classes are uniformly given by $\mu = 1$. In [1] it has been revealed that this approximation of the Markov-modulated overflow streams by Poisson flows with identical overflow rates yields a very accurate approximation of the blocking behavior.

For any fixed splitting probability $p \in (0,1)$ this overall Erlang loss model yields a system of fixed-point equations $F = (f_1 \circ f_2, f_2 \circ f_1) : I \to I$ on the compact unit square $I = [0,1]^2$ to determine the blocking probabilities $(B_1(p), B_2(p)) \in [0,1]^2$ by

$$B_1 = B_1(p) = E\left(N_1, \lambda \cdot [1 - (1-p) \cdot (1 - E(N_2, \lambda \cdot [1 - p \cdot (1 - B_1)]))]\right) \quad (6)$$

$$B_2 = B_2(p) = E\left(N_2, \lambda \cdot [1 - p \cdot (1 - E(N_1, \lambda \cdot [1 - (1-p) \cdot (1 - B_2)]))]\right). \quad (7)$$

The existence of a fixed point $B^*(p) = (B_1^*(p), B_2^*(p)) \in [0,1]^2$ is guaranteed by Brower's fixed-point theorem. In [1] it was shown that for fixed $p \in (0,1)$ the independent offered Poisson streams with the positive rates λ_1, λ_2 in (2), (3) determine even a unique fixed point $B^*(p)$ due to the monotonicity of Erlang's loss formula. Then they can be computed by a simple power iteration, e.g. $B_1^{(n)} = [f_1 \circ f_2]^n(B_1^{(0)}), B_1^{(0)} = E(N_1, L_1(p))$. Both blocking terms B_1, B_2 arising as fixed point $B^*(p) = (B_1, B_2)$ in (6, 7) are coupled by the common splitting probability $p = g(B_1, B_2)$ in (1) which depends in a nonlinear manner on the individual server utilizations

$$\rho_1(N_1) = g_1(N_1, L_1, B_1) = E(X_1)/N_1 = \frac{1}{N_1} \cdot L_1 \cdot (1 - B_1)$$

$$\rho_2(N_2) = g_2(N_2, L_2, B_2) = E(X_2)/N_2 = \frac{1}{N_2} \cdot L_2 \cdot (1 - B_2)$$

and, hence, blocking probabilities in both loss systems 1 and 2, and on the server utilization

$$\rho(N_1 + N_2) = \frac{E(X_1 + X_2)}{(N_1 + N_2)} = \frac{N_1}{N_1 + N_2} \cdot \frac{L_1}{N_1} \cdot (1 - B_1) + \frac{N_2}{N_1 + N_2} \cdot \frac{L_2}{N_2} \cdot (1 - B_2)$$

in the overall system. Hence, the splitting probability $p \in (0,1)$ in (1) is determined by the resulting fixed-point equation:

$$p = h(B_1, B_2, p) = \frac{\frac{N_1}{N_1+N_2}(1 - \frac{L_1}{N_1}(1 - B_1))}{1 - [\frac{N_1}{N_1+N_2}\frac{L_1}{N_1}(1 - B_1) + \frac{N_2}{N_1+N_2}\frac{L_2}{N_2}(1 - B_2)]}$$

$$= \frac{N_1 - L_1(p)(1 - B_1(p))}{N_1 - L_1(p)(1 - B_1(p)) + N_2 - L_2(p)(1 - B_2(p))} \quad (8)$$

$$= \frac{N_1 - \lambda \cdot [1 - (1-p) \cdot (1 - B_2(p))](1 - B_1(p))}{\begin{array}{r}(N_1 - \lambda \cdot [1 - (1-p) \cdot (1 - B_2(p))](1 - B_1(p)) \\ +N_2 - \lambda \cdot [1 - p \cdot (1 - B_1(p))](1 - B_2(p)))\end{array}} \quad (9)$$

The corresponding fixed-point model (4, 5, 6, 7, 9) of the combined splitting-blocking model $X = (B_1(p), B_2(p), p) \in [0,1]^3$ is simple, but analytically complex due to the ratio structure of the term p. It can be solved by a power iteration $p^{(n)} = h(B_1, B_2, p^{(n-1)})$, $n \in \mathbb{N}$, $p^{(0)} = 0.5$ whose local convergence to a fixed point $X^* = (B_1^*(p^*), B_2^*(p^*), p^*)$ is guaranteed by Brower's fixed-point theorem. Starting with the outcome of our previous analysis [1], we can reveal the dependence on the splitting parameter p^* by a

Steady-state representation result.
We consider the basic IaaS component of two servers that host N_1 and N_2 virtual machines (VMs) and serves Poisson arrival streams with offered loads $\lambda_1 = \lambda p^*$ and $\lambda_2 = \lambda(1 - p^*)$, respectively. They are coupled by mutual overflow routing combined with a least-load balancing scheme with splitting probability p^* in (1). The steady-state distribution $\Pi = (\pi_{i,j}), \pi_{i,j} = \lim_{t\to\infty} \mathbb{P}(X(t) = (i, j))$ of the resulting ergodic loss model $X = (X_1, X_2) \in [0, N_1] \times [0, N_2] \subset \mathbb{N}^2$ that describes the number of active VMs in server 1 and 2 is determined by a perturbed product-form solution

$$\begin{aligned}\pi_{i,j} \quad = \quad & \sum_{k=0}^{N_1} c_k(p^*) g_i(\rho_k(p^*), \lambda_1, p^*) \cdot g_j(-\rho_k(p^*), \lambda_2, p^*) \\ & + \sum_{l=0}^{N_2} d_l(p^*) g_i(\gamma_l(p^*), \lambda_1, p^*) \cdot g_j(-\gamma_l(p^*), \lambda_2, p^*) \end{aligned} \tag{10}$$

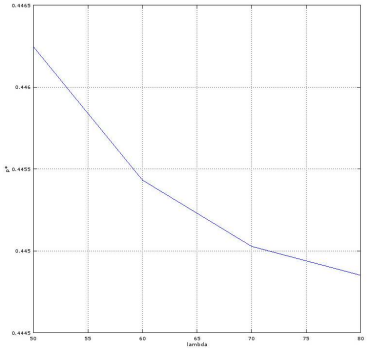
for $0 \le i \le N_1, 0 \le i \le N_2$. The parameters $\rho_k(p^*) \in (-\infty, -1), 1 \le k \le N_1$, are the distinct zeros of the Brockmeyer polynomial $g_{N_1}(1 + \rho(p^*), \lambda_1, p^*)$ and $\gamma_l(p^*) \in (1, \infty), 1 \le l \le N_2$, the distinct zeros of the Brockmeyer polynomial $g_{N_2}(1-\gamma(p^*), \lambda_2, p^*)$ and $\rho_0(p^*) = \gamma_0(p^*) = 0$. The $N_1 + N_2$ coefficients $c_k(p^*), 0 \le k \le N_1, d_l(p^*), 0 \le l \le N_2$ are the unique solution of a linear system determined by these Brockmeyer polynomials. All terms dependent in unique manner on the existing fixed-point $p^* \in [0, 1]$ of the non-linear system (9). They uniquely determine the IaaS blocking probability:

$$\begin{aligned}\pi_{N_1,N_2} \quad = \quad & E(N_1, \lambda_1) \cdot E(N_2, \lambda_2) + \sum_{k=1}^{N_1} c_k(p^*) g_i(\rho_k(p^*), \lambda_1, p^*) \cdot g_j(-\rho_k(p^*), \lambda_2, p^*) \\ & + \sum_{l=1}^{N_2} d_l(p^*) g_i(\gamma_l(p^*), \lambda_1, p^*) \cdot g_j(-\gamma_l(p^*), \lambda_2, p^*) \end{aligned} \tag{11}$$

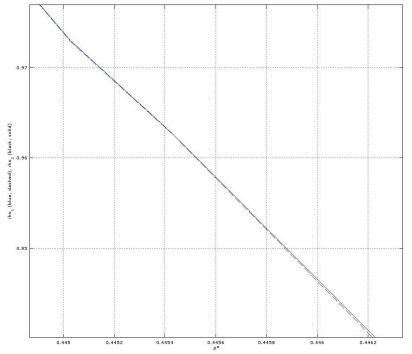
Combining all these coupled loss models arising from the hierarchical binary load balancing tree, we are able to determine by these analytic means the vector of the server-dependent blocking probabilities B_{ij} of class j on server i and the overall blocking probabilities $B_{i,*}, B_{*,j}$ of all servers and classes as fundamental performance metrics of the sketched interconnected service computing system.

3. Performance Study

We have investigated the loss and utilization performance of a basic building block of the proposed IaaS computing model. It consists of two loss systems with $C_1 = 20$ and $C_2 = 25$ virtual machines of the same class coupled by mutual overflow routing with a least-load balancing policy in (1). In figure 2a) we depict the dependence relation of the adaptive splitting probability $p^* \in (0, 1)$ on the overall rate λ of the arriving Poisson traffic for a heavily loaded system. In figure 2b) we illustrate the dependence of the utilization ρ_1 of system 1 (blue, dashed line) and the corresponding utilization ρ_2 of system 2 (black, solid line) on the fixed-point of the splitting probability $p^* \in (0, 1)$.



a) Dependence relation of the splitting probability $p^* \in (0, 1)$ on the overall rate λ of the arriving Poisson traffic.



b) Dependence of the utilizations ρ_1, ρ_2 of system 1 and 2 on the fixed-point splitting probability $p^* \in (0, 1)$.

Figure 2: Dependence relations in two interconnected loss systems with $C_1 = 20$ and $C_2 = 25$ VMs and dynamic MOR-based load balancing.

4. Conclusions

We have investigated the basic component of a new IaaS computing model with dynamic load balancing that is derived from mutual overflow routing (MOR) among two physical servers with virtualized computing resources and a state-dependent splitting of the offered traffic based on a least-load policy. We have first derived a fixed-point model for the MOR scheme on an underlying binary tree of Erlang loss systems that can reflect the state-dependent load balancing policy. Then we have developed an analysis method to compute the loss performance of the service system. The outcome has been demonstrated by a case study of a single class IaaS block illustrating the dependence relation of the traffic splitting and the utilization vector of the system.

REFERENCES

1. Krieger, U.R. Analysis of a Loss System with Mutual Overflow. In Proc. ITC-Seminar, Peking, September 1988.
2. Maguluri, S. T., Srikant, R., and Ying, L. Stochastic models of load balancing and scheduling in cloud computing clusters. In Proceedings IEEE INFOCOM 2012.
3. Mukhopadhyay, A., Mazumdar, R. et al. The Power of Randomized Routing in Heterogeneous Loss Systems. In Proc. ITC 2015, Ghent, 2015.
4. Mitzenmacher, M. The power of two choices in randomized load balancing. IEEE Transactions on Parallel and Distributed Systems, vol. 12, pp. 1094–1104, 2001.
5. Turner, S. R. E. The effect of increasing routing choice on resource pooling. Probability in the Engineering and Informational Sciences, vol. 12, 109–124, 1998.

ON THE WAITING TIME IN THE DISCRETE CYCLIC–WAITING SYSTEM OF $GEO/G/1$ TYPE

L. Lakatos[1]

[1] Eotvos Lorand University, Budapest, Hungary
lakatos1948@freemail.hu

Abstract

We continue to examine a discrete time queueing system where the service of a customer may start at the moment of arrival or at moments differing from it by the multiples of a given cycle time. We find the distribution and the mean value of waiting time in the case of general service time distribution.

Keywords: Geo/G/1 cyclic-waiting system, optical signals

1. Introduction

Earlier we have considered a single-server queueing system where an entering customer might be accepted for service either at the moment of arrival or at moments differing from it by the multiples of a given cycle time T. Such problem was motivated by the transmission of optical signals: optical signals enter a node and they should be transmitted according to the FCFS rule. This information cannot be stored, if it cannot be served at once is sent to a delay line and returns to the node after having passed it. So, the signal can be transmitted from the node at the moment of its arrival or at moments that differ from it by the multiples of time necessary to pass the delay line. The original problem had been raised in connection with the landing of airplanes, later it appeared to be an exact model for the transmission of optical signals where because of the lack of optical RAM the fiber delay lines were used.

First this system was considered from the viewpoint of the number of present customers [1]. By using Koba's results [2] in [3] we investigated the distribution and characteristics of waiting time for the continuous time model. [4] solved this problem for the discrete time case if the service time had geometrical distribution. In the present paper we investigate the case of general discrete service time distribution.

2. Preliminaries and theorem

We shortly repeat Koba's results [2] to find the waiting time distribution in the cyclic-waiting system.

Let t_n denote the time of arrival of the nth customer; its service will begin at the moment $t_n + T \cdot X_n$, where T is the cycle time and X_n is a nonnegetive integer. Let $\xi_n = t_{n+1} - t_n$ and η_n the service time of nth customer. Furthermore, let $X_n = i$, if

$$(k-1)T < iT + \eta_n - \xi_n \leq kT \qquad (k \geq 1),$$

then $X_{n+1} = k$, and if $iT + \eta_n - \xi_n \le 0$, then $X_{n+1} = 0$. Hence, X_n is a homogeneous Markov chain with transition probabilities p_{ik}, where

$$p_{ik} = P\{(k-i-1)T < \eta_n - \xi_n \le (k-i)T\}$$

if $k \ge 1$, and

$$p_{i0} = P\{\eta_n - \xi_n \le -iT\}.$$

Introduce the notations

$$f_j = P\{(j-1)T < \eta_n - \xi_n \le jT\}, \tag{1}$$

$$p_{ik} = f_{k-i} \quad \text{if} \quad k \ge 1, \quad p_{i0} = \sum_{j=-\infty}^{-i} f_j = \hat{f}_i. \tag{2}$$

The ergodic distribution of this chain satisfies the system of equations

$$p_j = \sum_{i=0}^{\infty} p_i p_{ij} \qquad (j \ge 0),$$

$$\sum_{j=0}^{\infty} p_j = 1.$$

Let us divide the cycle time T into n equal parts called slots. For each slot a new customer may enter with probability r, there is no entry with probability $1 - r$; the service time is k slots with probability q_k. Denoting the interarrival time by ξ, the service time by η, their distributions are

$$P\{\xi = k\} = (1-r)^{k-1} r, \qquad P\{\eta = k\} = q_k \quad (k \ge 1),$$

i.e. they have geometrical and general distributions, respectively.

Our result is formulated in the following

Theorem 1. *Let us consider the above described system and introduce a Markov chain whose states correspond to the waiting time (in the sense the waiting time is the number of actual state multiplied by T) at the arrival time of customers. The matrix of transition probabilities for this chain is*

$$\begin{bmatrix}
\sum_{j=-\infty}^{0} f_j & f_1 & f_2 & f_3 & f_4 & \cdots \\
\sum_{j=-\infty}^{-1} f_j & f_0 & f_1 & f_2 & f_3 & \cdots \\
\sum_{j=-\infty}^{-2} f_j & f_{-1} & f_0 & f_1 & f_2 & \cdots \\
\sum_{j=-\infty}^{-3} f_j & f_{-2} & f_{-1} & f_0 & f_1 & \cdots \\
\vdots & \vdots & \vdots & \vdots & \vdots & \ddots
\end{bmatrix}$$

its elements are defined by (1) and (2). The generating function of the ergodic distribution is

$$P(z) = \left[1 - \frac{\sum_{i=1}^{\infty} q_i \left\{ \left\lceil \frac{i}{n} \right\rceil - (1-r)^{i-1 \pmod n} \frac{1-(1-r)^{\lceil \frac{i}{n} \rceil n}}{1-(1-r)^n} \right\}}{\frac{Q_1(1-r)^n}{(1-r)[1-(1-r)^n]}}\right] \times$$

$$\times \frac{\frac{Q_1}{1-r} - \frac{Q_1[1-(1-r)^n]}{1-r} \frac{z}{z-(1-r)^n}}{1 - F_+(z) - \frac{Q_1[1-(1-r)^n]}{1-r} \frac{z}{z-(1-r)^n}}, \tag{3}$$

where

$$F_+(z) = \sum_{j=1}^{\infty} f_j z^j, \qquad Q_1 = \sum_{j=1}^{\infty} q_j (1-r)^j;$$

the condition of existence of ergodic distribution is

$$\frac{\sum_{i=1}^{\infty} q_i \left\{ \left\lceil \frac{i}{n} \right\rceil - (1-r)^{i-1 \pmod n} \frac{1-(1-r)^{\lceil \frac{i}{n} \rceil n}}{1-(1-r)^n} \right\}}{\frac{Q_1(1-r)^n}{(1-r)[1-(1-r)^n]}} < 1. \tag{4}$$

3. Proof of the Theorem

We find the transition probabilities f_j. We have

$$P\{\xi = k\} = (1-r)^{k-1} r, \qquad P\{\eta = k\} = q_k,$$

and we look for the distribution of $\eta - \xi$. Let $\eta - \xi = j > 0$. Then $\eta = \xi + j$, and the probability of this event is

$$\hat{q}_j = \sum_{i=1}^{\infty} (1-r)^{i-1} r q_{i+j} = \frac{r}{(1-r)^{j+1}} Q_{j+1},$$

where

$$Q_i = \sum_{k=i}^{\infty} q_k (1-r)^k.$$

If $\eta - \xi = -j \le 0$ $(j \ge 0)$, then $\eta + j = \xi$, and we have

$$\hat{q}_{-j} = P\{\eta - \xi = -j\} = \sum_{i=1}^{\infty} q_i (1-r)^{i+j-1} r = \frac{r(1-r)^j}{1-r} Q_1.$$

The transition probabilities f_j are obtained from these values, namely

$$f_j = \sum_{k=(j-1)n+1}^{jn} \hat{q}_k = \sum_{k=(j-1)n+1}^{jn} \frac{r}{(1-r)^{k+1}} \sum_{i=k+1}^{\infty} q_i (1-r)^i$$

for the positive jumps, and

$$f_{-j} = \sum_{k=jn}^{(j+1)n-1} \hat{q}_{-k} = \sum_{k=jn}^{(j+1)n-1} r(1-r)^{k-1} Q_1 = \frac{rQ_1}{1-r} \sum_{k=jn}^{(j+1)n-1} (1-r)^k =$$

$$= \frac{Q_1[1-(1-r)^n]}{1-r}(1-r)^{jn}$$

for the nonpositive jumps. Later we show that in the case of positive jumps

$$f_j = \sum_{i=(j-1)n+2}^{jn} q_i \left[1-(1-r)^{i-(j-1)-1}\right] + \sum_{i=jn+1}^{\infty} q_i [1-(1-n)^n](1-r)^{i-jn-1}.$$

Furthermore, let

$$\hat{f}_j = \sum_{i=-\infty}^{-j} f_i = \sum_{i=j}^{\infty} \frac{Q_1[1-(1-r)^n]}{1-r}(1-r)^{in} = \frac{Q_1}{1-r}(1-r)^{jn}.$$

By using the transition probabilities f_j, the ergodic probabilities are the solution of the system of equations

$$\begin{aligned} p_0 &= p_0\hat{f}_0 + p_1\hat{f}_1 + p_2\hat{f}_2 + p_3\hat{f}_3 + \dots \\ p_1 &= p_0 f_1 + p_1 f_0 + p_2 f_{-1} + p_3 f_{-2} + \dots \\ p_2 &= p_0 f_2 + p_1 f_1 + p_2 f_0 + p_3 f_{-1} + \dots \\ &\vdots \end{aligned}$$

Multiplying the j-th equation by z^j, summing up from zero to infinity, for the generating function $\sum_{j=0}^{\infty} p_j z^j$ we have (as in [1])

$$P(z) = P(z)F_+(z) + \sum_{j=1}^{\infty} p_j z^j \sum_{i=0}^{j-1} f_{-i} z^{-i} + \sum_{j=0}^{\infty} p_j \hat{f}_j, \tag{5}$$

where $F_+(z)$ is the generating function of positive jumps, because of its complexity its determination is given later.

In this expression

$$\sum_{i=0}^{j-1} f_{-i} z^{-i} = \sum_{i=0}^{j-1} \frac{Q_1[1-(1-r)^n]}{1-r}(1-r)^{in} z^{-i} = \\ = \frac{Q_1[1-(1-r)^n]}{1-r} \frac{1-\left(\frac{(1-r)^n}{z}\right)^j}{1-\frac{(1-r)^n}{z}},$$

$$\sum_{j=1}^{\infty} p_j z^j \sum_{i=0}^{j-1} f_{-i} z^{-i} = \sum_{j=1}^{\infty} p_j z^j \frac{Q_1[1-(1-r)^n]}{1-r} \frac{1-\left(\frac{(1-r)^n}{z}\right)^j}{1-\frac{(1-r)^n}{z}} = \\ = \frac{Q_1[1-(1-r)^n]}{1-r} \frac{z}{z-(1-r)^n} \left[P(z) - P((1-r)^n)\right],$$

$$\sum_{j=0}^{\infty} p_j \hat{f}_j = \sum_{j=0}^{\infty} p_j \frac{Q_1}{1-r}(1-r)^{jn} = \frac{Q_1}{1-r} P((1-r)^n).$$

Substituting these expressions, (5) yields

$$P(z)\left[1 - F_+(z) - \frac{Q_1[1-(1-r)^n]}{1-r}\frac{z}{z-(1-r)^n}\right] = \\ = P((1-r)^n)\left[\frac{Q_1}{1-r} - \frac{Q_1[1-(1-r)^n]}{1-r}\frac{z}{z-(1-r)^n}\right].$$

The value of $P((1-r)^n)$ can be found from the condition $P(1) = 1$. By using l'Hospital's rule we get

$$P((1-r)^n) = 1 - F'_+(1)\frac{(1-r)[1-(1-r)^n]}{Q_1(1-r)^n},$$

and the generating function takes on the form

$$P(z) = \left[1 - F'_+(1)\frac{(1-r)[1-(1-r)^n]}{Q_1(1-r)^n}\right] \times \\ \times \frac{\frac{Q_1}{1-r} - \frac{Q_1[1-(1-r)^n]}{1-r}\frac{z}{z-(1-r)^n}}{1 - F_+(z) - \frac{Q_1[1-(1-r)^n]}{1-r}\frac{z}{z-(1-r)^n}}. \qquad (6)$$

From this generating function

$$p_0 = \left[1 - F'_+(1)\frac{(1-r)[1-(1-r)^n]}{Q_1(1-r)^n}\right]\frac{Q_1}{1-r}.$$

It is positive if

$$F'_+(1)\frac{(1-r)[1-(1-r)^n]}{Q_1(1-r)^n} < 1, \tag{7}$$

which serves as stability condition.

4. The generating function of positive jumps

As we have mentioned the probability of $\eta - \xi = j$ $(j \geq 1)$ is

$$\hat{q}_j = \frac{r}{(1-r)^{j+1}} \sum_{i=j+1}^{\infty} q_i(1-r)^i = r \sum_{i=j+1}^{\infty} q_i(1-r)^{i-j-1}.$$

By using $\hat{q}_j$ we compute the transition probabilities f_j. We have

$$f_1 = \sum_{i=1}^{n} \hat{q}_i = \hat{q}_1 + \hat{q}_2 + \ldots + \hat{q}_n,$$

this sum is represented in the table (for the sake of simplicity we omit the factor r in all elements)

$$\begin{array}{lllllll}
q_2 & q_3(1-r) & \ldots & q_n(1-r)^{n-2} & q_{n+1}(1-r)^{n-1} & q_{n+2}(1-r)^n & \ldots \\
 & q_3 & \ldots & q_n(1-r)^{n-3} & q_{n+1}(1-r)^{n-2} & q_{n+2}(1-r)^{n-1} & \ldots \\
 & & & \vdots & \vdots & \vdots & \\
 & & & q_n & q_{n+1}(1-r) & q_{n+2}(1-r)^2 & \ldots \\
 & & & & q_{n+1} & q_{n+2}(1-r) & \ldots
\end{array}$$

Consequently,

$$f_1 =$$

$$= rq_2 + rq_3[1 + (1-r)] + rq_4[1 + (1-r) + (1-r)^2] + \ldots + rq_n[1 + \ldots + (1-r)^{n-2}] +$$

$$+ rq_{n+1}[1 + (1-r) + \ldots + (1-r)^{n-1}] + rq_{n+2}[(1-r) + (1-r)^2 + \ldots + (1-r)^n] +$$

$$+ rq_{n+3}[(1-r)^2 + (1-r)^3 + \ldots + (1-r)^{n+1}] + \ldots =$$

$$= \sum_{i=2}^{n} q_i[1-(1-r)^{i-1}] + \sum_{i=n+1}^{\infty} q_i[1-(1-r)^n](1-r)^{i-n-1}.$$

On a similar way one gets

$$f_2 =$$

$$\sum_{i=n+2}^{2n} q_i\left[1-(1-r)^{i-n-1}\right] + \sum_{i=2n+1}^{\infty} q_i\left[1-(1-r)^n\right](1-r)^{i-2n-1},$$

and in the general case

$$f_k = \sum_{i=(k-1)n+2}^{kn} q_i\left[1-(1-r)^{i-(k-1)n-1}\right] + \sum_{i=kn+1}^{\infty} q_i\left[1-(1-r)^n\right](1-r)^{i-kn-1}.$$

After some arithmetics the corresponding generating function will be

$$F_+(z) = \sum_{k=1}^{\infty} z^k \left\{ \sum_{i=(k-1)n+2}^{kn} q_i + \frac{1}{(1-r)^{kn+1}} \sum_{i=kn+1}^{\infty} q_i(1-r)^i - \right.$$

$$\left. - \frac{1}{(1-r)^{(k-1)n+1}} \sum_{i=(k-1)n+2}^{\infty} q_i(1-r)^i \right\}.$$

Its derivative at $z = 1$ is

$$F'_+(1) = \sum_{i=2}^{n} q_i + 2 \sum_{i=n+2}^{2n} q_i + 3 \sum_{i=2n+2}^{3n} + \ldots + k \sum_{i=(k-1)n+2}^{kn} q_i + \ldots +$$

$$+ \sum_{k=1}^{\infty} k q_{(k-1)n+1} - \sum_{k=1}^{\infty} \frac{1}{(1-r)^{(k-1)n+1}} \sum_{i=(k-1)n+1}^{\infty} q_i(1-r)^i =$$

$$= \sum_{k=1}^{\infty} k \sum_{i=(k-1)n+1}^{kn} q_i - \sum_{k=1}^{\infty} \frac{1}{(1-r)^{(k-1)n+1}} \sum_{i=(k-1)n+1}^{\infty} q_i(1-r)^i.$$

The second term of this expression is

$$\sum_{k=1}^{\infty} \sum_{i=(k-1)n+1}^{\infty} q_i(1-r)^{i-(k-1)n-1},$$

or

$$\begin{array}{cccccccc} q_1 & q_2(1-r) & \ldots & q_n(1-r)^{n-1} & q_{n+1}(1-r)^n & \ldots & q_{2n+1}(1-r)^{2n} & \ldots \\ & & & & q_{n+1} & \ldots & q_{2n+1}(1-r)^n & \ldots \\ & & & & & & q_{2n+1} & \ldots \end{array}$$

etc.

From each n columns one can factor out

$$q_i(1-r)^{i-1 \pmod n},$$

there remain the powers of $(1-r)^n$, their sums are of the form $\frac{1-(1-r)^{jn}}{1-(1-r)^n}$, i.e. in the first n columns $\frac{1-(1-r)^n}{1-(1-r)^n}$, in the second n columns $\frac{1-(1-r)^{2n}}{1-(1-r)^n}$, etc. The resulting sum will be

$$\sum_{i=1}^{\infty} q_i(1-r)^{i-1 \pmod n} \frac{1-(1-r)^{\lceil \frac{i}{n} \rceil n}}{1-(1-r)^n},$$

and

$$F'_+(1) = \sum_{k=1}^{\infty} k \sum_{i=(k-1)n+1}^{kn} q_i - \sum_{i=1}^{\infty} q_i(1-r)^{i-1 \pmod n} \frac{1-(1-r)^{\lceil \frac{i}{n} \rceil n}}{1-(1-r)^n},$$

which can be written in the form

$$F'_+(1) = \sum_{i=1}^{\infty} q_i \left\{ \left\lceil \frac{i}{n} \right\rceil - (1-r)^{i-1 \pmod n} \frac{1-(1-r)^{\lceil \frac{i}{n} \rceil n}}{1-(1-r)^n} \right\}.$$

The substitution of this value into (6) leads to the expression (3), the condition of ergodicity (7) gives (4).

Remark 1. By using the generating function one can compute the mean value of waiting time (measured in cycles). In our case it is equal to

$$\overline{C} = P'(1) = \frac{F''_+(1) - \dfrac{2Q_1(1-r)^n}{(1-r)[1-(1-r)^n]}}{2\left\{ \dfrac{Q_1(1-r)^n}{(1-r)[1-(1-r)^n]} - F'_+(1) \right\}} - \frac{1}{1-(1-r)^n},$$

where

$$F''_+(1) = \sum_{k=2}^{\infty} k(k-1) \sum_{i=(k-1)n+1}^{kn} q_i - \sum_{k=1}^{\infty} \frac{2k}{(1-r)^{kn+1}} \sum_{i=kn+1}^{\infty} q_i (1-r)^i.$$

Remark 2. One can check that in the case of geometrical service time distribution $q_i = q^{i-1}(1-q)$ the above formulas lead to the results of [4].

REFERENCES

1. Lakatos L., Szeidl L., Telek M. Introduction to Queueing Systems with Telecommunication Applications. Springer, 2013.
2. Koba E.V., On a GI/G/1 Queueing System with Repetition of Requests for Service and FCFS Service Discipline // Dopovidi NAN Ukrainy. 2000. (6). P. 101-103. (in Russian)
3. Lakatos L., Efrosinin D. Some Aspects of Waiting Time in Cyclic-Waiting Systems // Communications in Computer and Information Science. 2013. V. 356. P. 115-121.
4. Lakatos L., Efrosinin D. A Discrete Time Probability Model for the Waiting Time of Optical Signals // Communications in Computer and Information Science. 2014. V. 279. P. 114-123.

FORMALIZING SET OF PRE-EMPTION BASED MODELS OF MULTISERVICE 3GPP LTE NETWORKS

K. Samouylov, I. Gudkova, and E. Markova

Peoples' Friendship University of Russia, Moscow, Russia
{ksam, igudkova, emarkova}@sci.pfu.edu.ru

Abstract

Users of 3GPP LTE networks are provided with a wide range of multimedia services with varying QoS requirements; due to this fact a problem of an effective network resources' distribution arises and, consequently, a task of the optimal RAC schemes development. According to the international standards, two types of services are defined within LTE networks – GBR services and non-GBR services. The GBR services generate streaming traffic and non-GBR services – elastic traffic the bit rate of which can dynamically change depending on the cell load. Also, the service priorities differ and are organized with the help of different mechanisms, e.g. service interruption mechanism and mechanism of bit rate degradation. The paper proposes a formal unique description of RAC schemes that is used to developing an example set of models realizing three possible pre-emption based scenarios in multiservice LTE networks.

Keywords: LTE, radio admission control, RAC, guaranteed bit rate, GBR, non-GBR, service interruption, bit rate degradation

1. Introduction

Mobile 4G (4th generation) networks based on LTE (Long Term Evolution) technology [1] provide a wide range of multimedia services with varying quality of service (QoS) requirements. This is one of the most important trends in the modern telecommunication systems and networks modernization. A new classification of multiservice traffic identification has been proposed in LTE networks for reflecting the varying demands of users. According to the 3GPP (3rd Generation Partnership Project) consortium recommendations (TS 36.300, TS 23.401, TS 23.203), nine types of services are singled out in LTE. They are characterized by different service priorities and network resource requirements: guaranteed bit rate (GBR) services and non-GBR services.

The GBR services are real time services, e.g. voice telephony, video telephony, or real time gaming, for which a minimum bit rate value is specified, i.e. guaranteed bit rate. Nevertheless, when free resources are available in the network, the instantaneous bit rate can exceed the minimum specified one, but can not exceed some threshold value known as the maximum bit rate (MBR).

The non-GBR services are those for which no minimum (guaranteed) bit rate value is specified, the instantaneous bit rate can vary depending on the cell

load. It is determined with the help of the so-called aggregate maximum bit rate (AMBR), allowing to differentiate services by service priority levels. So, the bit rate cannot exceed the maximum value specified for user equipments (UE-AMBR) or determined by network characteristics (access point name AMBR, APN-AMBR). Examples of such services include e-mail, web browsing, or interactive gaming.

Various radio resource management (RRM) mechanisms are used to guarantee QoS. One of these mechanisms is the radio admission control (RAC) [2, 4, 5, 6, 7, 8, 9, 10] aimed at admitting or rejecting user requests for service taking into account limited frequency bands and varying QoS requirements. For this purpose different priority levels (allocation and retention priority, ARP) are assigned to multimedia services and define pre-emption RAC algorithms within the corresponding RAC schemes. According to 3GPP TS 23.203: "The range of the ARP priority level is 1 to 9 with 1 as the highest level of priority. The pre-emption capability information defines whether a service data flow can get resources that were already assigned to another service data flow with a lower priority level. The pre-emption vulnerability information defines whether a service data flow can lose the resources assigned to it in order to admit a service data flow with a higher priority level." In accordance with this definition of ARP, it is evident that in case of the lack of resources already occupied by lower priority services a new arrived higher priority request could be accepted, at best, through service bit rate degradation (partial pre-emption) [4, 6, 7, 9, 10], or at worst, through service interruption (full pre-emption) [4, 7, 8]. Other mechanisms such as reservation [5, 6] threshold [5] and probabilistic management [5, 9] could also be applied to realize the priority-service discipline.

Since the number of services LTE users are interested in varies as well as the services themselves within this specified range varies, more than 200 RAC schemes could be set up considering all possible pairwise influences of users on each others due to various pre-emption algorithms. Taking into account this fact and the annual growth of mobile traffic, it seems that the increasing need for the development of a general model considering overall possible priority mechanisms could be transformed into the need for a set of specific models reflecting the predefined required services and mechanisms. Nevertheless, a unique notation is needed to describe at least, firstly each service within a RAC scheme, and secondly, the pre-emption mechanism applied to each pair of services. This is exactly the main purpose of the paper, i.e. to propose a such notation (Section 2) and to illustrate its usage with an example set of specific models of RAC schemes (Section 3).

2. Formal Description of RAC Models

Thereby, each RAC scheme is characterized by the number $K \in \{1, \ldots, 9\}$ of services (Table 1), as well as by the pre-emption scenario that is realized with the use of bit rate degradation and service interruption mechanisms.

Note that for GBR services increasing or decreasing bit rate from a GBR value to a MBR one does not affect the mean service time. At the same time, changing bit rate for non-GBR services results in changing the mean transfer time. So, two types of traffics can be singled out: streaming traffic [11, 12] that is characterized by fixed values of bit rate and time and elastic traffic [2] that is characterized by a fixed value of file size and a variable data transfer time [3]. Traffic can be transferred in two modes – unicast mode or multicast mode. For the last one two disciplines have been analyzed [4]: T1 – multicast session is closed when the first user who has opened the session leaves it, – and T2 – multicast session is closed when the last user leaves the session. T stands for the so-colled "transparent" service discipline in queuing theory.

QCI	Resource type	Priority level	Examples of services	
1		2	Conversational voice	
2	GBR	4	Conversational live streaming video	
3		3	Real time gaming	
4		5	Non-conversational buffered streaming video	
5		1	IMS signalling	
7		7	Voice, live streaming video, interactive gaming	
6	Non-GBR	6	Buffered streaming video, TCP-based applications	for multimedia priority services subscribers
8		8	(e.g., www, e-mail, chat,	for premium subscribers
9		9	FTP, P2P file sharing, progressive video, etc.)	for non-privileged subscribers

Table 1: Characteristics of LTE Service Types (TS 23.203)

Each of nine services could be described with the help of the following parameters:

$$\mathcal{S} = <\text{Priority level, Resource type, Data transfer mode}>.$$

Assume that any RAC scheme is defined by the following mechanisms: bit rate degradation (partial pre-emption) and service interruption (full pre-emption). If k is the pre-emption capable service and k' is the pre-emption vulnerable service, than the pre-emption algorithm could be defined by the fol-

lowing matrix: $\mathbf{P} = (\mathbf{p}(k,k'))_{k,k'=1,\ldots,K} = (d(k,k'), i(k,k'))_{k,k'=1,\ldots,K}$, where

$$d(k,k') = \begin{cases} 1, & \text{if users of service } k' \text{ will be degraded,} \\ & \text{when a request for service } k \text{ arrives,} \\ 0, & \text{otherwise,} \end{cases}$$

$$i(k,k') = \begin{cases} 1, & \text{if users of service } k' \text{ will be interrupted,} \\ & \text{when a request for service } k \text{ arrives,} \\ 0, & \text{otherwise.} \end{cases}$$

3. Example Set of Markov Models of RAC Schemes

Let us illustrate the proposed formal description by an example set of models namely three Markov models (Figure 1) with different traffic types, namely: unicast streaming, multicast streaming, and elastic traffics, as well as some combinations of different pre-emption mechanisms – first of all, bit rate degradation and service interruption.

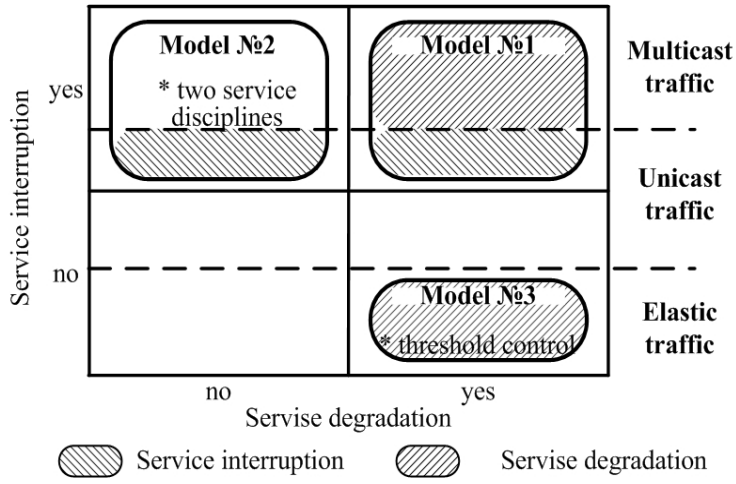


Fig. 1: Set of the RAC schemes

Let us consider an LTE cell having the capacity of C bps. The incoming requests for service generate unicast, multicast, and elastic traffics and arrive as Poisson processes. Resource occupancy times by unicast and multicast traffics as well as elastic file sizes are exponentially distributed. Bit rates are $d = 1$ and b bps for unicast and multicast traffics and e corresponds to the MBR for elastic traffic. The system states of models will be described by vectors with the following components: n – number of unicast users, m – state of a multicast connection ($m = 1$, if at least one user is being served; $m = 0$, otherwise), u – number of transferred elastic files.

3.1. Model of RAC scheme with multicast traffic degradation. Let us consider the first model (Figure 1) [4]. Suppose that users are provided with two services that generate streaming traffic. The service that has a higher priority level ($\mathcal{S}_1$) (e.g. video conference) generates multicast traffic, and the service that has a lower priority level ($\mathcal{S}_2$) (e.g. video on demand) generates unicast traffic. The number of resources allocated for the establishment of a multicast connection is adaptively changed along a certain specified value set of $b_1 > ... > b_k > ... > b_K$ bps and is determined in accordance with the number of free resources – $\max\{b_k,\ k = 1, ..., K:\ b_k \leq C - n\}$. The RAC scheme is realized by two mechanisms of priority admission control – degradation of multicast bit rate and interruption of unicast users. So, the scheme could be defined in the following manner:

$K = 2$, $\mathcal{S}_1$ =<4, GBR, multicast >, $\mathcal{S}_2$ =<5, GBR, unicast>,

P	1	2
1	$(0,0)$	$(0,1)$
2	$(1,0)$	$(0,0)$

The admission control is realized in two stages. The first stage is the degradation of number of resources occupied by multicast traffic down to the minimum value of b_K. The second stage is the interruption of $b_K - (C - dn)$ unicast users in case of the lack of resources. The state of a multicast connection is described as $m_k \in \{0, 1\}$, $k = 1, ..., K$: $m_k = 1$ if the connection is established and occupies b_k bps, $m_k = 0$ if the connection does not occupy b_k bps. The system state space of the corresponding Markov model has the following form

$$\mathcal{X}_1 = \{(\mathbf{m}, n):\quad \mathbf{m} = \mathbf{0},\ n = 0, \ldots, C,\ \ \mathbf{m} = \mathbf{e}_1,\ n = 0, \ldots, C - b_1,$$
$$\mathbf{m} = \mathbf{e}_k,\ n = C - b_{k-1} + 1, \ldots, C - b_k,\ k = 2, \ldots, K\}.$$

3.2. Model of RAC scheme with unicast traffic interruption. Let us consider the second scheme of the set (Figure 1) with two multicast traffic disciplines – T2 (e.g. real time gaming $\mathcal{S}_1$), and T1 (e.g. video conference $\mathcal{S}_2$), – and unicast traffic (e.g. video on demand $\mathcal{S}_3$).

The RAC scheme realizes the mechanism of unicast traffic interruption in case of the lack of resources required for the establishment of a multicast T1 or T2 connection; the admission control is defined by the following services parameters and matrix:

$K = 3$, $\mathcal{S}_1$ =<3, GBR, multicast>,
$\mathcal{S}_2$ =<4, GBR, multicast >, $\mathcal{S}_3$ =<5, GBR, unicast>,

P	1	2	3
1	$(0,0)$	$(0,0)$	$(0,1)$
2	$(0,0)$	$(0,0)$	$(0,1)$
3	$(0,0)$	$(0,0)$	$(0,0)$

Let us denote as $l \in \{0,1\}$ the state of a multicast connection with discipline T2, as b_1 and b_2 bps the number of resources necessary for the establishment of multicast T1 and T2 connections, respectively. Then models state space has the following form

$$\mathcal{X}_2 = \{(l,m,n) \in \{0,1\} \times \{0,1\} \times \{0,1,...,C\} : b_1 m + b_2 l + dn \leq C\}.$$

3.3. Model of RAC scheme with elastic traffic degradation. Unlike the previous sections where we have proposed the models with streaming traffic, let us construct a scheme with two types of services that generate elastic traffic, and with MBR thresholds $e_1 > e_2$ bps [10]. The number of resources occupied by files can dynamically vary from a maximum to a minimum value that is necessary to guarantee requirements for the mean transfer time and is realized through a threshold U of the number of transferred files. In case of low cell load, the maximum number or resources is occupied for files transfer. If the load rises so that it is no longer possible to guarantee a MBR for each transferred file, the instantaneous bit rate is degraded proportionally to an individual MBR, in accordance with some coefficient of bit rate degradation. Because of this, the RAC scheme and corresponding state space of the Markov model have the following forms:

$K = 2$, $\mathcal{S}_1 =<k$, non-GBR, elastic$>$, $\mathcal{S}_2 =<k$, non-GBR, elastic$>$, $k = \overline{6,9}$,

P	1	2
1	$(0,0)$	$(1,0)$
2	$(1,0)$	$(0,0)$

$$\mathcal{X}_4 = \{\mathbf{u} \geq \mathbf{0} : \quad u_1 + u_2 \leq U\}.$$

4. Conclusion

The paper discusses the principle of RAC schemes construction within multiservice networks with pre-emption based admission control. A set of three Markov models of RAC schemes with unicast streaming, multicast streaming, and elastic traffics was proposed. Different scenarios of priority admission control were analyzed, that are based, first and foremost, on the mechanism of bit rate degradation and service interruption of users with lower priority levels. In addition to the results described in the paper a unique input data for numerical experiments is developed for further performance analysis.

REFERENCES

1. Stasiak M., Glabowski M., Wisniewski A., & Zwierzykowski P. Modelling and dimensioning of mobile wireless networks: from GSM to LTE. Willey, 2010.

2. Samouylov K., Gudkova I. Recursive computation for a multi-rate model with elastic traffic and minimum rate guarantees // Proc. of the International Congress on Ultra Modern Telecommunications and Control Systems ICUMT-2010 (October 18–20, 2010, Moscow, Russia). IEEE. 2010. P. 1065-1072.
3. Basharin G., Gaidamaka Yu., & Samouylov K. Mathematical theory of teletraffic and its application to the analysis of multiservice communication of next generation networks // Automatic Control and Computer Sciences. 2013. V. 47, No. 2. P. 62-69.
4. Borodakiy V., Gudkova I., Samouylov K., & Markova E. Modelling and performance analysis of pre-emption based radio admission control scheme for video conferencing over LTE // Proc. of the 6th International Conference ITU Kaleidoscope: Living in a converged world – impossible without standards? K-LCW-2014 (June 3–5, 2014, St. Petersburg, Russia). Switzerland, Geneva, ITU. 2014. P. 53-59.
5. Ghaderi M., Boutaba R. Call admission control in mobile cellular networks: a comprehensive survey // Wireless Communications and Mobile Computing. 2006. V. 6, No. 1. P. 69–93.
6. Gudkova I. A., Samouylov K.E. Modelling a radio admission control scheme for video telephony service in wireless networks // Lecture Notes in Computer Science. 2012. V. 7469. P. 208–215.
7. Khabazian M., Kubbar O., & Hassanein H. A fairness-based pre-emption algorithm for LTE-Advanced // Proc. of the 10th IEEE Global Telecommunications Conference GLOBECOM-2012 (December 3–7, 2012, Anaheim, California, USA). IEEE. 2012. P. 5320-5325.
8. Kwan R., Arnott R., Trivisonno R., & Kubota M. On pre-emption and congestion control for LTE systems // Proc. of the 72nd Vehicular Technology Conference VTC2010-Fall (September 6–9, 2010, Ottawa, Canada). IEEE. 2010. P. 6-9.
9. Liao H., Wang X., & Chen H. H. Adaptive call admission control for multiclass services in wireless networks // Proc. of the IEEE International Conference on Communications ICC-2008 (May 19-23, 2008, Beijing, China). IEEE. 2008. P. 2840–2844.
10. Shorgin S. Y., Samouylov K. E., Gudkova I. A., Markova E. V., & Sopin E. S. Approximating performance measures of radio admission control model for non real-time services with maximum bit rates in LTE // AIP Conference Proceedings. USA, AIP Publishing. 2015. V. 1648. P. 1-4. (DOI 10.1063/1.4912508).
11. Gudkova I., Plaksina O. Performance measures computation for a single link loss network with unicast and multicast traffics // Lecture Notes in Computer Science. 2010. V.6294. P. 256-265.
12. Iversen V. Teletraffic engineering and network planning. Technical University of Denmark, 2011.

A SIMPLIFIED MODEL FOR PERFORMANCE ANALYSIS OF CLOUD COMPUTING SYSTEMS WITH DINAMIC SCALING

Yuliya Gaidamaka, Eduard Sopin, Margarita Talanova
Peoples' Friendship University of Russia, Moscow, Russia

Abstract

Cloud computing paradigm proved itself to be a good approach for efficient high-performance computational infrastructure design. Besides performance, one of the most important measures for cloud service eases performance of the system, but leads to increase in service costs. To avoid ping-pong effect in dynamic scaling, hysteresis approach is applied. Considering number of virtual machines in modern cloud platforms, special attention should be paid to efficiency of computing algorithm. In this work, we describe behavior of cloud computing system in terms of queuing system with threshold-based hysteretic control of active servers number and noninstantaneous server activation. We use state elimination technique to develop efficient algorithm for stationary probabilities calculations and estimation of its performance measures.

УПРОЩЕННАЯ МОДЕЛЬ ДЛЯ АНАЛИЗА ПОКАЗАТЕЛЕЙ ПРОИЗВОДИТЕЛЬНОСТИ СИСТЕМ ОБЛАЧНЫХ ВЫЧИСЛЕНИЙ С ДИНАМИЧЕСКИМ МАСШТАБИРОВАНИЕМ

Ю. Гайдамака, Э. Сопин, М. Таланова
Российский университет дружбы народов, г. Москва, Россия,
{ygaidamaka, esopin}@sci.pfu.edu.ru, matalanova@gmail.com

Аннотация

Cloud computing paradigm proved itself to be a good approach for efficient high-performance computational infrastructure design. Besides performance, one of the most important measures for cloud service providers is energy efficiency of the system. For better energy efficiency, in case of light load on the system additional virtual machines are switched off, and will be switched on again when load increases. Dynamic scaling increases performance of the system, but leads to increase in service costs. To avoid ping-pong effect in dynamic scaling, hysteresis approach is applied. Considering number of virtual machines in modern cloud platforms, special attention should be paid to efficiency of computing algorithm. In this work, we describe behavior of cloud computing system

in terms of queuing system with threshold-based hysteretic control of active servers number and noninstantaneous server activation. We use state elimination technique to develop efficient algorithm for stationary probabilities calculations and estimation of its performance measures.

Ключевые слова: cloud computing, hysteretic load control, dynamic scaling, state elimination method.

1. Introduction

Cloud computing grants to a user computing resources and infrastructure in the form of Internet service [1]. Cloud computing systems are applied to data storage and processing, to distributed computing for scientific and business solutions. Modern cloud systems are usually designed scalable so that the system can cope with high load and may reduce energy consumption during the periods of low load. One of the ways for system scalability realization is dynamic activation-deactivation of servers and virtual machines [2, 3, 4]. Analysis of such systems can be performed using queuing systems with threshold based control of active servers number [5, 6]. But the main problem of cloud systems modeling in terms of teletraffic theory is high computing complexity of derived algorithms [7, 8, 9]. In [5], a queuing system with hysteretic dynamic scaling and noninstantaneous additional server activation is applied for video-on-demand service, stationary characteristics of system are obtained by means of matrix-geometric methods which are not applicable for cloud computing because of extremely large amount of servers in modern cloud platforms. In [3, 4], cloud computing system with dynamic scaling was described in terms of queuing system with hysteretic control of active servers number, and for a case of three servers, the effective computing algorithm for stationary probability distribution was developed. Complexity of received algorithm is linear because of state elimination technique [11] used for queuing system analysis. In [10], the same approach was applied for any number of servers in the system. The weak side of proposed in [10] model is extremely large state space, which has quadratic dependence on number of servers. Moreover, the majority of states has a negligible effect on overall system behavior. Therefore, in this article, we consider a simplified model with reduced state space and perform its analysis using the same state elimination approach. Finally, in numerical analysis section we provide a comparative analysis of the initial and simplified models.

2. Simplified model description

For performance measures evaluation we describe cloud computing system dynamic adding and removing additional servers in terms of queuing system with K servers and hysteretic control of active servers number (figure 1). An arriving customer enter the system if the total amount of customers is less than system capacity R, otherwise the customer is considered lost. Customers

arrive according to the Poisson process with rate λ. We consider servers to be uniform, the service times are exponentially distributed with rate μ.

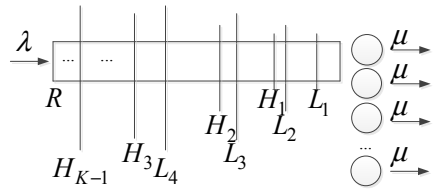


Рис. 1: Queuing model with thresholds $\vec{H}$ and $\vec{L}$.

In the empty system, there is only one active server. Server activation/deactivation procedures are governed by number of customers in the system upper threshold vector $\vec{H} = (H_1, H_2, ..., H_{K-1})$, $H_1 < H_2 < ... < H_{K-1})$ and lower threshold vector $\vec{L} = (L_1, L_2, ..., L_{K-1})$, $L_1 < L_2 < ... < L_{K-1})$ where $L_{k+1} < H_k$, $k = \overline{1, K-2}$ and $L_k < H_k$, $k = \overline{1, K-1}$. Customers are served in FCFS (First Come First Served) order. System functions as follows:

- if there is H_k customers in the system, arriving customer causes an additional $(k+1)$th server activation procedure, server activation time is not instant and is exponentially distributed with rate α;
- if there is L_k customers, then on departure of a customer either server shuts down in case of no running server activation procedure or one of server activation procedures stops.

For the simplification of initial model [10] we assume that in case of two running server activation procedures, third server activation procedure is comleted immediately on start. Thus, it is impossible to have 3 running activation procedures at once. Provided assumption significantly reduces state space of the system for big values of K.

Functioning of system is described by Markov process with a set of states

$$S = (k, i, n) = \begin{cases} 0 \le n \le H_1 & \text{for } k = 1, i = 1, \\ L_{k-1} \le n \le H_k & \text{for } k = 2, i = \overline{k-1, k}, \\ L_{k-1} \le n \le H_k & \text{for } k = \overline{3, K-1}, i = \overline{k-2, k}, \\ L_{k-1} \le n \le R & \text{for } k = K, i = \overline{k-2, k}, \end{cases}$$

where k is necessary number of servers; i is number of the active servers; n is number of customers in the system.

For the chosen arrangement of threshold values, figure 2 depicts state transition diagram for the system with $K = 5$ servers. Let us denote state levels k, i.

3. Stationary probability distribution calculation algorithm

For calculation of stationary probabilities, the method of state elimination is used. We introduce several auxilary probabilities and intensity, similar to [4]:

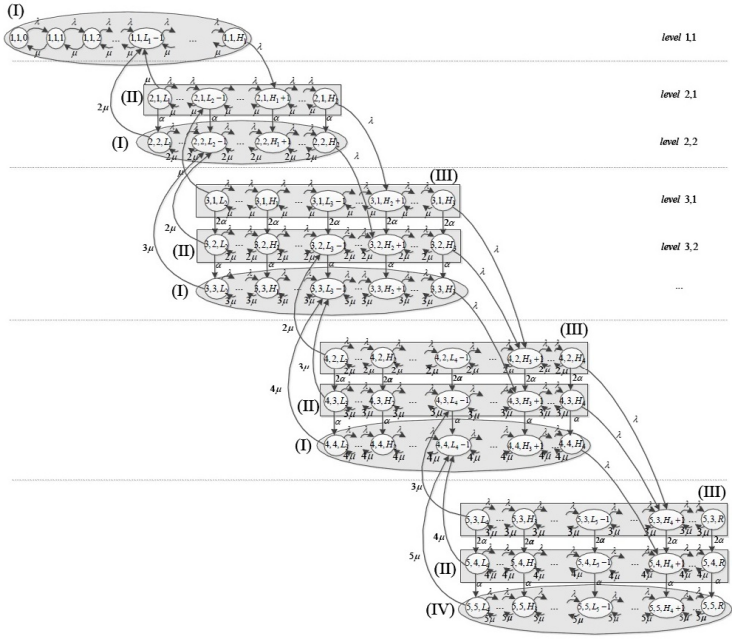


Рис. 2: State transition diagram for the system with $K = 5$ devices.

- $a_{k,i,n}$ is probability that starting from state (k, i, n) the system will pass to $(k, i, n-1)$ earlier, than to states on levels $k+1, i$ or $k, i+1$,
- $a^*_{k,i}$ is probability that starting from state (k, i, L_k) the system will pass to $(k, i, L_k - 1)$ earlier, than to state $(k, i+1, L_k - 1)$,
- $\Lambda_{k,i,n}$ is intensity of direct transition from states of level $k, i-1$ to a state (k, i, n) considering that states $(k, i, n+1)$, $(k, i, n+2)$, ... are eliminated,
- $A_{k,i,n}$ is probability that starting from state $(k, i, H_{k-1} + 1)$ the system will pass to (k, i, n) earlier, than to states of levels $k+1, i$ or $k, i+1$,
- $A^*_{k,i,n}$ is probability that starting from state $(k, i, H_{k-1} + 1)$ the system will pass to $(k, i, L_k - 1)$ earlier, than to state $k, i+1, L_k - 1$,
- $C_{k,i}$ is probability that starting from a state $(k+1, i, H_k + 1)$ the system will pass to $(k, i, L_k - 1)$ earlier, than to states of levels $k, i+1$.

For these auxilary probabilities and intensities recurrent relations (2)-(8), (11)-(24), (26)-(30), (32) hold true.

Stationary probabilities for states (k, i, n) for $k = \overline{1, K-1}, i = k$ are calculated by formula (1).

$$\pi_{k,i,n} = \frac{\Lambda_{k,i,n} + \lambda \pi_{k,i,n-1}}{i\mu + \lambda(1 - a_{k,i,n+1})}, L_{k-1} \le n \le H_k, \tag{1}$$

where threshold value $L_0 := 0$, initial probabilities are

$$\begin{cases} \pi_{k,i,L_{k-1}-1} := 1 \quad \text{for } k = 1, \\ \pi_{k,i,L_{k-1}-1} := 0 \quad \text{for } k = \overline{2, K-1}. \end{cases}$$

The auxiliary probabilities and intensities used in (1) are calculated using recurrent formulas (2)-(8).

$$a_{k,i,H_k+1} := 0; \tag{2}$$

$$a_{k,i,n} = \frac{i\mu}{i\mu + \lambda(1 - a_{k,i,n+1})}, L_k + 1 \leq n \leq H_k; \tag{3}$$

$$a_{k,i,n+1} := 1, L_{k-1} \leq n \leq L_k - 1, \tag{4}$$

$$\Lambda_{k,i,H_k+1} := 0; \tag{5}$$

$$\Lambda_{k,1,n} := 0, L_{k-1} \leq n \leq H_k; \tag{6}$$

$$\Lambda_{k,i,n} = \alpha\pi_{k,i-1,n} + a_{k,i,n+1}\Lambda_{k,i,n+1}, L_k \leq n \leq H_k, \tag{7}$$

$$\Lambda_{k,i,n} = \alpha \sum_{v=n}^{H_k} \pi_{k,i-1,v} + \lambda(1 - C_{k,i-1})\pi_{k,i-1,H_k}, L_{k-1} \leq n \leq L_k - 1, \tag{8}$$

Stationary probabilities for states (k, i, n) for $k = \overline{2, K-1}, i = k - 1$ are calculated by formulas (9) and (10).

$$\pi_{k,i,n} = \frac{\Lambda_{k,i,n} + \lambda\pi_{k,i,n-1} + \lambda A_{k,i,n}\pi_{k-1,i,H_{k-1}}}{i\mu + \alpha + \lambda(1 - a_{k,i,n+1})}, L_{k-1} \leq n \leq L_k - 2, L_k \leq n \leq H_k, \tag{9}$$

$$\pi_{k,i,n} = \frac{\Lambda_{k,i,n} + \lambda\pi_{k,i,n-1} + \lambda A^*_{k,i}\pi_{k-1,i,H_{k-1}}}{i\mu + \alpha + \lambda(1 - a^*_{k,i})}, n = L_k - 1, \tag{10}$$

where initial probabilities are $\pi_{k,i,L_{k-1}-1} := 0$.

The auxiliary probabilities and intensity used in formulas (9) and (10) are calculated using recurrent relations (11)-(24).

$$a_{k,i,H_k+1} := 0; \tag{11}$$

$$a_{k,i,n} = \frac{i\mu}{i\mu + \alpha + \lambda(1 - a_{k,i,n+1})}, L_{k-1} + 1 \leq n \leq L_k - 2, L_k \leq n \leq H_k; \tag{12}$$

$$a_{k,i,L_k-1} = \frac{i\mu}{i\mu + \alpha + \lambda(1 - a^*_{k,i})}; \tag{13}$$

$$a^*_{k,i} = a_{k,i,L_k} + (1 - a_{k,i,L_k})\frac{\lambda C_{k,i}}{\lambda + \alpha}, \tag{14}$$

$$\Lambda_{k,i,H_k+1} := 0; \tag{15}$$

$$\Lambda_{k,1,n} := 0, L_{k-1} \le n \le H_k; \tag{16}$$

$$\Lambda_{k,i,n} = 2\alpha\pi_{k,i-1,n} + a_{k,i,n+1}\Lambda_{k,i,n+1}, L_k \le n \le H_k; \tag{17}$$

$$\Lambda_{k,i,n} = 2\alpha\sum_{v=n}^{H_k}\pi_{k,i-1,v} + \lambda(1 - C_{k,i-1})\pi_{k,i-1,H_k}, L_{k-1} \le n \le L_k - 1 \tag{18}$$

$$C_{k,i} = \frac{i\mu}{i\mu + \alpha + \lambda(1 - a_{k+1,i,L_k+1})}\prod_{v=L_k+1}^{H_k+1} a_{k+1,i,v}, \tag{19}$$

$$A_{k,i,H_{k-1}+1} := 1; \tag{20}$$

$$A_{k,i,n} := 0, H_{k-1} + 2 \le n \le H_k; \tag{21}$$

$$A_{k,i,n} = \prod_{v=n+1}^{H_{k-1}+1} a_{k,i,v}, L_k - 1 \le n \le H_{k-1}; \tag{22}$$

$$A^*_{k,i} = A_{k,i,L_k-1} + (1 - A_{k,i,L_k-1})\frac{\lambda C_{k,i}}{\lambda + \alpha}; \tag{23}$$

$$A_{k,i,n} = A^*_{k,i}\prod_{v=n+1}^{L_k-1} a_{k,i,v}, L_{k-1} \le n \le L_k - 2. \tag{24}$$

Stationary probabilities for states (k, i, n), $k = \overline{3, K}, i = k-2$ are calculated by formula (25).

$$\pi_{k,i,n} = \frac{\lambda\pi_{k,i,n-1} + \lambda A_{k,i,n}(\pi_{k-1,i,H_{k-1}} + \pi_{k-1,i-1,H_{k-1}})}{i\mu + 2\alpha + \lambda(1 - a_{k,i,n+1})}, L_{k-1} \le n \le H_k, \tag{25}$$

where threshold value $H_K := R$ probability $\pi_{2,0,H_2} := 0$, initial probabilities are $\pi_{k,i,L_{k-1}-1} := 0$.

The auxiliary probabilities and intensity used in a formula (25) are calculated by recurrent relations (26)-(31).

$$a_{k,i,H_k+1} := 0, k \leq K-1; \tag{26}$$

$$a_{K,i,R+1} := 1, \tag{27}$$

$$a_{k,i,n} = \frac{i\mu}{i\mu + 2\alpha + \lambda(1 - a_{k,i,n+1})}, L_{k-1} + 1 \leq n \leq H_k, \tag{28}$$

$$A_{k,i,H_{k-1}+1} := 1; \tag{29}$$

$$A_{k,i,n} := 0, H_{k-1} + 2 \leq n \leq H_k; \tag{30}$$

$$A_{k,i,n} = \prod_{v=n+1}^{H_{k-1}+1} a_{k,i,v}, L_{k-1} \leq n \leq H_{k-1}. \tag{31}$$

$$C_{k,i} = \frac{i\mu}{i\mu + 2\alpha + \lambda(1 - a_{k+1,i+1,L_k+1})} \prod_{v=L_k+1}^{H_k+1} a_{k+1,i+1,v}, \tag{32}$$

Stationary probabilities for states (k, i, n) for $k = K, i = \{k, k-1\}$ are calculated on a formula (33).

$$\pi_{k,i,n} = \frac{\Lambda_{k,i,n} + \lambda\pi_{k,i,n-1} + \lambda A_{K,i,n}\pi_{K-1,i,H_{K-1}}}{i\mu + (K-i)\alpha + \lambda(1 - aK, i, n+1)}, L_{k-1} \leq n \leq R, \tag{33}$$

where initial probabilities are $\pi_{k,i,L_{k-1}-1} := 0$.

The auxiliary probabilities and intensities used in formula (33) are calculated by recurrent relations (34) - (38).

$$a_{K,i,R+1} := 1, \tag{34}$$

$$a_{K,i,n} = \frac{i\mu}{i\mu + (k-i)\alpha + \lambda(1 - a_{k,i,n+1})}, L_{K-1} + 1 \leq n \leq R, \tag{35}$$

$$a_{K,K,N} := 0, L_{K-1} + 1 \leq n \leq R, \tag{36}$$

$$A_{K,K-1,n} = \prod_{v=n+1}^{H_{K-1}+1} a_{K,K-1,v}, L_{k-1} \leq n \leq H_{k-1}. \tag{37}$$

$$\Lambda_{k,i,n} = \alpha \sum_{v=n}^{R} \pi_{k,i-1,v}, L_{k-1} \leq n \leq R, \tag{38}$$

To recieve stationary probabilities let $\widetilde{\pi}_{1,1,0} = 1$. Then, using formulas (1), (9), (10), (25), (33) we will calculate nonnormalized probabilities $\widetilde{\pi}_{k,i,n}$, $(k,i,n) \in S$ and the normalizing constant $G = \sum_{(k,i,n)\in S} \widetilde{\pi}_{k,i,n}$. Then stationary probabilities are $\pi_{k,i,n} = \frac{\widetilde{\pi}_{k,i,n}}{G}$, $(k,i,n) \in S$.

4. Stationary characteristics of queueing system

Knowing stationary probability distribution of the system, it is possible to estimate some probabilistic and time characteristics of the considered system, formulas are given below.

Average number of customers in the system:
$N = \sum_{(k,i,n)\in S} n\pi_{k,i,n}$.
Average number of required servers:
$MD = \sum_{k=1}^{K} k \sum_{i=1}^{K} \sum_{n=1}^{R} \pi_{k,i,n}$.
Average number of busy servers:
$MF = \sum_{i=1}^{K} i \sum_{k=1}^{K} \sum_{n=1}^{R} \pi_{k,i,n}$.
Blocking probability:
$\pi = \sum_{i=1}^{K} \pi_{K,i,R}$.
Average sojourne time:
$T = \frac{N}{\lambda(1-\pi)}$.

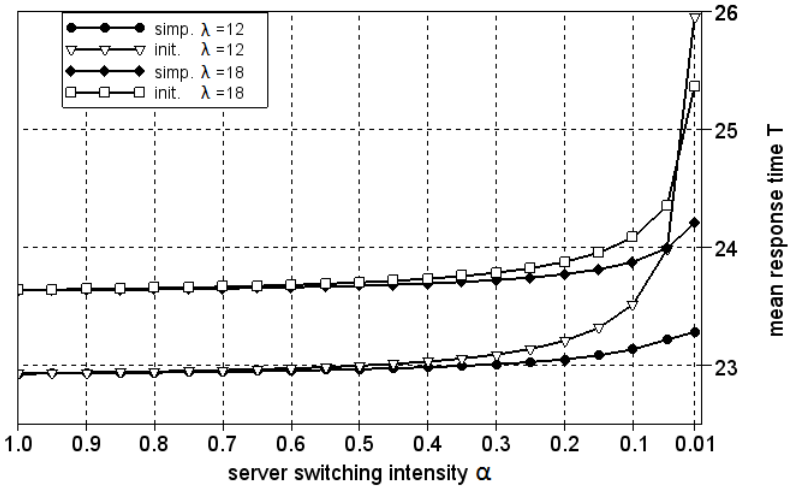


Рис. 3: Mean response time comparison for thresholds arrangement (39)

Note that three last characteristics correspond to performance measures of cloud computing system with dynamic scaling, namely to average number of the activated virtual machines, blocking probability of user request and average response time of the system. The average number of the activated virtual machines MF can serve for assessment of energy efficiency of cloud computing system at known energy consumption on maintenance of the virtual computer in an active state.

5. Numerical analysis

In this section we perform a comparative analysis of initial system, proposed in [10] and simplified model, described above. We compare two main performance characteristics for both models, i.e. mean response time and blocking probability. We provide calculations for the system with $K = 20$ servers, serving rate $\mu = 1$ and two different arrival rate intensities $\lambda = \{12, 18\}$ for various values of server switching rate α, which is key parameter that makes behavior of the systems differ from each other.

We used two threshold arrangements for numerical analysis:

$$\begin{cases} H_i = 25i, \quad i = \overline{1, K-1}, R = 500; \\ L_i = H_{i-1} - 1, \quad i = \overline{2, K-1}, L_1 = 12; \end{cases} \tag{39}$$

and

$$\begin{cases} H_i = 50i, \quad i = \overline{1, K-1}, R = 1000; \\ L_i = H_{i-1} - 1, \quad i = \overline{2, K-1}, L_1 = 12. \end{cases} \tag{40}$$

Figures 3 and 4 show mean response time of both models for thresholds arrangements (39) and (40) respectively. As it is seen of figure 3, simplified and initial model have almost the same response time for $\frac{\mu}{\alpha} \approx 1$. With the growth of $\frac{\mu}{\alpha}$ ratio to 10, relative error of the simplified model increases to approximately 2% for low load and 1% for high load case and reaches 10% and 5% when $\frac{\mu}{\alpha} = 100$ for light and heavy loads respectively. Note that simplified model show lower mean response time because with low additional server activation rate we assumed activation to complete immediately in some cases.

With wider inter-threshold range arrangement (40), relative error of the simplified significantly decreases (figure 4). It becomes nearly 0.2% in $\frac{\mu}{\alpha} = 10$ case and 3% in $\frac{\mu}{\alpha} = 100$ case.

Figures 5 and 6 show blocking probability of both models for thresholds arrangement (39) and (40) respectively. Figures show that in light load case, blocking probability relative error of the simplified model is extremely high even for $\frac{\mu}{\alpha} = 5$, but it may be considered feasible in heavy load case. The reason of that is more frequent immediate server activation, which increases system utilization. However, in heavy load cases system shows almost maximum performance already, and additional factors do not have significant effect.

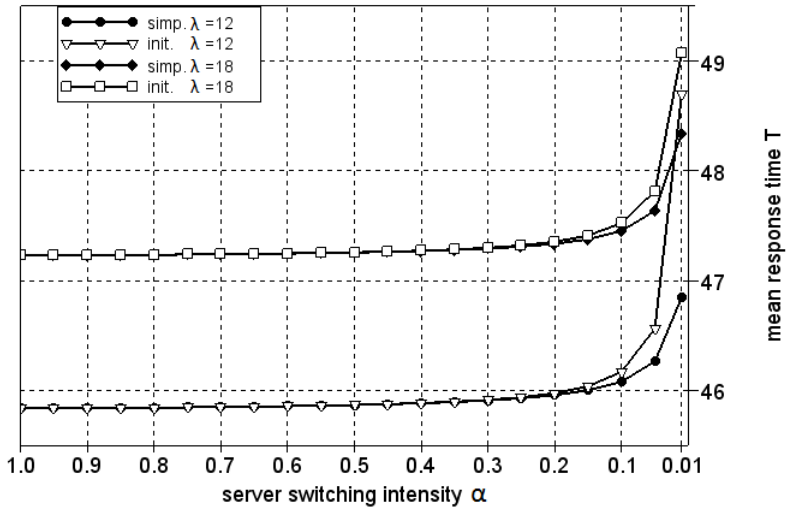


Рис. 4: Mean response time comparison for thresholds arrangement (40)

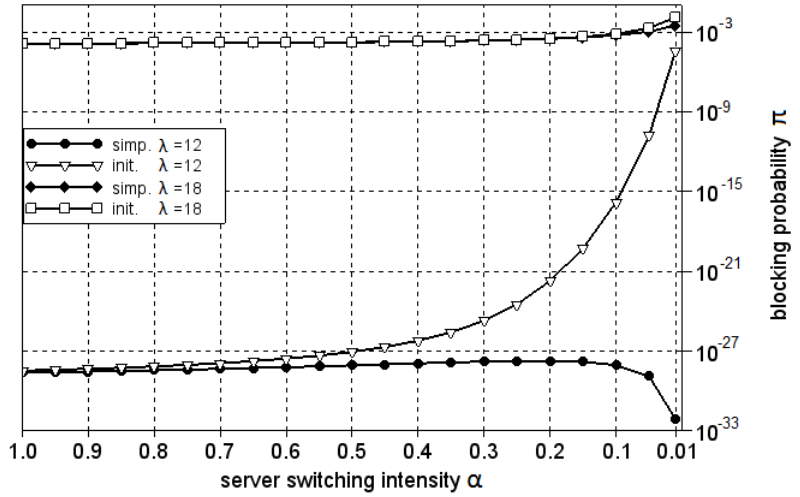


Рис. 5: Blocking probability comparison for thresholds arrangement (39)

6. Conclusion

In this paper, we proposed the simplified model for performance analysis of cloud computing system with dynamic server activation. For the considered model, we developed an efficient stationary probabilities computing algorithm based on state elimination technique. Simplificaiton of the model significantly

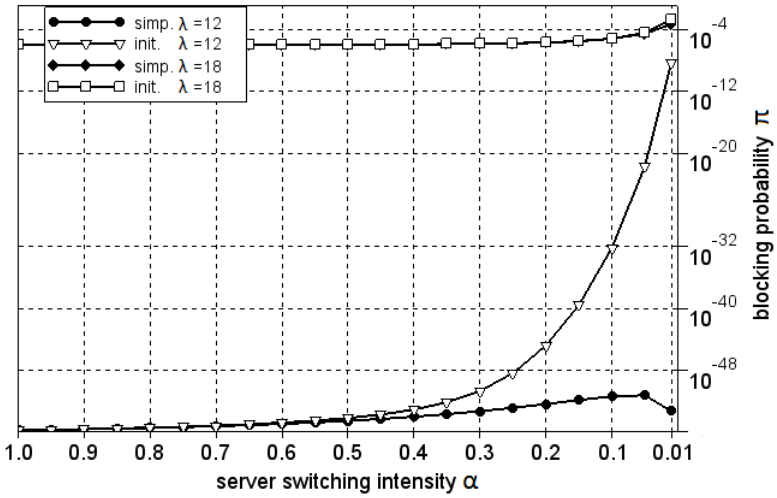


Рис. 6: Blocking probability comparison for thresholds arrangement (40)

decreases state space complexity of corresponding Markov process and, consequently, computing complexity of the algorithm.

Our analysis showed that the simplified system gives good precision of the mean response time for a variety of system and load parameters. Moreover, increasing distance between corresponding lower and upper thresholds, improves the degree of calculation accuracy. However, blocking probability precision can be considered feasible only in heavy load cases.

Another way to increase computation accuracy of the simplified model is to increase maximum simultaneously running server activation procedures to three or four and it will be done if our further research.

7. Acknowledgments

This work was supported in part by the Russian Foundation for Basic Research, projects No. 14-07-00090, 15-07-03051, 15-07-03608. Authors thank master student of the Applied Probability and Informatics Department Vasiliev I. for participation in receiving numerical results.

ЛитеРатуРа

1. ETSI Cloud Standards Coordination. Final Report 2013, ver. 1.0, `http://www.etsi.org/images/files/Events/2013/2013_CSC_Delivery_WS/CSC-Final_report-013-CSC_Final_report_v1_0_PDF_format-.PDF`
2. Goswami V., Patra S. S., Mund G. B. Performance Analysis of Cloud with Queue-Dependent Virtual Machines // Proc. of 1st Internationall Conf.

on Recent Advances in Information Technology, 2012, Dhanbad, India. P. 357-362.
3. Pechinkin A., Gaidamaka, Yu., Sopin E., Talanova M. Performance analysis of cloud computing systems with dynamic scaling // Collection of the chosen works IX of the International annual scientific and practical conference "Modern Information Technologies and IT Education, 2014, Moscow. Pp. 395-406.
4. Shorgin S. Y., Pechinkin A. V., Samouylov K. E., Gaidamaka Y. V., Gudkova I. A., Sopin E. S. Threshold-based Queuing System for Performance Analysis of Cloud Computing System with Dynamic Scaling // Proc. of the 12th International Conference of Numerical Analysis and Applied Mathematics ICNAAM-2014, Rhodes, Greece, 2014, USA, AIP Publishing, 2015, Vol. 1648. P. 1-3.
5. Golubchik L., Lui John C. S. Bounding of Performance Measures for Threshold-Based Queuing Systems: Theory and Application to Dynamic Resource Management in Video-on-Demand Servers // IEEE Trans. Computers, vol. 51, no. 4. P. 353-372.
6. Gaidamaka Yu. V., Pechinkin A. V., Razumchik R.V., Samuylov A. K., Samouylov K. E., Sokolov I. A., Sopin E. S., Shorgin S.. Ya. The distribution of the return time from the set of overload states to the set of normal load states in a system M|M|1|<L,H>|<H,R>with hysteretic load control // Informatics and its applications, vol. 7, no. 4, 2013. Pp.20-33.
7. Basharin G. P., Gaidamaka Yu. V. , Samouylov K. E. Mathematical Theory of Teletraffic and Its Application to the Analysis of Multiservice Communication of Next Generation Networks // Automatic Control and Computer Sciences, 2013, Vol. 47, No. 2, pp. 62–69..
8. Mokrov E. V, Samouylov K. E. Modeling of cloud computing as a queuing system with batch arrivals, T-Comm – Telecommunication and Transport, No. 11, 2013. Pp. 139-141.
9. Mokrov E. V., Chukarin A. V. Performance analysis of cloud computing system with live migration, T-Comm – Telecommunication and Transport,No.8, 2014. Pp. 64-67.
10. Gaidamaka Yu. V., Sopin E. S.,Talanova M. O. Performance measures analysis of cloud computing systems with hysteretic control // T-Comm - Telecommunication and Transport, in print.
11. Bocharov P. P., D'Apice C., Pechinkin A. V., Salerno S. Queueing Theory, Ultrecht, Boston: VSP Publishing, 2004.

$MMAP/M_2/N/\infty$ system with preemptive priority and servers reservation

A. Dudin, S. Dudin, O. Dudina
Belarusian State University, 4, Nezavisimosti Ave., Minsk, 220030, Belarus

A multi-server queueing system with an infinite buffer, two types of customers and servers reservation is considered. The flow of customers is described by the marked Markovian arrival process ($MMAP$). Type 1 customers have preemptive priority over type 2 customers. The ergodicity condition of the system is derived. The main performance measures are calculated.

СИСТЕМА $MMAP/M_2/N/\infty$ С АБСОЛЮТНЫМ ПРИОРИТЕТОМ И РЕЗЕРВИРОВАНИЕМ ПРИБОРОВ

А. Дудин, С. Дудин, О. Дудина
Белорусский государственный университет, Минск, Беларусь,
dudin@bsu.by, dudin85@mail.ru, dudina_olga@email.com

Аннотация

Исследуется система обслуживания с бесконечным буфером и двумя типами запросов. Входной поток запросов моделируемся с помощью маркированного марковского входного потока. Запросы первого типа имеют абсолютный приоритет над запросами второго типа. Получено условие эргодичности. Найдены основные характеристики производительности.

Ключевые слова: маркированный марковский входной поток, абсолютный приоритет

1. Введение

Важным разделом теории массового обслуживания является теория приоритетных систем обслуживания. В таких системах запросы разделены на несколько классов, например, в порядке убывания их важности (экономической, социальной и т.д.) для системы, и пользователи различных классов требуют различного обслуживания в системе. Запросы из классов с более высоким приоритетом имеют преимущества перед другими в отношении доступа к свободным приборам, если они имеются, или выбору запросов из очереди для обслуживания.

Существуют разные виды приоритетов, предоставляемых важным запросам. Наиболее известными являются относительный и абсолютный приоритет. Относительный приоритет предполагает, что прибывающий высокоприоритетный запрос не может прерывать обслуживание, предоставляемое низкоприоритетному запросу. Абсолютный приоритет предполагает, что поступающий высокоприоритетный запрос прерывает обслуживание, предоставляемое запросу с низким приоритетом.

В статье мы рассматриваем многолинейную систему с абсолютным приоритетом. Очевидно, что абсолютный приоритет является лучшим по сравнению с относительным приоритетом с точки зрения приоритетных запросов, но он является худшим вариантом с точки зрения запросов с низким приоритетом и с точки зрения использования системных ресурсов. В качестве компромисса может быть использована дисциплина с относительным приоритетом в сочетании с резервированием некоторых приборов исключительно для обслуживания приоритетных запросов. Системы массового обслуживания с относительным приоритетом и резервированием некоторых приборов были рассмотрены, например, в [1, 2]. Целью комбинации резервирования с *относительным* приоритетом в [1, 2] является желание предоставить больше преимуществ приоритетным запросам по сравнению с обычным относительным приоритетом. В данной работе мы предлагаем сочетание резервирования с *абсолютным* приоритетом. Так как абсолютный приоритет и так дает очень существенное преимущество для *приоритетных* запросов, они не нуждаются в дополнительном резервировании приборов, поэтому, мотивацией рассматриваемой дисциплины является улучшение качества обслуживания *низкоприоритетных* запросов путем уменьшения частоты прерываний их обслуживания.

Результаты анализа, представленного в данной статье, могут быть использованы для улучшения работы многих реальных систем, например, различных технических, производственных, обслуживающих систем, в которых, во избежание возможного простоя и для увеличения прибыли, приборы обеспечивают обслуживание некоторых фоновых или внешних клиентов в моменты отсутствия первичных запросов. Примерами таких систем могут служить системы когнитивного радио, см., например, [3, 4]

2. Математическая модель

Рассматривается N-линейная система массового обслуживания с бесконечным буфером и двумя типами запросов. Запросы поступают в соответствии с маркированным марковским входным потоком ($MMAP$-потоком) запросов, заданным неприводимой цепью Маркова ν_t, $t \geq 0$, с непрерывным временем и конечным пространством состояний $\{0, ..., W\}$. Время пребывания цепи в состоянии ν экспоненциально распределено с положительным параметром λ_ν. Когда время пребывания в состоянии ν истекло, с вероятностью $p^{(0)}_{\nu,\nu'}$ процесс ν_t переходит в состояние ν' без генерации за-

просов, $\nu \neq \nu'$, и с вероятностью $p_{\nu,\nu'}^{(r)}$ процесс ν_t переходит в состояние ν' с генерацией запроса r-го типа, $r = 1,\ 2$, $\nu, \nu' = \overline{0, W}$.

Поведение $MMAP$-потока полностью характеризуется матрицами D_0, $D_r,\ r = 1, 2$, которые определяются следующим образом: $(D_r)_{\nu,\nu'} = \lambda_\nu p_{\nu,\nu'}^{(r)}$, $\nu, \nu' = \overline{0, W}$, $r = 1, 2$, и $(D_0)_{\nu,\nu} = -\lambda_\nu, \nu = \overline{0, W},\ (D_0)_{\nu,\nu'} = \lambda_\nu p_{\nu,\nu'}^{(0)},\ \nu, \nu' = \overline{0, W}, \nu \neq \nu'$. Матрица $D(1) = D_0 + D_1 + D_2$ представляет собой инфинитезимальный генератор цепи $\nu_t, t \geq 0$. Средняя интенсивность поступления запросов λ имеет вид $\lambda = \boldsymbol{\theta}(D_1 + D_2)\mathbf{e}$, где $\boldsymbol{\theta}$ – вектор стационарного распределения цепи Маркова $\nu_t, t \geq 0$. Вектор $\boldsymbol{\theta}$ является единственным решением системы линейных алгебраических уравнений $\boldsymbol{\theta} D(1) = \mathbf{0},\ \boldsymbol{\theta}\mathbf{e} = 1$. Здесь и далее $\mathbf{e}$ – вектор-столбец, состоящий из единиц, $\mathbf{0}$ – вектор-строка, состоящая из нулей. Средняя интенсивность λ_r поступления запросов r-го типа определяется формулой $\lambda_r = \boldsymbol{\theta}\, D_r\mathbf{e}$, $r = 1,\ 2$.

Мы предполагаем, что запросы первого типа имеют абсолютный приоритет. Если в момент поступления запроса первого типа есть свободный прибор, запрос немедленно начинает обслуживание. Если в момент прихода запроса первого типа все приборы заняты и имеются запросы второго типа, получающие обслуживание, обслуживание одного запроса второго типа прекращается, а запрос первого типа занимает освободившийся прибор. Запрос второго типа, обслуживание которого прекратилось, возвращается в очередь с вероятностью p или уходит из системы с дополнительной вероятностью $1 - p$. Если в момент прихода запроса первого типа все приборы заняты запросами первого типа, то этот запрос покидает систему навсегда.

Мы предполагаем, что зафиксирован некоторый параметр (порог) M, $0 < M \leq N$. Запросы второго типа принимаются на обслуживание, если число занятых приборов меньше M. Если в момент прибытия произвольного запроса второго типа число занятых приборы больше, чем $M-1$, то этот запрос идет в буфер с вероятностью q, а с дополнительной вероятностью покидает систему.

Время обслуживания запросов типа r имеет экспоненциальное распределение с параметром $\mu_r,\ r = 1, 2$.

3. Процесс изменения состояний системы

Пусть

i_t, $i_t \geq 0$, – число запросов второго типа в буфере,

n_t, $n_t = \overline{(1 - \delta_{i_t,0})M, N}$, – число занятых приборов,

l_t, $l_t = \overline{0, \min\{n_t, M\}}$, – число запросов второго типа на обслуживании,

ν_t, $\nu_t = \overline{0, W}$, – состояние управляющего процесса $MMAP$ в момент времени $t,\ t \geq 0$. Здесь $\delta_{i_t,0}$ – это кронекерова дельта.

Поведение изучаемой системы описывается в терминах регулярной неприводимой цепи Маркова с непрерывным временем $\xi_t = \{i_t, n_t, l_t, \nu_t\}$, $t \geq 0$.

Введем следующие обозначения:

- $\bar{W} = W + 1$;
- I – единичная матрица, O – нулевая матрица соответствующего размера. Если размерность матрицы или вектора не ясна из контекста, она указывается как нижний индекс;
- $\oplus$ и $\otimes$ обозначают кронекеровы сумму и произведение матриц, соответственно, смотри, например, [5];
- $\mathrm{diag}\{A_1, \ldots, A_l\}$ – диагональная матрица с диагональными элементами или блоками $A_1, \ldots, A_l$, $\mathrm{diag}^-\{A_1, \ldots, A_l\}$ – матрица с поддиагональными элементами или блоками $A_1, \ldots, A_l$, $\mathrm{diag}^+\{A_1, \ldots, A_l\}$ – матрица с наддиагональными элементами или блоками $A_1, \ldots, A_l$;
- $C_l = \mathrm{diag}\{0, 1, \ldots, l\}, \bar{C}_l = \mathrm{diag}\{l, l-1, \ldots, 0\}$, $l = \overline{0, M}$;
- $\tilde{C}_l = \mathrm{diag}\{l, l-1, \ldots, l-M+1, l-M\}$, $l = \overline{M, N}$;
- E_n^+, $\hat{E}_n^+$, $n = \overline{0, M-1}$, – матрицы размера $(n+1) \times (n+2)$ со всеми нулевыми элементами, кроме элементов $(E_n^+)_{l,l+1}$, $l = \overline{0, n+1}$, и $(\hat{E}_n^+)_{l,l}$, $l = \overline{0, n+1}$, которые равны 1;
- $E_n^-, \hat{E}_n^-$, $n = \overline{1, M}$, – матрицы размера $(n+1) \times n$ со всеми нулевыми элементами, кроме элементов $(E_n^-)_{l,l}$, $l = \overline{0, n}$, и $(\hat{E}_n^-)_{l,l-1}$, $l = \overline{1, n+1}$, которые равны 1;
- E^- – квадратная матрицы размера $M+1$ со всеми нулевыми элементами, кроме элементов $(E^-)_{l,l-1}$, $l = \overline{1, M}$, которые равны 1;
- $\mathcal{C}_n = \mu_1 \tilde{C}_n + \mu_2 C_M$, $\mathcal{C}_n^- = \mu_1 \tilde{C}_n + \mu_2 C_M E^-$, $n = \overline{M+1, N}$;
- $\hat{I}_l$, $l = \overline{M+1, N-M+1}$, – квадратная матрицы размера l со всеми нулевыми элементами, кроме элемента $(\hat{I}_l)_{0,0}$, который равен 1.

Перенумеруем состояния цепи ξ_t в лексикографическом порядке компонент (i, n, l, ν). Множество состояний, имеющих значение (i, n) двух первых компонент цепи, будем называть макросостоянием (i, n).

Пусть Q – генератор цепи Маркова ξ_t, $t \geq 0$, сформированный из блоков $Q_{i,j}$, состоящих из матриц $(Q_{i,j})_{n,n'}$ интенсивностей переходов цепи ξ_t, $t \geq 0$, из макросостояния (i, n) в макросостояние (j, n'), n, $n' = \overline{0, N}$. Диагональные элементы матрицы $Q_{i,i}$ отрицательны и их модули определяют интенсивность выхода из соответствующего состояния цепи Маркова ξ_t, $t \geq 0$.

Лемма 1. *Инфинитезимальный генератор Q цепи Маркова ξ_t, $t \geq 0$, имеет блочно-трехдиагональную структуру:*

$$Q = \begin{pmatrix} Q_{0,0} & Q_{0,1} & O & O & \ldots \\ Q_{1,0} & Q_1 & Q_2 & O & \ldots \\ O & Q_0 & Q_1 & Q_2 & \ldots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}.$$

Ненулевые блоки $Q_{i,j}$, $i, j \geq 0$, имеют следующий вид:

$$Q_{0,0} = \mathrm{diag}\{A^{(0)}, \ldots, A^{(N)}\} + \mathrm{diag}^-\{F^{(1)}, \ldots, F^{(N)}\} +$$

$$+\mathrm{diag}^+\{B^{(0)},\dots,B^{(N-1)}\} + I_{(M+1)(N-M/2+1)} \otimes D_0,$$

$$Q_1 = \mathrm{diag}\{A^{(M)},\dots,A^{(N)}\} + \mathrm{diag}^-\{F^{(M+1)},\dots,F^{(N)}\}+$$

$$+\mathrm{diag}^+\{B^{(M)},\dots,B^{(N-1)}\} + I_{(N-M+1)(M+1)} \otimes D_0,$$

$$Q_{0,1} = \begin{pmatrix} O_{\frac{M(M+1)}{2}\bar{W}\times(N-M+1)(M+1)\bar{W}} \\ Q^+ \end{pmatrix},$$

$$Q_2 = \mathrm{diag}\{\underbrace{qI_{(M+1)} \otimes D_2,\dots,qI_{(M+1)} \otimes D_2}_{N-M}, pE^- \otimes D_1 + qI_{(M+1)} \otimes D_2\},$$

$$Q_{1,0} = \begin{pmatrix} O_{(N-M+1)(M+1)\bar{W}\times\frac{M(M+1)}{2}\bar{W}} & Q_0 \end{pmatrix},$$

$$Q_0 = \mathrm{diag}\{\mu_1\bar{C}_M E^+ + \mu_2 C_M, O_{M+1},\dots,O\} \otimes I_{\bar{W}},$$

где

$$A^{(n)} = \begin{cases} -\mathcal{C}_n \otimes I_{\bar{W}}, & 0 \le n < M, \\ -\mathcal{C}_n \otimes I_{\bar{W}} + (1-q)I_{(M+1)} \otimes D_2, & M \le n < N, \\ -\mathcal{C}_n \otimes I_{\bar{W}} + \\ +(\hat{I}_{M+1} + (1-p)E^-) \otimes D_1 + (1-q)I_{(M+1)} \otimes D_2, & n = N,\ \ i \ge 0; \end{cases}$$

$$B^{(n)} = \begin{cases} E_n^+ \otimes D_2 + \hat{E}_n^+ \otimes D_1, & 0 \le n < M, \\ I_{M+1} \otimes D_1, & M \le n < N; \end{cases}$$

$$F^{(n)} = \begin{cases} (\mu_1\bar{C}_n E_n^- + \mu_2 C_n \hat{E}_n^-) \otimes I_{\bar{W}}, & 0 < n \le M, \\ \mathcal{C}_n^- \otimes I_{\bar{W}}, & M < n \le N. \end{cases}$$

Доказательство леммы опирается на анализ переходов цепи Маркова ξ_t, $t \ge 0$, за бесконечно малый интервал времени с последующей группировкой интенсивностей соответствующих переходов в блоки матрицы Q.

Блоки генератора не зависят от компоненты i при $i > 1$ и цепь Маркова ξ_t, $t \ge 0$, принадлежит классу квазитеплицевых цепей Маркова (КТЦМ) с непрерывным временем, см. [6].

Как показано в [6], КТЦМ ξ_t, $t \ge 0$, является эргодической тогда и только тогда, когда выполнено неравенство

$$\mathbf{x}Q_0\mathbf{e} > \mathbf{x}Q_2\mathbf{e}, \tag{1}$$

где вектор $\mathbf{x}$ является единственным решением системы линейных алгебраических уравнений

$$\mathbf{x}(Q_0 + Q_1 + Q_2) = \mathbf{0},\ \mathbf{x}\mathbf{e} = 1. \tag{2}$$

Легко видеть, что система (2) может быть переписана в виде

$$\mathbf{0} = \mathbf{x}(Q_0 + Q_1 + Q_2) = \tag{3}$$

$$= \mathbf{x}\Bigg[\begin{pmatrix} I_{M+1}\otimes D_0 & I_{M+1}\otimes D_1 & \dots & O & O \\ O & I_{M+1}\otimes D_0 & \dots & O & O \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ O & O & \dots & I_{M+1}\otimes D_0 & I_{M+1}\otimes D_1 \\ O & O & \dots & O & (E^- + \hat{I})\otimes D_1 + I_{M+1}\otimes D_0 \end{pmatrix} +$$

$$+ I_{(N-M+1)(M+1)} \otimes D_2 +$$

$$+ \begin{pmatrix} -\mu_1\tilde{C}_M(I-E^+) & O & O & \dots & O & O \\ \mathcal{C}^-_{M+1} & -\mathcal{C}_{M+1} & O & \dots & O & O \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ O & O & O & \dots & -\mathcal{C}_{N-1} & O \\ O & O & O & \dots & \mathcal{C}^-_N & -\mathcal{C}_N \end{pmatrix} \otimes I_{\bar{W}}\Bigg],$$

$$\mathbf{x}\mathbf{e} = 1.$$

Домножив справа на $\mathbf{e}_{(N-M+1)(M+1)} \otimes I_{\bar{W}}$ левую и правую части системы (3), получим, что

$$\mathbf{x}(\mathbf{e}_{(N-M+1)(M+1)} \otimes D(1)) = \mathbf{0}, \mathbf{x}\mathbf{e} = 1. \tag{4}$$

Из (4) следует, что вектор $\mathbf{x}$ может быть представлен в виде

$$\mathbf{x} = \mathbf{z} \otimes \boldsymbol{\theta}, \tag{5}$$

где $\mathbf{z}$ – стохастический вектор размера $(N-M+1)(M+1)$, $\boldsymbol{\theta}$ – вектор стационарного распределения $MMAP$-потока запросов.

Подставляя вектор $\mathbf{x}$ в виде (5) в систему (3), домножая левую и правую части (3) справа на $I_{(N-M+1)(M+1)} \otimes \mathbf{e}_{\bar{W}}$, учитывая, что $\boldsymbol{\theta}D_1\mathbf{e} = \lambda_1$, $\boldsymbol{\theta}(D_0 + D_2)\mathbf{e} = -\boldsymbol{\theta}D_1\mathbf{e} = -\lambda_1$, $\boldsymbol{\theta}\mathbf{e} = 1$, мы получим, что вектор $\mathbf{z}$ является решением системы

$$\mathbf{z}\Omega = \mathbf{0}, \ \ \mathbf{z}\mathbf{e} = 1, \tag{6}$$

где матрица Ω имеет блочно-трехдиагональную структуру с поддиагональными блоками $\mathcal{C}^-_{M+1}, \dots, \mathcal{C}^-_N$, диагональными блоками $\tilde{A}^{(M)}, \dots, \tilde{A}^{(N)}$ и наддиагональными блоками $\lambda_1 I_{M+1}, \dots, \lambda_1 I_{M+1}$, где

$$\tilde{A}^{(n)} = \begin{cases} -\mu_1\tilde{C}_M(I-E^+) - \lambda_1 I_{M+1}, & n = M, M < N, \\ -\mu_1\tilde{C}_M(I-E^+) - \lambda_1(I - E^- - \hat{I}), & n = M = N, \\ -\mathcal{C}_n - \lambda_1 I_{M+1}, & M < n < N, \\ -\mathcal{C}_N - \lambda_1(I - E^- - \hat{I}_{M+1}), & n = N > M. \end{cases}$$

Легко видеть, что матрица Ω – это генератор двумерной цепи Маркова $\{n_t, l_t\}$, $t \geq 0$, определяющей число занятых приборов n_t, $n_t = \overline{M, N}$, и число приборов, занятых запросами второго типа, l_t, $l_t = \overline{0, M}$, в момент времени t при условии перегрузки системы. Очевидно, что при условии перегрузки рассматриваемой системы число занятых приборов n_t изменяется в интервале $[M, N]$. Из системы (6) следует, что вектор $\mathbf{z} =$

$(\mathbf{z}_M, \ldots, \mathbf{z}_N)$, $\mathbf{z}_n = (\mathbf{z}(n,0), \ldots, \mathbf{z}(n,M)), n = \overline{M,N}$, задает совместное стационарное распределение числа занятых приборов и числа запросов второго типа на обслуживании когда система перегружена.

С учетом этого, подставляя вектор $\mathbf{x}$ в виде (5) в неравенство (2), выполняя преобразования, получим неравенство:

$$\mathbf{z}_M(\mu_2 C_M + \mu_1 \bar{C}_M)\mathbf{e} > q\lambda_2 + p\lambda_1 \mathbf{z}_N \hat{\mathbf{e}}, \tag{7}$$

где $\hat{\mathbf{e}}$ – вектор размера $M+1$, первая компонента которого равна 0, а остальные равны 1.

Таким образом, мы доказали следующее утверждение.

Теорема 1. *Цепь ξ_t, $t > 0$, эргодична, тогда и только тогда, когда выполнено условие (7).*

Если условие эргодичности системы выполнено, то существуют пределы (стационарные вероятности) $\pi(i,n,l,\nu,) = \lim\limits_{t\to\infty} P\{i_t = i,\ n_t = n,\ l_t = l,\ \nu_t = \nu\},\ i \geq 0,\ n = \overline{(1-\delta_{i,0})M, N},\ l = \overline{0, \min\{n, M\}},\ \nu = \overline{0, W}$.

Сформируем вектор-строки $\boldsymbol{\pi}_i$ следующим образом:

$$\boldsymbol{\pi}(i,n,l) = (\pi(i,n,l,0), \pi(i,n,l,1), \ldots, \pi(i,n,l,W)),\ l = \overline{0, \min\{n,M\}},$$

$$\boldsymbol{\pi}(i,n) = (\boldsymbol{\pi}(i,n,0), \boldsymbol{\pi}(i,n,1), \ldots, \boldsymbol{\pi}(i,n,\min\{n,M\})),\ n = \overline{(1-\delta_{i,0})M, N},$$

$$\boldsymbol{\pi}_i = (\boldsymbol{\pi}(i,0), \boldsymbol{\pi}(i,1), \ldots, \boldsymbol{\pi}(i,N)),\ i \geq 0.$$

Общеизвестно, что векторы $\boldsymbol{\pi}_i$, $i \geq 0$, удовлетворяют следующей системе линейных алгебраических уравнений:

$$(\boldsymbol{\pi}_0, \boldsymbol{\pi}_1, \ldots)Q = \mathbf{0},\ \ (\boldsymbol{\pi}_0, \boldsymbol{\pi}_1, \ldots)\mathbf{e} = 1.$$

Для решения данной бесконечной системы, может быть использован алгоритм, предложенный в [6].

4. Характеристики производительности системы

Найдя векторы стационарных вероятностей $\boldsymbol{\pi}_i,\ i \geq 0$, можно вычислить различные характеристики производительности системы.

Среднее число запросов в системе $L = \sum\limits_{n=1}^{N} n\boldsymbol{\pi}(0,n)\mathbf{e} + \sum\limits_{i=1}^{\infty}\sum\limits_{n=M}^{N}(i+n)\boldsymbol{\pi}(i,n)\mathbf{e}$. Среднее число запросов в буфере $N^{buffer} = \sum\limits_{i=1}^{\infty} i\boldsymbol{\pi}_i\mathbf{e}$. Среднее число занятых приборов $N^{server} = \sum\limits_{i=0}^{\infty}\sum\limits_{n=1}^{N} n\boldsymbol{\pi}(i,n)\mathbf{e}$. Среднее число занятых приборов, обслуживающих запросы первого типа, $N^{server-1} = \sum\limits_{n=1}^{N}\sum\limits_{l=0}^{\min\{n,M\}}(n-l)\boldsymbol{\pi}(0,n,l)\mathbf{e} + \sum\limits_{i=1}^{\infty}\sum\limits_{n=M}^{N}\sum\limits_{l=0}^{M}(n-l)\boldsymbol{\pi}(i,n,l)\mathbf{e}$. Среднее число занятых приборов,

обслуживающих запросы второго типа $N^{server-2} = N^{server} - N^{server-1}$. Средняя интенсивность выходящего потока обслуженных запросов первого типа $\lambda_{out}^{(1)} = \mu_1 N^{server-1}$. Средняя интенсивность выходящего потока обслуженных запросов второго типа $\lambda_{out}^{(2)} = \mu_2 N^{server-2}$. Средняя интенсивность выходящего потока обслуженных запросов $\lambda_{out} = \lambda_1^{out} + \lambda_2^{out}$. Вероятность потери запроса первого типа вычисляется как $P_1^{loss} = \lambda_1^{-1} \sum_{i=0}^{\infty} \boldsymbol{\pi}(i, N, 0) D_1 \mathbf{e} = 1 - \frac{\lambda_1^{out}}{\lambda_1}$. Вероятность потери запроса второго типа вычисляется как $P_2^{loss} = 1 - \frac{\lambda_2^{out}}{\lambda_2}$. Вероятность потери произвольного запроса вычисляется как $P^{loss} = 1 - \frac{\lambda_{out}}{\lambda_1 + \lambda_2}$. Вероятность потери запроса второго типа на входе в систему из-за занятости более $M - 1$ приборов $P^{ent-loss} = (1 - q)\lambda_2^{-1} \sum_{i=0}^{\infty} \sum_{n=M}^{N} \boldsymbol{\pi}(i, n)(I_{(M+1)} \otimes D_2)\mathbf{e}$. Вероятность того, что запрос второго типа будет выбит запросом первого типа и покинет систему $P^{knock-out-loss} = (1 - p)\lambda_1^{-1} \sum_{i=0}^{\infty} \sum_{l=1}^{M} \boldsymbol{\pi}(i, N, l) D_1 \mathbf{e}$.

5. Заключение

В работе исследована многолинейная система обслуживания с бесконечным буфером и двумя типами запросов. Найдено условие эргодичности. Получены основные характеристики производительности.

Данная работа выполнена при частичной поддержке Белорусского республиканского фонда фундаментальных исследований.

ЛИТЕРАТУРА

1. Kim C.S., Klimenok V., Dudin A. Optimization of Guard Channel Policy in Cellular Mobile Networks with Account of Retrials // Computers and Operation Research. 2014. V. 43. P. 181-190.
2. Kim C.S., Klimenok V., Taramin O. A tandem retrial queueing system with two Markovian flows and reservation of channels // Computers and Operations Research. 2010. V. 37. P. 1238-1246.
3. J. Mitola, G.Q. Maguire, Cognitive radio: making software radios more personal // IEEE Personal Communications. 1999. V. 6. P. 13-18.
4. Akyildiz I.F., Lee W.Y., Vuran M.C., Mohanty S. Next generation dynamic spectrum access cognitive radio wireless networks: A survey // Computer Networks. 2006. V. 50. P. 2127-2159.
5. Graham A. Kronecker products and matrix calculus with applications. Cichester, Ellis Horwood, 1981.
6. Neuts M. F. Structured Stochastic Matrices of M/G/1 Type and Their Applications. Marcel Dekker, New York, 1989.

TWO-SERVER QUEUEING SYSTEM WITH PHASE-TYPE SERVICE TIME DISTRIBUTION AND COMMON PHASES OF SERVICE

Ch. Kim[1], *A. Dudin*[2], *S. Dudin*[2], *O. Dudina*[2]

[1] Sangji University, Wonju, Kangwon, 220-702, Korea

[2] Belarusian State University, 4, Nezavisimosti Ave., Minsk, 220030, Belarus

dowoo@sangji.ac.kr, dudin@bsu.by, dudin85@mail.ru, dudina_olga@email.com

Abstract

We consider two server queueing system with infinite buffer. Customers arrive to the system according to the Markovian arrival process. Service time of a customer has a phase-type distribution. The servers use the same equipment (phases of PH) for customers processing. So, if service of a customer transits to the phase, at which another server is currently providing the service, the service of the customer is suspended until the phase will become available. Behavior of the system is described by the multi-dimensional Markov chain. The generator of this Markov chain is derived. The main performance measures are calculated.

Keywords: Markovian arrival process, phase-type service time distribution

1. Introduction

The simplest queueing models assume that service time is exponentially distributed. This allows to ignore the elapsed service time when some Markovian process describing behavior of the queue is constructed. It is quite obvious that in many real world applications of queueing theory assumption about the exponential distribution fails and more general distributions of service time have to be considered. If service time has an arbitrary distribution, it is mandatory to take into account the elapsed or residual service times, e.g., by introducing supplementary variables. This may lead to huge analytical difficulties in analysis of the Markovian process describing dynamics of the queue, especially when multi-server queue is under study. To avoid such difficulties, so called phase type (PH) distribution, as natural extension of previously popular Erlangian and hyper-exponential distribution, was offered, see, e.g., [1]. Nice property of this distribution is its generality. Any distribution can be approximated, in sense of weak convergence, by the PH type distribution, see, e.g., [2]. Random time having the PH type distribution is interpreted as the sequence of phases duration of which has exponential distribution. Generally speaking the number of such phases is random. Formal definition of PH type distribution is given in the next section. For purposes of this paper, we slightly rephrase this definition as follows. There is a virtual network with nodes (phases), say,

$\{1, \ldots, M\}$. Random time having the PH type distribution with an irreducible representation $(\boldsymbol{\beta}, S)$ is the time during which some virtual customer stays in this network conditional on the fact that this virtual customer starts its staying in the network from the visit to the state m of the network with probability $\boldsymbol{\beta}_m,\ m = \overline{1, M}$, then it makes transitions inside of the set $\{1, \ldots, M\}$ with intensities given by the entries of the matrix S or leaves the network from any state m with intensity, which is the mth component of column vector $\mathbf{S}_0 = -S\mathbf{e}$, where $\mathbf{e}$ denotes unit column vector. In brief, as it was already noted above, random time having the PH type distribution consists of the random number of *virtual* phases, duration of which is exponentially distributed. This allows to replace keeping track of the **continuous** elapsed or residual service time by the keeping track of the **discrete** current phase of the service what greatly simplifies the analysis. However, in many real world situations, random time having PH type distribution represents the sequence of *real* phases. E.g., processing time of the query in data base consists of implementation of a sequence of input/output operations alternating with the use of CPU. Processing of a car in service station consists of a sequence of technological operations. If the service is provided by the single server, no collisions occur. However, if there are several servers operating in parallel, collisions may occur because the servers use some common resources, e.g. tables of indices of data base of equipment of service station.

Traditionally, in analysis of multi-server queues with PH type service time distribution it is assumed that service processes in the servers are independent. To the best of our knowledge, the systems with interference of phases of service in different servers are not considered in literature. In this paper, we start research of such systems from a relatively simple model where there are two servers, state spaces of the networks, in terms of which PH type service time distribution is interpreted, coincide and if a phase of the service by a server is required while it is busy by another server this phase of service is postponed until this phase will be released by another server. In some sense, in this model we somehow unite the problems from two popular theories – queueing theory and scheduling theory which consider the similar objects (service systems) but with emphasis to different aspects of the problem.

2. Mathematical Model

Queueing system with two servers and an infinite buffer is considered.

Customers arrive at the system according to the Markovian arrival process. We will code this process as MAP. Arrivals in the MAP are directed by an irreducible continuous time Markov chain ν_t, $t \geq 0$, with the finite state space $\{0, 1, ..., W\}$. The sojourn time of the Markov chain ν_t, $t \geq 0$, in the state ν has an exponential distribution with the parameter $\lambda_\nu,\ \nu = \overline{0, W}$. Here, notation such as $\nu = \overline{0, W}$ means that ν assumes values from the set $\{0, 1, ..., W\}$. After this sojourn time expires, with probability $p_k(\nu, \nu')$ the process ν_t transits to the state ν' and k customers, $k = 0, 1$, arrive at the system. The intensities of

transitions from one state to another, that are accompanied by the arrival of k customers, are combined to the square matrices $D_k,\ k = 0, 1$, of size $W + 1$. The matrix generating function of these matrices is $D(z) = D_0 + D_1 z,\ |z| \le 1$. The matrix $D(1)$ is an infinitesimal generator of the process $\nu_t,\ t \ge 0$. The stationary distribution vector $\boldsymbol{\theta}$ of this process satisfies the system of equations $\boldsymbol{\theta} D(1) = \mathbf{0},\ \boldsymbol{\theta}\mathbf{e} = 1$. Here and throughout this paper, $\mathbf{0}$ is a zero row vector. In case if the dimension of a vector is not clear from the context, it is indicated as a lower index.

The average intensity λ (fundamental rate) of the MAP is defined by $\lambda = \boldsymbol{\theta} D_1 \mathbf{e}$.

The service time of a customer for the server has a PH (phase-type) distribution with an irreducible representation $(\boldsymbol{\beta}, S)$. This service time can be interpreted as the time until the underlying Markov process $m_t, t \ge 0$, with a finite state space $\{1, \dots, M, M + 1\}$ reaches the single absorbing state $M + 1$ condition on the fact that the initial state of this process is selected among the states $\{1, \dots, M\}$ according to the probabilistic row vector $\boldsymbol{\beta} = (\beta_1, \dots, \beta_M)$. The transition rates of the process m_t within the set $\{1, \dots, M\}$ 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 $b_1 = \boldsymbol{\beta}(-S)^{-1}\mathbf{e}$.

We assume that the service of customers by two servers is not independent. The state space of the underlying Markov processes for both the servers is the same. If there is no collision, these processes make transitions according to definition given above, independently of each other. However if the service of a customer by the l-th server, $l = 1, 2$ is at some phase while the underlying service process of other customer by l'th, $l' = 1, 2$, $l' \ne l$, server should transit to the same phase, the l'-th server is blocked and service process is stopped until the l-th server finishes this phase of the service. The servers are identical and are enumerated in arbitrary order. However, if two servers need the same phase of the service, they are assumed being enumerated in order of occupation of the phase. Number 1 has the server that currently occupies the phase under the conflict. Number 2 has the server that currently waits for releasing this phase.

The customers from the buffer are impatient, i.e., the customer leaves the buffer and the system after an exponentially distributed waiting time described by the parameter α, $0 < \alpha < \infty$.

3. Process of system states

It is easy to see that the behavior of the system under study is described in terms of the following regular irreducible continuous-time Markov chain

$$\xi_t = \{i_t, r_t, \nu_t, n_t,\ m_t\},\ t \ge 0,$$

where, during the epoch $t,\ t \ge 0$,

- i_t is the number of customers in the system, $i_t \geq 0$;
- r_t is an indicator that indicates whether some server is blocked or not: $r_t = 0$ corresponds to the case when a server isn't blocked and $r_t = 1$ otherwise;
- ν_t is the state of the underlying process of the MAP_1, $\nu_t = \overline{0, W}$;
- n_t is the state of PH service process on the first server, $n_t = \overline{1, M}$.
- m_t is the state of PH service process on the second server, $m_t = \overline{1, M}$, $m_t \neq n_t$.

The Markov chain ξ_t, $t \geq 0$, has the following state space:

$$\Big(\{0, 0, \nu\}\Big) \bigcup \Big(\{1, 0, \nu, n\},\, n = \overline{1, M}\Big) \bigcup$$

$$\Big(\{i, 0, \nu, n, m\},\, i \geq 2,\, n = \overline{1, M},\, m = \overline{1, M},\, m \neq n\Big) \bigcup$$

$$\Big(\{i, 1, \nu, n\},\, i \geq 2,\, n = \overline{1, M}\Big),\, \nu = \overline{0, W}.$$

For further use throughout this paper, we introduce the following notation:

- I is the identity matrix, and O is a zero matrix of appropriate dimension;
- $\otimes$ and $\oplus$ indicate the symbols of Kronecker product and sum of matrices, respectively;
- $\bar{W} = W + 1$;
- I_{l_1,l_2}, $l_1, l_2 = \overline{1, M}$, $l_1 \neq l_2$, is the matrix of size $(M-1) \times (M-1)$ with all zero entries except the entries $(I_{l_1,l_2})_{k,k}$, $k = \overline{0, M-2}$, $k \neq l_2 - 2$, in the case $(l_1 < l_2)$ and $(I_{l_1,l_2})_{k,k}$, $k = \overline{0, M-2}$, $k \neq l_2 - 1$, in the case $(l_1 > l_2)$ which are equal to 1;
- $\tilde{S}_l$, $l = \overline{1, M}$, is the square matrix of size $M-1$ that is obtained from matrix S by removing the $l-1$-th column and the l-th row;
- $\mathbf{e}_{l_1,l_2}$, $l_1, l_2 = \overline{1, M}$, $l_1 \neq l_2$, is the column vector of size $(M-1)$ with all zero entries except the entry $(\mathbf{e}_{l_1,l_2})_{l_2-1}$ in the case $(l_1 > l_2)$ and $(\mathbf{e}_{l_1,l_2})_{l_2-2}$, in the case $(l_1 < l_2)$ which are equal to 1;
- $\mathbf{c}_{l_1,l_2}$, $l_1, l_2 = \overline{1, M}$, $l_1 \neq l_2$, is the row vector of size $(M-1)$ with all zero entries except the entry $(\mathbf{c}_{l_1,l_2})_{l_1-2}$ in the case $(l_1 > l_2)$ and $(\mathbf{c}_{l_1,l_2})_{l_1-1}$, in the case $(l_1 < l_2)$ which are equal to 1;
- $\boldsymbol{\beta}_l$, $l = \overline{1, M}$, - the row vector that obtained from the vector $\boldsymbol{\beta}$ by delating $l-1$-th component;
- $\mathbf{a}_l$, $l = \overline{1, M}$, is the column vector of size $M-1$ that is obtained from the $l-1$-th column of the matrix S by removing the $l-1$-th entry;
- I_l^+, $l = \overline{1, M}$, is the matrix of size $(M-1) \times M$ which obtained from the identity matrix of size $M-1$ by adding the zero column in position $l-1$;
- $\mathbf{S}_0^l$, $l = \overline{1, M}$ is a column vector of size $M-1$ which is obtained from the vector $\mathbf{S}_0$ by removing the $l-1$-th component.

- $\tilde{\mathbf{a}}_l$, $l = \overline{1, M}$, is a row vector of size M with all zero components except the component $(\tilde{\mathbf{a}}_l)_{l-1}$ which is equal to 1;
- B_l, $l = \overline{1, M}$, is the matrix of size $(M-1) \times M(M-1)$ which obtained from the matrix $\mathrm{diag}\{\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_M\}$ by deleting the $l-1$-th row;
- C_l, $l = \overline{1, M}$, is the matrix of size $(M-1) \times M$ which obtained from the matrix $\mathrm{diag}\{\beta_1, \ldots, \beta_M\}$ by deleting the $l-1$-th row;

Let us enumerate the states of the Markov chain ξ_t, $t \geq 0$, in the direct lexicographic order of the components r, k, ν, ζ, η and refer to the set of the states of the chain having values (i, r) of the first two components of the Markov chain as a macro-state (i, r). Let Q be the generator of the Markov chain ξ_t, $t \geq 0$, consisting of the blocks $Q_{i,j}$, which, in turn, consist of the matrices $(Q_{i,j})_{r,r'}$ of the transition rates of this chain from the macro-state (i, r) to the macro-state (j, r'), $r, r' = 0, 1$. The diagonal entries of the matrices $Q_{i,i}$ are negative, and the modulus of the diagonal entry of the blocks $(Q_{i,i})_{r,r}$ defines the total intensity of leaving the corresponding state of the Markov chain ξ_t, $t \geq 0$.

Analysing all transitions of the Markov chain $\xi_t, t \geq 0$, during an interval of an infinitesimal length and rewriting the intensities of these transitions in the block matrix form we obtain the following result.

Theorem 1. *The infinitesimal generator $Q = (Q_{i,j})_{i,j \geq 0}$ of the Markov chain ξ_t, $t \geq 0$, has a block-tridiagonal structure:*

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

The non-zero blocks $Q_{i,j}$, $i, j \geq 0$, have the following form:

$$Q_{0,0} = D_0,\ Q_{1,1} = D_0 \oplus S,\ Q_{i,i} = \begin{pmatrix} Q_{i,i}^{(0,0)} & Q_{i,i}^{(0,1)} \\ Q_{i,i}^{(1,0)} & Q_{i,i}^{(1,1)} \end{pmatrix},\ i > 1,$$

$$Q_{i,i}^{(0,0)} = D_0 \otimes I_{M(M-1)} + I_{\bar{W}} \otimes (\mathcal{S} + \mathrm{diag}\{\tilde{S}_1, \ldots, \tilde{S}_M\}) - (i-2)\alpha I_{\bar{W}M(M-1)},\ i > 1,$$

$$\mathcal{S} = \begin{pmatrix} S_{1,1}I_{M-1} & S_{1,2}I_{1,2} & \ldots & S_{1,M}I_{1,M} \\ S_{2,1}I_{2,1} & S_{2,2}I_{M-1} & \ldots & S_{2,M}I_{2,M} \\ \vdots & \vdots & \ldots & \vdots \\ S_{M,1}I_{M,1} & S_{M,2}I_{M,2} & \ldots & S_{M,M}I_{M-1} \end{pmatrix},$$

$$Q_{i,i}^{(0,1)} = I_{\bar{W}} \otimes \left(\begin{pmatrix} \mathbf{0}^T & S_{1,2}\mathbf{e}_{1,2} & \ldots & S_{1,M}\mathbf{e}_{1,M} \\ S_{2,1}\mathbf{e}_{2,1} & \mathbf{0}^T & \ldots & S_{2,M}\mathbf{e}_{2,M} \\ \vdots & \vdots & \ldots & \vdots \\ S_{M,1}\mathbf{e}_{M,1} & S_{M,2}\mathbf{e}_{M,2} & \ldots & \mathbf{0}^T \end{pmatrix} + \mathrm{diag}\{\mathbf{a}_1, \ldots, \mathbf{a}_M\} \right),\ i > 1,$$

$$Q_{i,i}^{(1,0)} = I_{\bar{W}} \otimes \begin{pmatrix} \mathbf{0} & S_{1,2}\mathbf{c}_{1,2} & \dots & S_{1,M}\mathbf{c}_{1,M} \\ S_{2,1}\mathbf{c}_{1,M} & \mathbf{0} & \dots & S_{2,M}\mathbf{c}_{1,M} \\ \vdots & \vdots & \dots & \vdots \\ S_{M,1}\mathbf{c}_{1,M} & S_{M,2}\mathbf{c}_{1,M} & \dots & \mathbf{0} \end{pmatrix},$$

$$Q_{i,i}^{(1,1)} = D_0 \oplus \mathrm{diag}\{S_{1,1}, \dots, S_{M,M}\} - (i-2)\alpha I_{\bar{W}M},\ i > 1,$$

$$Q_{0,1} = D_1 \otimes \boldsymbol{\beta},\ Q_{1,2} = \begin{pmatrix} Q_{1,2}^{(0,0)} & Q_{1,2}^{(0,1)} \end{pmatrix},\ Q_{1,2}^{(0,0)} = D_1 \otimes \mathrm{diag}\{\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_M\},$$

$$Q_{1,2}^{(0,1)} = D_1 \otimes \mathrm{diag}\{\beta_1, \dots, \beta_M\},$$

$$Q^{+} = \begin{pmatrix} D_1 \otimes I_{M(M-1)} & O \\ O & D_1 \otimes I_M \end{pmatrix},$$

$$Q_{1,0} = I_{\bar{W}} \otimes S_0,\ Q_{2,1} = \begin{pmatrix} Q_{2,1}^{(0,0)} \\ Q_{2,1}^{(1,0)} \end{pmatrix},\ i > 1,$$

$$Q_{2,1}^{(0,0)} = I_{\bar{W}} \otimes \Bigg(\begin{pmatrix} (\mathbf{S_0})_1 I_1^{+} \\ \vdots \\ (\mathbf{S_0})_M I_M^{+} \end{pmatrix} + \mathrm{diag}\{\mathbf{S}_0^l,\ l = \overline{1, M}\} \Bigg),\ Q_{2,1}^{(1,0)} = I_{\bar{W}} \otimes \begin{pmatrix} (\mathbf{S_0})_1 \tilde{\mathbf{a}}_1 \\ \vdots \\ (\mathbf{S_0})_M \tilde{\mathbf{a}}_M \end{pmatrix},$$

$$Q_{i,i-1} = \begin{pmatrix} Q_{i,i-1}^{(0,0)} & Q_{i,i-1}^{(0,1)} \\ Q_{i,i-1}^{(1,0)} & Q_{i,i-1}^{(1,1)} \end{pmatrix},\ i > 2,$$

$$Q_{i,i-1}^{(0,0)} = I_{\bar{W}} \otimes \Bigg(\begin{pmatrix} (\mathbf{S_0})_1 B_1 \\ \vdots \\ (\mathbf{S_0})_M B_M \end{pmatrix} + \mathrm{diag}\{\mathbf{S}_0^l \boldsymbol{\beta}_l,\ l = \overline{1, M}\} \Bigg) + (i-2)\alpha I_{\bar{W}M(M-1)},$$

$$Q_{i,i-1}^{(0,1)} = I_{\bar{W}} \otimes \Bigg(\begin{pmatrix} (\mathbf{S_0})_1 C_1 \\ \vdots \\ (\mathbf{S_0})_M C_M \end{pmatrix} + \mathrm{diag}\{\mathbf{S}_0^l \beta_l,\ l = \overline{1, M}\} \Bigg),$$

.

$$Q_{i,i-1}^{(1,0)} = I_{\bar{W}} \otimes \mathrm{diag}\{(\mathbf{S_0})_l \boldsymbol{\beta}_l,\ l = \overline{1, M}\},$$

$$Q_{i,i-1}^{(1,1)} = I_{\bar{W}} \otimes \mathrm{diag}\{(\mathbf{S_0})_l \beta_l,\ l = \overline{1, M}\} + (i-2)\alpha I_{\bar{W}M}.$$

Corollary 1. The Markov chain $\xi_t,\ t \geq 0$, belongs to the class of continuous-time asymptotically quasi-Toeplitz Markov chains ($AQTMC$), see [3].

Let us analyze the properties of this Markov chain. This analysis should include derivation of conditions which should be imposed on the system parameters to guarantee existence of the stationary distribution of the states of the chain (ergodicity condition) and a procedure for computation of the stationary probabilities of the states.

As follows from [3], a sufficient condition for the existence of a stationary distribution of $AQTMC$ $\xi_t,\ t \geq 0$, is expressed in terms of the matrices Y_0, Y_1 and Y_2 defined as follows:

$$Y_0 = \lim_{i\to\infty} R_i^{-1} Q_{i,i-1},\ Y_1 = \lim_{i\to\infty} R_i^{-1} Q_{i,i} + I,\ Y_2 = \lim_{i\to\infty} R_i^{-1} Q_{i,i+1}$$

where the matrix R_i is a diagonal matrix with the diagonal entries which are defined as the moduli of the corresponding diagonal entries of the matrix $Q_{i,i},\ i \geq 0$.

It is easy to verify that in the considered case the matrices Y_0, Y_1 and Y_2 have the following form:

$$Y_0 = I,\ Y_1 = O,\ Y_2 = O.$$

It follows from [3] that sufficient condition of ergodicity of Markov chain $\xi_t,\ t \geq 0$, is fulfillment of inequality

$$\mathbf{y}Y_0\mathbf{e} > \mathbf{y}Y_2\mathbf{e} \tag{1}$$

where the vector $\mathbf{y}$ is the unique solution to the system

$$\mathbf{y}(Y_0 + Y_1 + Y_2) = \mathbf{y},\ \mathbf{y}\mathbf{e} = 1.$$

It is easy to see that here this ergodicity condition (1) is transformed to inequality $1 > 0$ which is true for all possible values of the system parameters.

Thus, the following limits (stationary probabilities) exist:

$$\pi(i, r, \nu, n, m) = \lim_{t\to\infty} P\{i_t = i,\ r_t = r,\ \nu_t = \nu,\ n_t = n, m_t = m\},$$

$$i \geq 0,\ r = \overline{0,1},\ \nu = \overline{0,W},\ n = \overline{1,M},\ m = \overline{1,M}.$$

Then let us form the row vectors $\boldsymbol{\pi}_i$ of the stationary probabilities as follows:

$$\boldsymbol{\pi}_0 = (\pi(0,0,0), \pi(0,0,1), \ldots, \pi(0,0,W)),$$

$$\boldsymbol{\pi}_1 = (\boldsymbol{\pi}(1,0,0), \boldsymbol{\pi}(1,0,1), \ldots, \boldsymbol{\pi}(1,0,W)),$$

where

$$\boldsymbol{\pi}(1,0,\nu) = (\pi(1,0,\nu,1), \pi(1,0,\nu,2), \ldots, \pi(1,0,\nu,M)),\ \nu = \overline{0,W}.$$

$$\boldsymbol{\pi}_i = (\boldsymbol{\pi}(i,0), \boldsymbol{\pi}(i,1)),\ i \geq 2,$$

where

$$\boldsymbol{\pi}(i,r) = (\boldsymbol{\pi}(i,r,0), \boldsymbol{\pi}(i,r,1), \ldots, \boldsymbol{\pi}(i,r,W)),\ \ i \geq 2,$$

$$\boldsymbol{\pi}(i,0,\nu) = (\boldsymbol{\pi}(i,0,\nu,1), \boldsymbol{\pi}(i,0,\nu,2), \ldots, \boldsymbol{\pi}(i,0,\nu,M)),\ \nu = \overline{0,W},$$

$$\boldsymbol{\pi}(i,0,\nu,n) = (\pi(i,0,\nu,n,1), \pi(i,0,\nu,n,2), \ldots, \pi(i,0,\nu,n,M)),\ n = \overline{1,M},$$

$$\boldsymbol{\pi}(i,1,\nu) = (\pi(i,1,\nu,1), \pi(i,1,\nu,2), \dots, \pi(i,1,\nu,M)),\ \nu = \overline{0,W}.$$

It is well known that the probability vectors $\boldsymbol{\pi}_i$, $i \geq 0$, satisfy the following system of linear algebraic equations:

$$(\boldsymbol{\pi}_0, \boldsymbol{\pi}_1, \dots)Q = \mathbf{0},\ \ (\boldsymbol{\pi}_0, \boldsymbol{\pi}_1, \dots)\mathbf{e} = 1. \tag{2}$$

To solve system (2), we use the numerically stable algorithm for computation of the probability vectors $\boldsymbol{\pi}_i$, $i \geq 0$, developed in [4] which effectively uses information about the asymptotic behavior of the Markov chain ξ_t, $t \geq 0$, and the sparse structure of the generator Q.

4. Performance measures

The average number N_{buffer} of customers in the buffer is computed by

$$N_{buffer} = \sum_{i=3}^{\infty} (i-2)\mathbf{p}_i\mathbf{e}.$$

The average number N_{busy} of busy servers at an arbitrary moment is computed by

$$N_{busy} = \sum_{i=1}^{\infty} \min\{\mathrm{i}, 2\}\mathbf{p}_i\mathbf{e}.$$

The probability $P_{blocked}$ that server is blocked at an arbitrary moment is computed by

$$P_{blocked} = \sum_{i=2}^{\infty} \mathbf{p}(i,1)\mathbf{e}.$$

The probability P_{loss} that an arbitrary customer will be lost (due to impatience) is computed by

$$P_{loss} = \frac{\alpha N_{buffer}}{\lambda}.$$

The average intensity λ_{out} of flow of customers who receive service is computed by

$$\lambda_{out} = \lambda(1 - P_{loss}).$$

5. Conclusion

Two server queueing model is analysed. Service times by two servers have phase type distribution with coinciding state spaces of underlying Markov chains. These service times are independent if underlying Markov chains currently have different states (phases). If the phases coincide, service by one of the servers is postponed until the phase will be released by the competitive

server. Generator of multi-dimensional Markov chain describing behavior of the system is written down. Formulas for computation of the key performance measure of the system in terms of stationary probabilities of the Markov chain are presented. It is planned to extend the results to the cases when the number of servers is more than two, when the state spaces of underlying processes of service coincide partially, when the number of active servers can be dynamically changed, etc.

6. Acknowledgements

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (Grant No.2014R1A1A4A01007517).

REFERENCES

1. Neuts, M.F.: Matrix-geometric solutions in stochastic models. Baltimore, The Johns Hopkins University Press (1981)
2. Asmussen, S.: Applied Probability and Queues. New York, Springer-Verlag (2003)
3. Klimenok, V.I., Dudin, A.N.: Multi-dimensional asymptotically quasi-Toeplitz Markov chains and their application in queueing theory. Queueing Systems 54 245-259 (2006)
4. Dudina, O., Kim, Ch., Dudin, S.: Retrial queueing system with Markovian arrival flow and phase-type service time distribution. Computers and Industrial Engineering 66 360-373 (2013)

TANDEM INFINITE-SERVER QUEUEING SYSTEM WITH HIGH-RATE MARKOVIAN ARRIVAL PROCESS

A. Moiseev, A. Nazarov
Tomsk State University, Tomsk, Russia

We consider a tandem queueing system with infinite number of servers and Markovian arrival process. Service times at the system stages are i.i.d. and given by distribution functions individually for each stage. The study is performed under the asymptotic condition of the arrivals' rate growth. It is shown that multi-dimensional probability distribution of customers number at the system stages can be approximated by multi-dimensional Gaussian distribution which parameters are obtained in the paper.

АНАЛИЗ МНОГОФАЗНОЙ СИСТЕМЫ ОБСЛУЖИВАНИЯ С НЕОГРАНИЧЕННЫМ ЧИСЛОМ ПРИБОРОВ И ВЫСОКОИНТЕНСИВНЫМ ВХОДЯЩИМ MAP-ПОТОКОМ

А. Моисеев, А. Назаров

Томский государственный университет, Томск, Россия
moiseev.tsu@gmail.com, nazarov.tsu@gmail.com

Аннотация

В работе представлено исследование многофазной системы массового обслуживания с входящим MAP-потоком, неограниченным числом приборов и произвольным обслуживанием на фазах системы. Показано, что в условии неограниченно растущей интенсивности входящего потока многомерное распределение вероятностей числа заявок на фазах системы в стационарном режиме функционирования является асимптотически нормальным, получены параметры соответствующего многомерного нормального распределения.

Ключевые слова: Многофазная система массового обслуживания, MAP-поток, асимптотический анализ

1. Введение

В приложениях теории массового обслуживания очень часто применяются модели многофазных систем обслуживания [1, 2], предполагающих поэтапную обработку поступающих сообщений (заявок). Но обычно в

литературе встречается исследование моделей с входящим пуассоновским потоком и/или экспоненциальным обслуживанием на фазах либо анализ многофазных систем специфической конфигурации (например, двухфазных систем [3]).

В настоящей работе представлено исследование многофазной системы обслуживания с входящим MAP-потоком [4], неограниченным числом приборов и произвольным обслуживанием на фазах системы. Показано, что в условии растущей интенсивности входящего потока многомерное распределение числа заявок на фазах системы в стационарном режиме может быть аппроксимировано многомерным нормальным распределением, параметры которого получены в работе.

Аналогичные исследования для многофазной системы с входящим рекуррентным потоком выполнено в [5].

2. Постановка задачи

Пусть имеется многофазная система массового обслуживания с неограниченным числом приборов на каждой фазе и входящим MAP-потоком. Система состоит из K фаз обслуживания. Время обслуживания на k-й фазе является случайной величиной с функцией распределения $B_k(x)$, $k = \overline{1, K}$. Заявка входящего потока поступает на первую фазу. По окончании обслуживания на k-й фазе, она переходит на следующую, $(k+1)$-ю фазу до тех пор, пока не получит обслуживания на последней K-й фазе, после чего покидает систему.

Входящий поток заявок является MAP-потоком [4], заданный представлением $(\mathbf{D}_0, \mathbf{D}_1)$. Здесь $\mathbf{D}_0$ и $\mathbf{D}_1$ – матрицы порядка M. Матрица $\mathbf{D} = \mathbf{D}_0 + \mathbf{D}_1$ является генератором управляющей цепи Маркова $m(t)$. Обозначим ее стационарное распределение вероятностей через $\mathbf{r}$. Вектор-строка $\mathbf{r}$ удовлетворяет системе уравнений:

$$\begin{cases} \mathbf{rD} = \mathbf{0}, \\ \mathbf{re} = 1. \end{cases} \tag{1}$$

Здесь $\mathbf{0}$ и $\mathbf{e}$ – вектор-строка из нулей и вектор-столбец из единиц соответственно. Величина

$$\lambda = \mathbf{rD}_1\mathbf{e} \tag{2}$$

называется интенсивностью MAP-потока (*fundamental rate*).

Обозначим $i_k(t)$ – число заявок, находящихся в момент времени t на обслуживании на k-й фазе системы. Ставится задача поиска многомерного стационарного распределения вероятностей вектора состояний $\mathbf{i}(t) = \{i_1(t), \ldots, i_K(t)\}$ – числа заявок на фазах системы.

3. Метод многофазного динамического просеивания

Прямое исследование случайного процесса $\mathbf{i}(t)$ достаточно затруднено. Поэтому для решения поставленной задачи воспользуемся методом мно-

гофазного динамического просеивания, подробно представленного в [5]. Кратко опишем его здесь.

Зафиксируем некоторый момент времени T. Обозначим через $S_k(t)$, $k = \overline{1,K}$, вероятность того, что заявка входящего потока, поступившая в момент времени $t < T$, в момент T будет обслуживаться на k-й фазе системы. Через

$$S_0(t) = 1 - \sum_{k=1}^{K} S_k(t) \tag{3}$$

обозначим вероятность того, что указанная заявка покинет систему до момента T.

Определим K так называемых просеянных потоков событий. Будем считать, что заявка входящего потока, поступающая в систему в момент времени t, с вероятностью $S_k(t)$ генерирует событие в k-м просеянном потоке, а с вероятностью $S_0(t)$ – не генерирует события ни в одном из потоков.

Пусть в начальный момент времени $t_0 < T$ система пуста. Обозначим $n_k(t)$ – число событий, наступивших в k-м просеянном потоке до момента времени t. Тогда для вектора $\mathbf{n}(t) = \{n_1(t), \ldots, n_K(t)\}$ в момент времени $t = T$ имеем равенства

$$\mathrm{P}\{\mathbf{i}(T) = \mathbf{i}\} = \mathrm{P}\{\mathbf{n}(T) = \mathbf{i}\} \tag{4}$$

для любых значений $\mathbf{i}$, то есть распределения вероятностей значений случайных процессов $\mathbf{i}(t)$ и $\mathbf{n}(t)$ в этот момент времени совпадают. Таким образом, получив выражение для распределения вероятностей многомерного процесса $\mathbf{n}(t)$ и подставив $t = T$, получим распределение вероятностей значений исследуемого процесса $\mathbf{i}(t)$ в момент времени T, который, вообще говоря, выбран произвольно.

В работе [5] получены следующие выражения для вероятностей $S_k(t)$:

$$S_k(t) = B^*_{k-1}(T-t) - B^*_k(T-t). \tag{5}$$

где $B^*_k(x) = (\underbrace{B_1 * \cdots * B_k})(x)$ есть свертка функций $B_1(x), \ldots, B_k(x)$ для значений $k = \overline{2,K}$ и $B^*_0(x) \equiv 1$, $B^*_1(x) = B_1(x)$.

4. Уравнения Колмогорова

Очевидно, что процесс $\{\mathbf{n}(t), m(t)\}$ является марковским (здесь $m(t)$ – состояние управляющей входящим MAP-потоком цепи Маркова), и для его распределения вероятностей $P(\mathbf{n}, m, t) = \mathrm{P}\{\mathbf{n}(t) = \mathbf{n}, m(t) = m\}$ можно записать систему дифференциальных уравнений Колмогорова

$$\frac{\partial P(\mathbf{n}, m, t)}{\partial t} = \sum_{\eta=1}^{M} P(\mathbf{n}, \eta, t)(\mathbf{D}_0)_{\eta m} +$$

$$+\sum_{\eta=1}^{M}\left[P(\mathbf{n},\eta,t)(\mathbf{D}_1)_{\eta m}S_0(t)+\sum_{k=1}^{K}P(\mathbf{n}-\mathbf{e}_k,\eta,t)(\mathbf{D}_1)_{\eta m}S_k(t)\right]$$

для всех значений $m=\overline{1,M}$ и неотрицательных $\mathbf{n}$. Здесь $\mathbf{e}_k$ – вектор, все элементы которого равны 0 за исключением k-го, который равен 1.

С использованием векторных обозначений эта система перепишется в виде

$$\frac{\partial\mathbf{P}(\mathbf{n},t)}{\partial t}=\mathbf{P}(\mathbf{n},t)\mathbf{D}_0+\mathbf{P}(\mathbf{n},t)\mathbf{D}_1S_0(t)+\sum_{k=1}^{K}\mathbf{P}(\mathbf{n}-\mathbf{e}_k,t)\mathbf{D}_1S_k(t), \qquad (6)$$

где $\mathbf{P}(\mathbf{n},t)=\{P(\mathbf{n},1,t),\ldots,P(\mathbf{n},M,t)\}$, причем $P(\mathbf{n},t)=0$, если хотя бы один элемент вектора $\mathbf{n}$ меньше нуля.

Рассмотрим векторную характеристическую функцию

$$\mathbf{H}(\mathbf{u},t)=\sum_{n_1=0}^{\infty}\cdots\sum_{n_K=0}^{\infty}e^{ju_1n_1+\cdots+ju_Kn_K}\mathbf{P}(\mathbf{n},t).$$

Для нее система (6) с учетом (3) перепишется в виде

$$\frac{\partial\mathbf{H}(\mathbf{u},t)}{\partial t}=\mathbf{H}(\mathbf{u},t)\left[\mathbf{D}+\mathbf{D}_1\sum_{k=1}^{K}\left(e^{ju_k}-1\right)S_k(t)\right] \qquad (7)$$

Начальное условие для уравнения (7) имеет вид

$$\mathbf{H}(\mathbf{u},t_0)=\mathbf{r}. \qquad (8)$$

Прямое решение задачи Коши (7), (8) с использованием матричной экспоненты не представляется возможным, так как матрицы $\mathbf{D}$ и $\mathbf{D}_1$ не перестановочны. В связи с этим в настоящей работе выполнено исследование свойств его решения в асимптотическом условии неограниченно растущей интенсивности входящего потока, которое мы называем условием высокой интенсивности входящего потока. Для этого в уравнении (7) выполним следующие подстановки: $N\mathbf{D}_0$ вместо матрицы $\mathbf{D}_0$ и $N\mathbf{D}_1$ вместо $\mathbf{D}_1$, где N – большой по величине параметр (в теоретических исследованиях будем полагать $N\to\infty$). Получим:

$$\frac{1}{N}\frac{\partial\mathbf{H}(\mathbf{u},t)}{\partial t}=\mathbf{H}(\mathbf{u},t)\left[\mathbf{D}+\mathbf{D}_1\sum_{k=1}^{K}\left(e^{ju_k}-1\right)S_k(t)\right]. \qquad (9)$$

Интенсивность входящего MAP-потока, заданного таким образом составляет $N\lambda$ и неограниченно растет при увеличении значения параметра N.

5. Асимптотика первого порядка

Рассматривая условие высокой интенсивности входящего потока $N \to \infty$, выполним в задаче (9), (8) следующие замены:

$$\frac{1}{N} = \varepsilon, \quad \mathbf{u} = \varepsilon \mathbf{w}, \quad \mathbf{H}(\mathbf{u}, t) = \mathbf{F}_1(\mathbf{w}, t, \varepsilon). \tag{10}$$

С использоваием этих замен задача (9), (8) перепишется в виде

$$\varepsilon \frac{\partial \mathbf{F}_1(\mathbf{w}, t, \varepsilon)}{\partial t} = \mathbf{F}_1(\mathbf{w}, t, \varepsilon) \left[\mathbf{D} + \mathbf{D}_1 \sum_{k=1}^{K} \left(e^{j\varepsilon w_k} - 1 \right) S_k(t) \right], \tag{11}$$

$$\mathbf{F}_1(\mathbf{w}, t_0, \varepsilon) = \mathbf{r}. \tag{12}$$

Относительно асимптотического решения $\mathbf{F}_1(\mathbf{w}, t) = \lim\limits_{\varepsilon \to 0} \mathbf{F}_1(\mathbf{w}, t, \varepsilon)$ этой задачи имеет место следующее утверждение.

Теорема 1. *Асимптотическое решение* $\mathbf{F}_1(\mathbf{w}, t)$ *задачи* (11)–(12) *имеет вид*

$$\mathbf{F}_1(\mathbf{w}, t) = \mathbf{r} \exp \left\{ \lambda \sum_{k=1}^{K} j w_k \int_{t_0}^{t} S_k(\tau)\, d\tau \right\}, \tag{13}$$

где $\mathbf{r}$ *и* λ *определяются выражениями* (1) *и* (2).

Доказательство. Доказательство выполним в два этапа.

Этап 1. Положим в выражении (11) $\varepsilon \to 0$. Получим:

$$\mathbf{F}_1(\mathbf{w}, t)\mathbf{D} = \mathbf{0}.$$

Сравнивая это уравнение с (1), можем сделать вывод, что функция $\mathbf{F}_1(\mathbf{w}, t)$ может быть представлена в виде

$$\mathbf{F}_1(\mathbf{w}, t) = \mathbf{r}\Phi_1(\mathbf{w}, t), \tag{14}$$

где $\Phi_1(\mathbf{w}, t)$ – некоторая скалярная функция, удовлетворяющая условию

$$\Phi_1(\mathbf{w}, t_0) = 1. \tag{15}$$

Этап 2. Умножим обе части матричного уравнения (11) справа на вектор $\mathbf{e}$, поделим на ε, подставим (14) и выполним предельный переход $\varepsilon \to 0$. Тогда с учетом того, что

$$\mathbf{De} = \mathbf{0}. \tag{16}$$

Получим следующее дифференциальное уравнение:

$$\frac{\partial \Phi_1(\mathbf{w}, t)}{\partial t} \mathbf{re} = \Phi_1(\mathbf{w}, t) \mathbf{r} \mathbf{D}_1 \mathbf{e} \sum_{k=1}^{K} j w_k S_k(t).$$

Учитывая (2) и $\mathbf{re} = 1$, получаем:

$$\frac{\partial \Phi_1(\mathbf{w}, t)}{\partial t} = \Phi_1(\mathbf{w}, t)\lambda \sum_{k=1}^{K} j w_k S_k(t).$$

С учетом начального условия (15) получаем следующее решение этого уравнения:

$$\Phi_1(\mathbf{w}, t) = \exp\left\{ \lambda \sum_{k=1}^{K} j w_k \int_{t_0}^{t} S(\tau)\, d\tau \right\}.$$

Подставляя это выражение в (14), для функции $\mathbf{F}_1(\mathbf{w}, t)$ получаем выражение (13). Теорема доказана. ■

6. Асимптотика второго порядка

Введем векторную функцию $\mathbf{H}_2(\mathbf{u}, t)$, удовлетворяющую равенству

$$\mathbf{H}(\mathbf{u}, t) = \mathbf{H}_2(\mathbf{u}, t) \exp\left\{ N\lambda \sum_{k=1}^{K} j u_k \int_{t_0}^{t} S_k(\tau)\, d\tau \right\}. \tag{17}$$

Подставляя это выражение в (9), (8), получим следующую задачу Коши относительно функции $\mathbf{H}_2(\mathbf{u}, t)$:

$$\begin{cases} \dfrac{1}{N}\dfrac{\partial \mathbf{H}_2(\mathbf{u}, t)}{\partial t} + \mathbf{H}_2(\mathbf{u}, t)\lambda \displaystyle\sum_{k=1}^{K} j u_k S_k(t) = \\ \qquad = \mathbf{H}_2(\mathbf{u}, t) \left[\mathbf{D} + \mathbf{D}_1 \displaystyle\sum_{k=1}^{K} \left(e^{j u_k} - 1 \right) S_k(t) \right], \\ \mathbf{H}_2(\mathbf{u}, t_0) = \mathbf{r}. \end{cases} \tag{18}$$

Выполним здесь следующие замены переменных:

$$\frac{1}{N} = \varepsilon^2, \quad \mathbf{u} = \varepsilon \mathbf{w}, \quad \mathbf{H}_2(\mathbf{u}, t) = \mathbf{F}_2(\mathbf{w}, t, \varepsilon). \tag{19}$$

Тогда задача (18) примет вид

$$\begin{cases} \varepsilon^2 \dfrac{\partial \mathbf{F}_2(\mathbf{w}, t, \varepsilon)}{\partial t} + \mathbf{F}_2(\mathbf{w}, t, \varepsilon)\lambda \displaystyle\sum_{k=1}^{K} j\varepsilon w_k S_k(t) = \\ \qquad = \mathbf{F}_2(\mathbf{w}, t, \varepsilon) \left[\mathbf{D} + \mathbf{D}_1 \displaystyle\sum_{k=1}^{K} \left(e^{j\varepsilon w_k} - 1 \right) S_k(t) \right], \\ \mathbf{F}_2(\mathbf{w}, t_0, \varepsilon) = \mathbf{r}. \end{cases} \tag{20}$$

Обозначим

$$\mathbf{F}_2(\mathbf{w}, t) = \lim_{\varepsilon \to 0} \mathbf{F}_2(\mathbf{w}, t, \varepsilon). \tag{21}$$

Докажем следующее утверждение.

Теорема 2. *Асимптотическое решение* $\mathbf{F}_2(\mathbf{w},t)$ *задачи* (20) *имеет вид*

$$\mathbf{F}_2(\mathbf{w},t)=\mathbf{r}\exp\left\{\lambda\sum_{k=1}^{K}\frac{(jw_k)^2}{2}\int\limits_{t_0}^{t}S_k(\tau)\,d\tau+\kappa\sum_{k=1}^{K}\sum_{\nu=1}^{K}\frac{jw_kjw_\nu}{2}\int\limits_{t_0}^{t}S_k(\tau)S_\nu(\tau)\,d\tau\right\},\tag{22}$$

где

$$\kappa=2\mathbf{g}(\mathbf{D}_1-\lambda\mathbf{I})\mathbf{e},\tag{23}$$

а вектор-строка $\mathbf{g}$ *удовлетворяет следующей системе линейных алгебраических уравнений:*

$$\mathbf{gD}=\mathbf{r}(\lambda\mathbf{I}-\mathbf{D}_1).\tag{24}$$

Доказательство. Доказательство выполним в три этапа.

Этап 1. Выполним в (20) предельный переход $\varepsilon\to 0$, получим:

$$\begin{cases}\mathbf{F}_2(\mathbf{w},t)\mathbf{D}=\mathbf{0},\\ \mathbf{F}_2(\mathbf{w},t_0)=\mathbf{r}.\end{cases}$$

Сравнивая этот результат с (1), можно сделать вывод о том, что функция $\mathbf{F}_2(\mathbf{w},t)$ может быть представлена в виде

$$\mathbf{F}_2(\mathbf{w},t)=\mathbf{r}\Phi_2(\mathbf{w},t),\tag{25}$$

где $\Phi_2(\mathbf{w},t)$ – некоторая скалярная функция, удовлетворяющая условию

$$\Phi_2(\mathbf{w},t_0)=1.\tag{26}$$

Этап 2. Учитывая (25) и (21), функцию $\mathbf{F}_2(\mathbf{w},t,\varepsilon)$ можно представить в виде разложения

$$\mathbf{F}_2(\mathbf{w},t,\varepsilon)=\Phi_2(\mathbf{w},t)\left[\mathbf{r}+\mathbf{g}\sum_{k=1}^{K}j\varepsilon w_kS_k(t)\right]+\mathbf{O}(\varepsilon^2),\tag{27}$$

где $\mathbf{g}$ – некоторая вектор-строка.

Подставим (27) и разложение $e^{j\varepsilon w_k}=1+j\varepsilon w_k+O\left(\varepsilon^2\right)$ в уравнение задачи (20), получим равенство:

$$\mathbf{r}\lambda\sum_{k=1}^{K}j\varepsilon w_kS_k(t)=\mathbf{rD}+\mathbf{rD}_1\sum_{k=1}^{K}j\varepsilon w_kS_k(t)+\mathbf{gD}\sum_{k=1}^{K}j\varepsilon w_kS_k(t)+\mathbf{O}(\varepsilon^2).$$

Выполнив здесь предельный переход $\varepsilon\to 0$, с учетом (1)получаем следующее матричное уравнение относительно неизвестного вектора $\mathbf{g}$:

$$\mathbf{gD}=\mathbf{r}(\lambda\mathbf{I}-\mathbf{D}_1),$$

которое совпадает с (24).

Этап 3. Умножим обе части дифференциального уравнения задачи (20) справа на вектор $\mathbf{e}$. Используя разложение $e^{j\varepsilon w_k} = 1 + j\varepsilon w_k + \frac{(j\varepsilon w_k)^2}{2} + O(\varepsilon^3)$, учитывая свойства (1), (16) и обозначения (2), (23), выполним в этом уравнении предельный переход при $\varepsilon \to 0$. В результате получим следующее линейное дифференциальное однородное по t уравнение относительно функции $\Phi_2(\mathbf{w}, t)$:

$$\frac{\partial \Phi_2(\mathbf{w}, t)}{\partial t} = \Phi_2(\mathbf{w}, t) \left[\lambda \sum_{k=1}^{K} \frac{(jw_k)^2}{2} S_k(t) + \kappa \sum_{k=1}^{K} \sum_{\nu=1}^{K} \frac{jw_k jw_\nu}{2} S_k(t) S_\nu(t) \right],$$

где величина κ определяется выражением (23). С учетом начального условия (26) получаем следующее решение этого уравнения:

$$\Phi_2(\mathbf{w}, t) = \exp \left\{ \lambda \sum_{k=1}^{K} \frac{(jw_k)^2}{2} \int_{t_0}^{t} S_k(\tau) d\tau + \kappa \sum_{k=1}^{K} \sum_{\nu=1}^{K} \frac{jw_k jw_\nu}{2} \int_{t_0}^{t} S_k(\tau) S_\nu(\tau) d\tau \right\}.$$

Подставляя полученное решение в (25), получаем окончательное выражение для функции $\mathbf{F}_2(\mathbf{w}, t)$ в виде (22). Теорема доказана. ■

7. Стационарное распределение вероятностей числа заявок в системе

Выполним в (22) замены, обратные к (19). Учитывая (17), получаем следующее выражение для векторной характеристической функции $\mathbf{H}(\mathbf{u}, t)$ многомерного распределения числа событий в просеянных потоках, наступивших до момента времени t (см. раздел 3):

$$\mathbf{H}(\mathbf{u}, t) = \mathbf{r} \exp \left\{ N\lambda \sum_{k=1}^{K} \left[ju_k + \frac{(ju_k)^2}{2} \right] \int_{t_0}^{t} S_k(\tau) d\tau + \right.$$

$$\left. + N\kappa \sum_{k=1}^{K} \sum_{\nu=1}^{K} \frac{ju_k ju_\nu}{2} \int_{t_0}^{t} S_k(\tau) S_\nu(\tau) d\tau \right\}.$$

Вернемся к исследуемому процессу $\mathbf{i}(t)$, который представляет число заявок на фазах рассматриваемой многофазной системы массового обслуживания. С помощью $h(\mathbf{u})$ обозначим характеристическую функцию сечения этого процесса в момент времени T. Применяя основную формулу (4) многофазного динамического просеивания, получаем следующую аппроксимацию $h^{(2)}(\mathbf{u})$ для характеристической функции $h(\mathbf{u})$ при условии достаточно больших значений параметра N:

$$h(\mathbf{u}) \approx h^{(2)}(\mathbf{u}) = \exp \left\{ N\lambda \sum_{k=1}^{K} \left[ju_k + \frac{(ju_k)^2}{2} \right] \int_{t_0}^{T} S_k(\tau) d\tau + \right.$$

$$+N\kappa\sum_{k=1}^{K}\sum_{\nu=1}^{K}\frac{ju_k ju_\nu}{2}\int_{t_0}^{T}S_k(\tau)S_\nu(\tau)d\tau\Bigg\}.$$

Рассматривая начальный момент $t_0 \to -\infty$ и выполнив в интегралах замену $t = T - \tau$, получаем следующую аппроксимацию для характеристической функции $h(\mathbf{u})$ числа заявок на фазах системы в стационарном режиме функционирования:

$$h(\mathbf{u}) \approx h^{(2)}(\mathbf{u}) = \exp\left\{j\mathbf{u}N\lambda\mathbf{S}\mathbf{e} + \frac{1}{2}j\mathbf{u}\left[N\lambda\mathbf{S} + N\kappa\mathbf{V}\right]j\mathbf{u}^T\right\}, \qquad (28)$$

где $\mathbf{S}$ – диагональная матрица с элементами главной диагонали, равными средним временам обслуживания на соответствующих фазах, а матрица $\mathbf{V}$ составлена из элементов $V_{k\nu} = \int\limits_0^\infty \left[B^*_{k-1}(t) - B^*_k(t)\right]\left[B^*_{\nu-1}(t) - B^*_\nu(t)\right]dt$, $k, \nu = \overline{1, K}$.

Таким образом, многомерное стационарное распределение числа заявок на фазах многофазной системы обслуживания с входящим MAP-потоком, неограниченным числом приборов и произвольным обслуживанием на фазах системы при достаточно большой интенсивности входящего потока может быть аппроксимировано многомерным нормальным распределением с вектором средних $N\lambda\mathbf{S}\mathbf{e}$ и матрицей ковариаций $N\left[\lambda\mathbf{S} + \kappa\mathbf{V}\right]$.

Для определения точности полученной гауссовской аппроксимации проведен ряд численных экспериментов, в каждом из которых для маргинальных стационарных распределений вероятностей числа заявок на фазах системы закон, составленный на основе нормального распределения с соответствующими параметрами, сравнивался с эмпирическим распределением, построенным на основе результатов имитационного моделирования. Для проведения сравнения использовалось расстояние Колмогорова [6], которое для дискретных распределений имеет вид

$$d = \max_{i\geq 0}\left|\sum_{l=0}^{i}\left[\tilde{P}(l) - P(l)\right]\right|,$$

где $P(l)$ – маргинальное стационарное распределение числа заявок на одной фазе системы, вычисленное на основе аппроксимации (28), $\tilde{P}(l)$ – маргинальное эмпирическое распределение числа заявок на соответствующей фазе, построенное на основе результатов имитационного моделирования.

Установлено, что полученная аппроксимация дает хорошие результаты (расстояние Колмогорова $d < 0,05$) при значениях параметра $N \geq 30$.

8. Заключение

В работе представлено исследование многофазной системы массового обслуживания с входящим MAP-потоком, неограниченным числом приборов и произвольным обслуживанием на фазах. Исследование выполнено в

асимптотическом условии неограниченно растущей интенсивности входящего потока. Показано, что при достаточно большой интенсивности входящего MAP-потока многомерное распределение вероятностей числа заявок на фазах системы в стационарном режиме с достаточной точностью может быть аппроксимировано многомерным нормальным распределением. В работе получены значения вектора математических ожиданий и матрицы ковариаций для этой гауссовской аппроксимации. Численные эксперименты с применением имитационного моделирования позволяют определить область применимости полученной аппроксимации.

ЛИТЕРАТУРА

1. Бочаров П. П., Печинкин А. В. Теория массового обслуживания: Учебник. М.: Изд-во РУДН, 1995.
2. Грачев В. В., Моисеев А. Н., Назаров А. А., Ямпольский В. З. Многофазная модель массового обслуживания системы распределенной обработки данных // Доклады ТУСУРа. 2012. № 2 (26), часть 2. С. 248–251.
3. Genadis T. The distribution of the passage time in a twoЦstation reliable production line: an exact analytic solution // International Journal of Quality and Reliability Management. 1997. V. 14, Iss. 9. P. 929–935.
4. Chakravarthy, S. R. Markovian arrival processes. Wiley Encyclopedia of Operations Research and Management Science, 2010.
5. Моисеев А. Н., Назаров А. А. Асимптотический анализ многофазной системы массового обслуживания с высокоинтенсивным рекуррентным входящим потоком // Автометрия. 2014. Т. 50, № 2, С. 67–76.
6. Рыков В. В., Иткин В. Ю. Математическая статистика и планирование эксперимента: уч. пособие. М.: МАКС Пресс, 2010.

ONE AND TWO STAGE SHORT RANGE DRIVE-THRU VEHICLE NETWORKS PERFORMANCE EVALUATION

E. Petersons, N. Bogdanovs, A. Ipatovs

Department of Transport Electronics and Telematics, Riga Technical University, Riga, Latvia

ernestspetersons@yahoo.com, Nikolajs.Bogdanovs@rtu.lv, Aleksandrs.Ipatovs@riga.lv

Abstract

In this paper were presented experimental data and model for evaluating the data transmission speed between the remote moving object-vehicle and base station in wireless networks 802.11n and 4G-LTE standarts. Network of 802.11n standart represent first stage of overall two stage network, involving 4G-LTE channel. The experimental resalts are used for evaluation of actual speed of data transmission between the moving vehicle and base station by using IxChariot program. Network Goodput depending on vehicles speed and vehicles density on road are represented too.

Keywords: 802.11n, 4G-LTE standarts, model, Goodput

1. Introduction

Short-range vehicle-roadside or V2I communication is expected to be part of the future intelligent transportation system (ITS) in order to increase the safety of the roads and efficiency of the traffic. Therefore, investigation on proper communication for ITS are increasing. Today, the IEEE 802.11n and 802.11ac standard for high data rate wireless networks is widespread and costs effective. Extension of this standard could be a part of V2I communication technology.

IEEE 802.11p standard was specially developed for Short Range Drive-thru vehicle Networks, but it use 5.85-5.925 GHz frequencies. These frequencies are paid and the equipment is not cheap. Unlike the specially developed standard 802.11p allows transferring short official reports of urgent character only. Protocols 802.11g/n/ac allow large-scale data transfer at high speeds, and passing to free range of frequencies. The problem is – it is necessary to transfer large-scale data in short time intervals during movement of vehicle. The losses appear during transferring from zone to zone.

Nowadays main wireless system standard is the 802.11n. This standard replaces the 802.11g standard. Significant improvement of quality of signal and goodput also should be observed in utilization of this standard in the Drive-thru Internet system. To prove this fact and to estimate improvements some experiments were conducted, the results and the analysis are described below.

We want once again to refer about usefulness of utilization of the 802.11n standard in Drive-thru Internet system, unless the utilization of specially designed 802.11p standard:

- Equipment for the 802.11n standard is cheap and available.
- Frequencies of the 802.11n standard are free of charge.
- The 802.11p standard does not provide transfer of data large amount and it is developed for transfer of short messages.

All these advantages indicate on feasibility of utilization of the 802.11n standard in the Drive-thru Internet system.

The uniqueness of this research is development of real Drive-thru Internet system for experiments. Experiments for the 802.11n standard were conducted in polygon Rumbula in Riga. Inter-connected roadside access points (APs) ASUS RT-N16 with firmware dd-wrt.v24 were used as workstations. This firmware supports WDS.

In plain terms, this technology allows the access points to establish wireless connection not only with wireless customers, but also between themselves, extending the zone of wireless net action. The main advantage to such net is that its access points are interconnected, comprising a drive-thru Internet system. In this case, there is no need to use landline nets for connection of access points. The net constructed using WDS technology allows the mobile stations to switch from one access point to another without losing connection with wireless net. Figure 1 shows the scheme of the Inter-connected roadside access points (APs) interconnection via WDS.

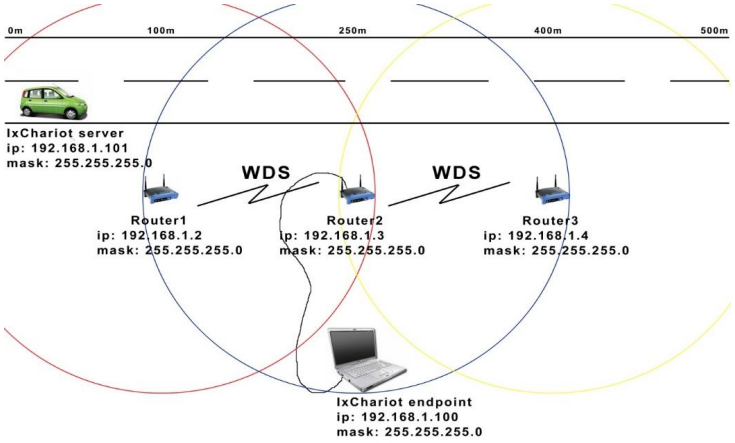


Figure 1: Connection scheme of Inter-connected roadside access points (APs) and zones of their activity.

The graph in Figure 2 shows goodput dependence on different speeds of Drive-thru vehicle (20km/h and 100km/h) with 802.11n standard. As WDS was used significant loss can be seen in left and right zones of the graph.

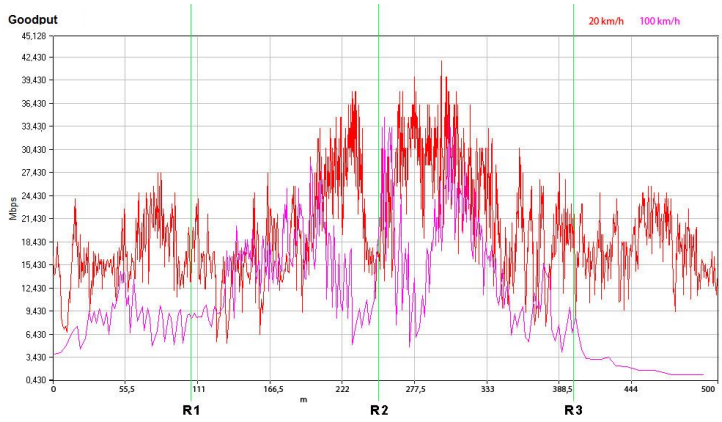


Figure 2: Analysis of Goodput (Mbps) at different speeds (20km/h and 100km/h) in short range drive-thru vehicle network with 802.11n standard.

The graph in Figure 3 shows goodput dependence on speeds of drive-thru vehicle (40km/h, 60km/h and 80km/h) with 802.11n standard.

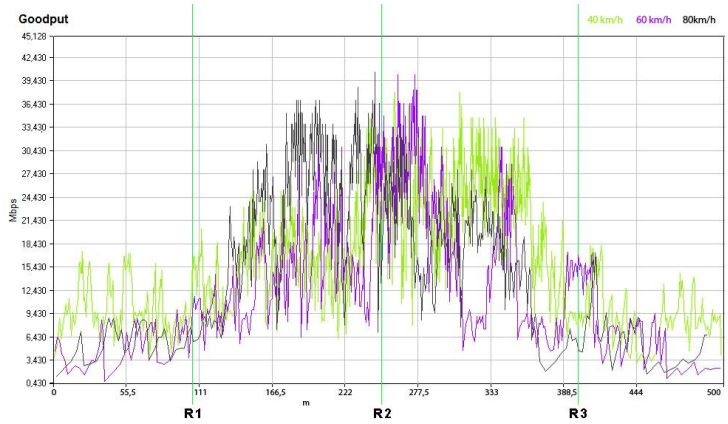


Figure 3: Analysis of Goodput (Mbps) at different speeds (40km/h, 60km/h and 80km/h) in short range drive-thru vehicle network with 802.11n standard.

Analyzing these graphs, it can be concluded that by increasing speed of the drive-thru vehicle, goodput reduced slightly, but at the same time it remains stabile. Significant growth of productivity can be seen not only on main Inter-connected roadside access point (APs), but also on secondary Inter-connected roadside access point (APs). Transitional processes during switching between base stations are not seen practically.

Main advantage of the 802.11n standard is stability of data transfer in Drive-thru Internet system at high speed of drive-thru vehicles.

Using Inter-connected roadside access points (APs) with the 802.11n standard, transition between workstations is performed without fading. Average goodput on all stages was 9,838 Mbps.

Figure 4 shows average goodput dependence on 802.11n at different speeds.

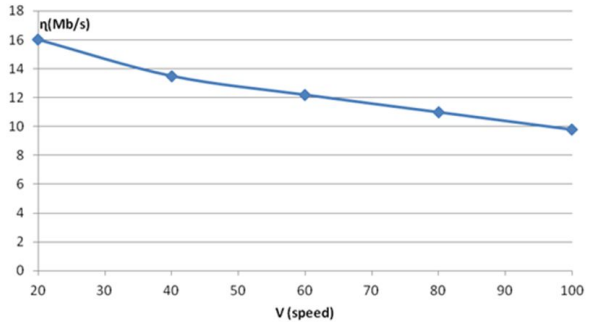


Figure 4: Average goodput dependence on 802.11n at different speeds.

2. Performance Evaluation of Vehicular Heterogeneous Wireless Networks

In second part of report, two-stage real data network and model of network are presented. In this case, differently from first part, server is placed fare away from position of AP concentrator. Therefore additional data transmission channel is solved in network. In our experiments LTE channel has been used.

The physical realization of communication for the transmission data from the vehicle to the user's server and back is the wireless network. At the first stage the data are transmitted from the mobile object to the nearest Access Point according to the protocol 802.11n.

However, the distance from the AP object should not exceed 200 meters. Further, from the AP the data are transmitted to the remote base station (server) by the channel according to the LTE. This variant provides the data transmission at the distance up to the several kilometers. Thus, the object of the research represents the two-stage system of the wireless networks. This system can be represented by the two-stage network model, as it is shown in Figure 5. The null node stimulates the data transmission from the movable object with the intensity of the data transmission ε_0.

The second node stimulates the AP wireless network providing the data reception and transmission from the mobile objects of the null node. The intensity of the data processing is equal to ε_1.

$$\varepsilon_1 = \beta_i, \tag{1}$$

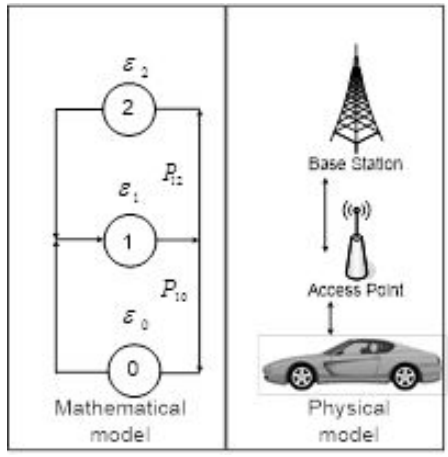


Figure 5: Two-stage vehicular network model.

where β_i – transmission rating in the wireless network 802.11n depending on the distance to the base station as it is shown in the Figure 5.

In order to determine the model parameters, such as the intensity of processing in nodes and transition probability, it is necessary to evaluate the physical parameters of prototype. For the purpose of creation of such prototype was used the equipment of the company Cisco. Using Cisco equipment was built the wireless two rank transport networks.

This prototype represents 'test-bed' for the research of dependence goodput on the travel speed of the vehicle. Moreover, the measurement of useful data transmission rate covers not only the first rank of the system: "mobile object" – AP and further the data transmission channels from AP to the remote server of the user.

Naturally, that the Goodput to a great extent will depend on the data transmission rate from the AP to server, i.e., from the transmission data characteristics. To carry out such researches has been taken the router CISCO C819 M2M which has two output channels. One channel provides the data transmission in GPRS mode. The second channel, being characterized by a high data transfer rate, uses LTE mode — the mode of the next generation of mobile communications. Scheme "test-bed" is presented below (see Figure 6:

The main task in this experiment is the approximate data transmission speed in accordance with the distance to the base station, as well as the second task should be resolved, when the Internet speed should be fixed in accordance with the N moving objects, which are located in the coverage of the wireless network base station.

In order to carry out such study, the program IxChariot should be used. Over the FTP protocol the file is being sent via base stations from the computer to the remote server moving along the base station. FTP protocol is being used to transfer large amounts of data. During the experiment, the actual speed Goodput will be measured

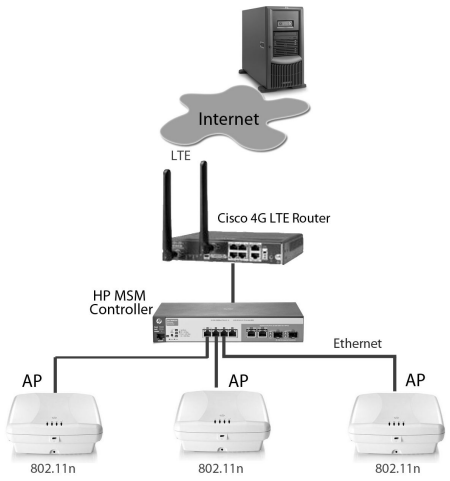


Figure 6: Two-stage network model.

(the recommended data transmission speed) in the program over FTP protocol in order to determine the data transmission speed at the speed of the certain vehicle.

The first experiment was carried out with the use of access point in each direction at the distance of 100 meters from the controller in accordance with the principle as shown in the scheme. This means that the access points are located at the distance of 100 meters from one another, we use three access points and the client was moving with a speed of 20 km/h. It should be mentioned that these experiments have no authority controllers. Goodput - permeability and Elapsed time are denoting the experimental measurement time (see Figure 7).

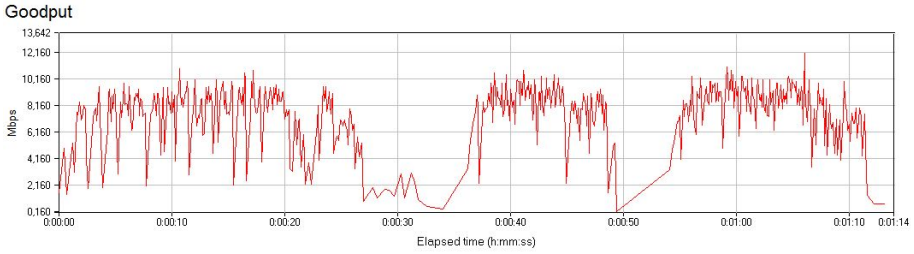


Figure 7: Two-stage goodput with 4G and 802.11n at the speed of 20km/h.

Response time from the device, seconds and Elapsed time are denoting the experimental measurement time (see Figure 8).

The next experiment was carried out at the speed of 50 km/h (Figure 9), traveling down the road along the access points. The distance between the access points is 100 meters. The Figure 10 shows the goodput at the speed of 90 km/h.

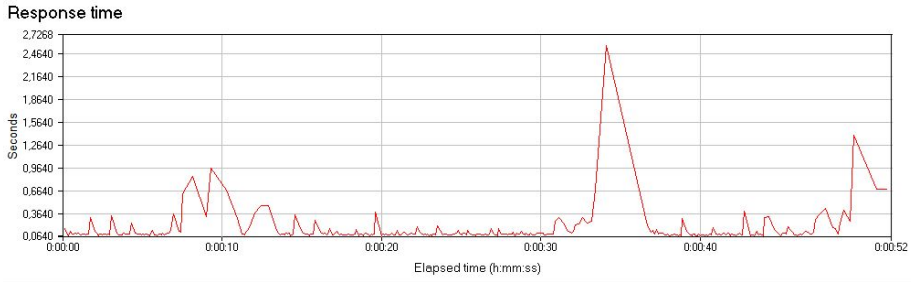


Figure 8: Response time at the client travel speed of 20 km/h along the access points.

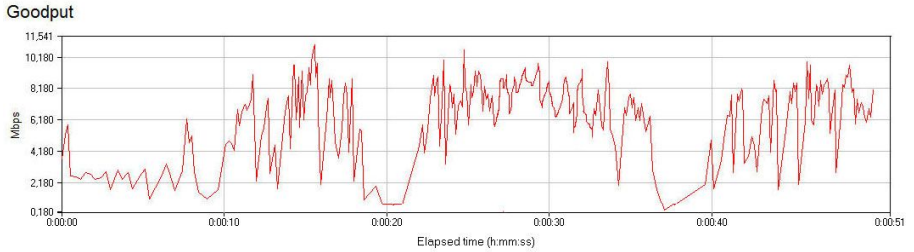


Figure 9: Two-stage goodput with 4G and 802.11n at the speed of 50km/h.

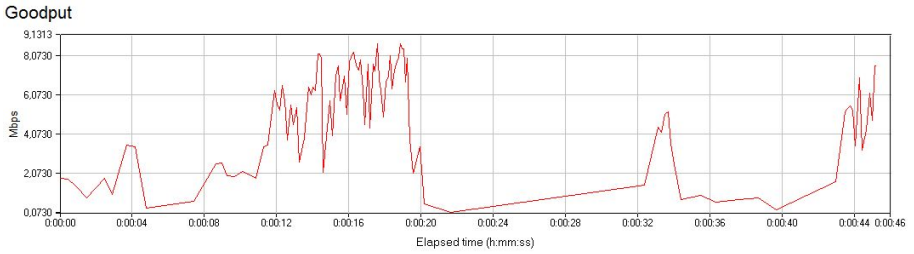


Figure 10: Two-stage goodput with 4G and 802.11n at the speed of 90km/h.

After experimental evaluation main parameters of network, let us return to another ones in model. Depending on the vehicle's proximity to the base station, the data processing rate and data processing intensity in the base station will be different. For the file transfer with packet length of 1500 bytes, the base stations goodput β_i was estimated experimentally.

In its turn, AP is connected with the remote base station along the wireless network with the LTE. The intensity of the data transmission of the second node is to be taken to be equal to ε_2.

The route of the data transmission keeps the track from the null node to the first node and then to the second one [1], if the file transfer is considered from the vehicle

to the BS. From the BS it is transmitted the ACK confirmations on the packet's transmission. In this case, the average time for the transmission will be different: more time is spent on the transmission of the data packets, which is denoted as $E(t_i)$

The ACK transmission takes less time and denotes it as $E(t_0)$. Then the average time of the data processing in the zero node will be:

$$E(t_1) = \frac{E(t_i) + E(t_0)}{2}. \tag{2}$$

If on the top of each transmitted packet the ACK confirmations are received, the intensity of the processing in the zero node will be:

$$\varepsilon_0 = \frac{1}{E(t_1)} \tag{3}$$

The model is present in the parameter N, determining the number of the data transmission initiators, which compete for the resource sharing of the 1 and 2 nodes [2]. In our case this is the number of automobiles in the AP coverage area. Then the three-node and two-stage model of the goodput can be expressed by the (12) formula. In this formula the parameters a, X_1 and X_2 are determined by the value from (4).

The valuation problem of the goodput provided by the model consists of the determination of the value N -– the number of vehicles in the AP coverage area. Moreover, in the wireless network standard 802.11n the speed of data transmission depends on the remoteness of the vehicles from AP. The terminal count in each vehicular wireless network is usually high [3, 4]. The bandwidth equation for a two-stage network being:

$$X_1 = \frac{\varepsilon_0}{\varepsilon_1 P_{10}}; X_2 = aX_1; a = \frac{\varepsilon_1}{\varepsilon_2} P_{12}. \tag{4}$$

The intensity for the ε_2:

$$\varepsilon_2 = \frac{1}{t} \tag{5}$$

$$t = \frac{l_p}{V_f} \tag{6}$$

Where V_f – the effective data transfer rate for the LTE. For the having the peak transfer rate V_n. The actual speed is determined in the following way:

$$V_f = \frac{V_n}{2} \tag{7}$$

The starting point for the calculation is the normalizing function $G(N)$, that is chosen from the principle of the sum of probabilities being one $p(n_0, n_1, n_2)$, where n_i in vector $\bar{n} = (n_1, n_{2,3})$ is the inquiry count in i-th node. The resulting equation for $G(N)$ calculation looks like this:

$$G(N) = \sum_{\bar{n}} \prod_{i=1}^{3} (X_i)^n i, \tag{8}$$

$$\bar{n} \in \left\{ n_1, n_2, n_3) / \sum_{i=1}^{3} n_i = N, n_i \geq 0 \forall i \right\}. \tag{9}$$

Where N – the number of vehicles.

The function of the studied two layer vehicular network looks like this:

$$G(N) = \frac{1}{1-a} \sum_{j=0}^{N} X_1^j (1 - a^{j+1}) \tag{10}$$

The Goodput η of the two-stage network is defined as the count of the processed inquiries in per unit of time [4, 5]. The finished task is put out trough the subsystem of input/output, and instantly a new task is loaded through it. The probability of a lack of inquiries in the i-th node will be [6, 7]:

$$p\{n_i = 0\} = \frac{G(N) - X_i G(N-1)}{G(N)} \tag{11}$$

The result is:

$$\eta = P_{10} \varepsilon_1 (1 - p\{n_i = 0\}) \tag{12}$$

The particular importance is paid to the assessment of parameter N – the number of "tasks" or requests, circulating in the network of communication [4, 11]. Taking into the consideration the fact that the number of mobile customers in the coverage area of the base station WiFi is limited or fixed, and then the model corresponds to the class of closed networks of queuing system [9, 10].

If, the average number of customers N_n, who present in the coverage area of the base station and the average number of packages Z received or transmitted by the customer in the base station can be determined and the number of inquiries in the network is:

$$N_n = N \cdot Z \tag{13}$$

For the determination the number of customers in the coverage area of the base station will be consider the highway on which there is a high way flow of cars – the worst case for the researched network [12]. Such high rate flow according refers to the class of mobile objects, following for the leaders. It has been affirmed that the distribution of time intervals for the consecutive vehicles can be approximated by the Erlang distribution of $(k-1)$ order.

According to an experimental data this value lies within the range of 5 to 7 . The intensity of input flow of vehicles:

$$\lambda_2 = \frac{1}{\overline{h_2}}, \tag{14}$$

where $\overline{h_2}$ – is the average value of the interval between the vehicle and the distribution of intervals itself between vehicles:

$$f(h_2) = \frac{(K\lambda_2)^K \cdot h_1^{K-1} \cdot e^{-\lambda_2 K h_2}}{(K-1)!}. \tag{15}$$

Erlang law of $k-1$ law is formed as a chain of stages on which occur the delays of requests, distributed exponentially with parameters $\lambda_2 K$.

The intensity of traffic of automobiles, passing through any given point of the road per unit time [8], including the location point of the base station constitutes:

$$q = K\vartheta. \tag{16}$$

According to some experimental data the intensity of traffic is 1800-1900 of vehicles/per hour for traffic lane [13]. Naturally, if the number of bands in one direction, then the service rate of vehicles will be z times more, i.e.

$$q_z = K \cdot z. \tag{17}$$

At that it is necessary to take into the consideration the density of vehicles per unit of length of band [14]. In this case the unit of length constitutes the diameter of coverage area of the base station – 200 meters.

The given initial data allow coming to the estimation of the average number of vehicles on the road [15]. For this we assume that the ingoing stream of automobiles Erlang $k-1$ order and $k = 5$. The service time of flow of vehicles of the road with the bands is a random value of distribution on exponentially law with the parameter q_z.

For the evaluation the number of vehicles which are in the coverage area of the road, use the $E_k/M/1$ model. According to the number of customers in the system of service, i.e., the average number of vehicles in segment of road:

$$L = \frac{\rho}{1 - r_0^k}. \tag{18}$$

Where ρ – the load ratio of the system.

In our case:

$$\rho = \frac{\lambda_2}{q_z}. \tag{19}$$

At that there r_0 is a root of solution of characteristic equation which lies in the interval 0,1.

$$q_z \cdot r^{k+1} - (k \cdot \lambda_2 + q_z)r + k\lambda_2 = 0. \tag{20}$$

In the considered situation $k = 5$. Note that r_0 is comparable with ρ, then the solution (20) for r_0 is sought iteratively. In the Table 1 are presented the results of

ρ	0.25	0.5	0.6	0.75	0.8	0.9
r_0	0.56	0.76	0.84	0.88	0.9	0.92
$L = N$	0.265	0.67	1.031	1.588	1.954	2.64

Table 1: The number of vehicles on the road section.

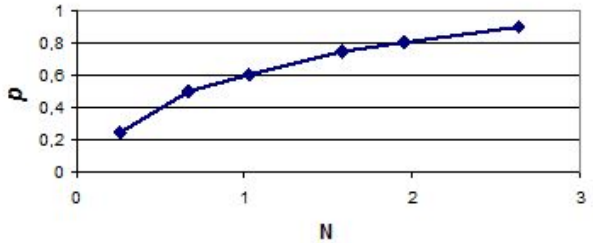


Figure 11: The number of vehicles with the different load factor ρ.

solution of equation for the different load factors as well as the resulting indices of number of vehicles on the road section included in the service area of the base station (see Figure 11).

From the equation (12) for each segment can be calculated. The network performance influences the probability of transmission of the confirmation ACK, as increases, the number of packages per unit of time increases too.

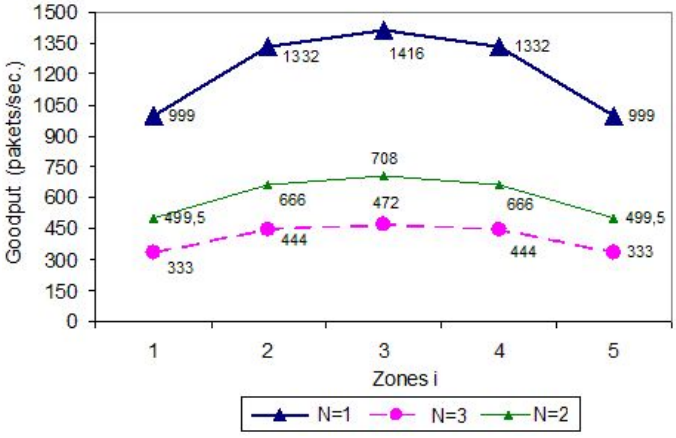


Figure 12: Goodput for two-stage network model with $P_{10} = 0.999$, $P_{12} = 0.001$ for 802.11n and LTE at the Erlang distribution.

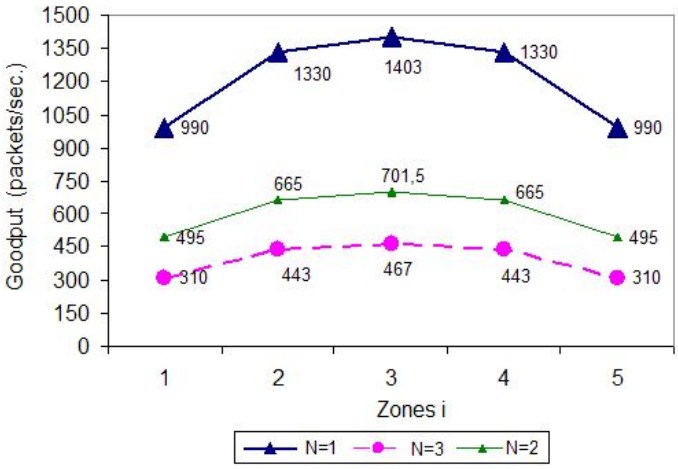


Figure 13: Goodput for two-stage network model with $P_{10} = 0.99$, $P_{12} = 0.01$ for 802.11n and LTE at the Erlang distribution.

3. Conclusion

In fact, we have derived practical analytical models for the distribution of the number of packets that a vehicle can download from a two-stage network system with access point and base station. In this work was developed the model to determine the actual speed of date transmission, depending on the number of mobile objects.

On the basis of obtained data it is possible to conclude that the performance of the base station is connected both with the traffic parameters and data transmission feature. In this paper were developed the model for the determination of the actual speed of data transmission, depending on the number N of the mobile objects, which are in the coverage area of the base station of the wireless network. On the basis of the paper the actual rate of datatransmission will depend on the number of objects, interacting with the base station and their remoteness from it.

REFERENCES

1. Lagershausen S. Performance Analysis of Closed Queueing Networks, Springer. 2013. pp.5-26.
2. B.S.Kerner, Introduction to Modern Traffic Flow Theory and Control, Publisher: Springer (2009), 265 p.
3. Ipatovs, Aleksandrs. Experimental and Analytical Goodput Evaluationof Drive-thru Internet Systems: promocijas darbs / A.Ipatovs zinatniskais vaditajs E.Petersons ; Rigas Tehniska Universitate. ELEKTRONIKAS UN TELEKOMUNIKACIJU FAKULTATE. Transporta elektronikas un telematikas katedra. Riga: [RTU], 2012. 127lp.
4. D. Gross, J. Shortle, J. Thompson, C. Harris. Fundamentals of Queuering Theory. Willy. 2008, 500 p.

5. Garber N. J., Hoel L. A. Traffic & Highway Engineering, Cengage Learning. 2009. 1230 p.
6. R. Jain and J. M. Smith, “Modeling vehicular traffic flow using m/g/c/c state dependent queueing models,” Transportation Science, vol. 31, pp.324–336, 1997.
7. I.A. Ismail, G.S. Mokaddis, S.A. Metwally and Mariam K. Metry, Optimal Treatment of Queueing Model for Highway,Journal of Computations & Modelling, vol.1, no.1, 2011, pp 61-71,ISSN: 1792-7625
8. J. P.Singh, N. Bambos, B. Srinivassan and D. Clawin, Wireless LAN performance under varied stress conditions in vehicular traffic scenarios, proceedings of Vehicular Technology Conference, (2002), Vol. 2, pp. 24-28.
9. K. K. Leung, W. A. Massey, and W. Whitt, “Traffic models for wireless communication networks,” IEEE J. Sel. Areas Commun., vol. 12, pp. 1353–1364, 1994.
10. Window-based rate adaptation in 802.11n wireless networks. Pefkianakis I., Hu Y., Lee S.B. bez viet. : Springer, 2012. DOI: 10.1007/s11036-011-0347-x.
11. A. Matsumoto, K. Yoshimura, S. Aust, T. Ito, Y. Kondo, Performance evaluation of IEEE 802.11n devices for vehicular networks, LCN 2009, The 34th Annual IEEE Conference on Local Computer Networks, LCN 2009, 20-23 October 2009, Zurich, Switzerland, Proceedings (2009), 669-670.
12. V. Bychkovsky , B. Hull , A. Miu , H. Balakrishnan and S. Madden ”A measurement study of vehicular Internet access using in situ Wi-Fi networks”, Proc. ACM MobiCom, 2006, pp.50 –61.
13. Luan, T.H., Xinhua Ling, Xuemin Shen. MAC in Motion: Impact of Mobility on the MAC of Drive-Thru Internet. Mobile Computing, IEEE Transactions on Volume: 11, Issue: 2, 2012, pp. 305-319
14. Marc Emmelmann, Bernd Bochow, C. Christopher Kellum: Vehicular Networking Automotive Applications and Beyond. A John Wiley and Sons, Ltd, Publication, 2010, pp. 227-255.
15. Jorg Ott and Dirk Kutscher, “Drive-thru Internet: IEEE 802.11b for „Automobile“ Users,” in Proceedings of the IEEE Infocom 2004 Conference, Hong Kong, March 2004.

A CLUSTER CACHING RULE IN NEXT GENERATION NETWORKS

N. Markovich

V.A. Trapeznikov Institute of Control Sciences of Russian Academy of Sciences, Moscow, Russia
nat.markovich@gmail.com, markovic@ipu.rssi.ru

Abstract

Probabilistic aspects of caching are considered. The caching serves to keep popular contents inside a memory unit called 'cache' to be able to access them quickly. Using extreme value theory we propose a caching strategy called Cluster Caching Rule driven by content popularity that may change in time. A non-Poisson request arrival process is used when requests are statistically correlated. The idea of the new approach is to locate in cache only contents whose popularity exceeds a sufficiently large threshold. Due to dependence and possible heavy-tailed distribution of inter-requests and inter-arrival times of documents, the popularity process builds clusters of exceedances. The cluster and inter-cluster sizes are geometrically distributed as derived in Markovich (2014). We use it to calculate means of the cache utilization and occupancy. We escape assumptions like a constant size of content and a Poisson request process that are typical in the literature.

Keywords: Caching, content, popularity process, cluster of exceedances, cache utilization, cache occupancy

1. Introduction

The paper is devoted to probabilistic aspects of caching. A caching policy serves to keep popular contents inside a short memory unit called 'cache' to be able to access them as soon as possible. We consider it from new perspectives using achievements of extreme value theory.

Let C be the size of a cache and $d_1, ..., d_M$ be a catalog of documents. Since the popularity of contents may change over time and place, the hitting of documents inside the cache has to be changed, too. Cache filling depends on the input (the arrival) process of document requests, the cache size and the replacement policy. The latter may be controled by the time that is allowed for the document to be in the cache. This called TTL (the Time-to-Live) may be individual for each document depending on its popularity [1] or cache dependent [2], or the TTL may be fixed, [3]. If a requested document is found in the cache then the cache content remains unchanged. Then we say that the document request hits the cache. If the document is new, i.e. it cannot be found in the cache, then the missing content is brought in from the outside world (a long memory). In this case, we say that the document request misses

the cache. The previous cache content may remain unchanged in this case too if the cache was not full. Otherwise, the new document evicts some content from the cache according to the replacement rule.[1]

There are several identification problems in the caching such as size and location of caches, statistical estimation of the content popularity, a convenient modeling of the inter-request time (IRT) process, the optimality evaluation of the replacement rule, e.g., by hit/miss probabilities as well as the occupancy and the utilization of the cache. The most popular replacement rule is the Least-Recently-Used (LRU) caching, [3], [4], where the requested document hits the first position of the cache. The object at the last position may leave the cache if the requested document is new or it stays longer otherwise.

Despite of an easy idea, the analytical analysis of the LRU is difficult. Usually, it is assumed for simplicity that the content size is deterministic despite it is naturally random. However, one can combine several content units into chunks of a constant size. Such chunks are assumed to be claimed.

The requesting point process of the object is often assumed to be a renewal Poisson process and the IRTs are then independent exponentially distributed, see [1], [4] among others. A superposition of renewal Poisson processes is also renewal Poisson. This simplifies the use of cache trees where superposed request processes from leaf-caches arrive to the root-cache, [1]. Generally, the superposition of two renewal processes is not a renewal process. Markov modulated processes are closed under superposition. Hence, Markov and semi-Markov modulated processes are typically used as alternative models of arrival request processes, [1], [5], [6]. In contrast to renewal Poisson processes such superposed request processes are correlated.

Usually, the probability of the content popularity is modeled by Zipf's law that has a Pareto tail, [7], [8]. In [9] the mixture of Zipf and heavy-tailed Weibull distributions is found as an appropriate model of the web content popularity. In [4] real web traces from the National Laboratory for Applied Network Research (NLANR) were analyzed. It was found that about 70% of all documents are one-time requested. The caching of such documents is not reasonable. However, the LRU and TTL rules allow to cache such documents.

Assuming that the document popularity can be modeled by a generalized Zipf's law, in [5] it was derived that the cache missing probability is asymptotically, for sufficiently large cache sizes, the same for the LRU replacement both with dependent and independent identically distributed IRTs. In [6] similar result was proved for the Least-Frequently-Used (LFU) rule which only caches the most popular objects. The IRTs of each object may be dependent and heavy-tailed distributed which reflects the heavy-tailed popularity of the documents.

According to our new caching rule called *Cluster Caching Rule* (CCR), only clusters of frequently requested contents whose popularity exceeds a sufficiently

[1] If the document size is fixed then a new document evicts one document from the full cache.

large threshold may hit the cache. This is similar to TTL and LFU rules, [10] but there are novelties based on the extreme value theory. The latter allows us to explain the impact of the dependence and heavy-tailed quantities on the caching. Due to dependence and a possibly heavy-tailed distribution of IRTs of the object, the popularity and IRT processes build clusters of exceedances, see Fig. 1. Such assumptions like (1) the content size is constant; (2) the IRT process is renewal Poisson; (3) IRTs are independent are avoided.
We consider clusters of exceedances of a popularity process of random sized content whose IRTs may be statistically correlated. Inter-arrival times of documents can be heavy-tailed. The clusters provide objects to hit/miss the cache.

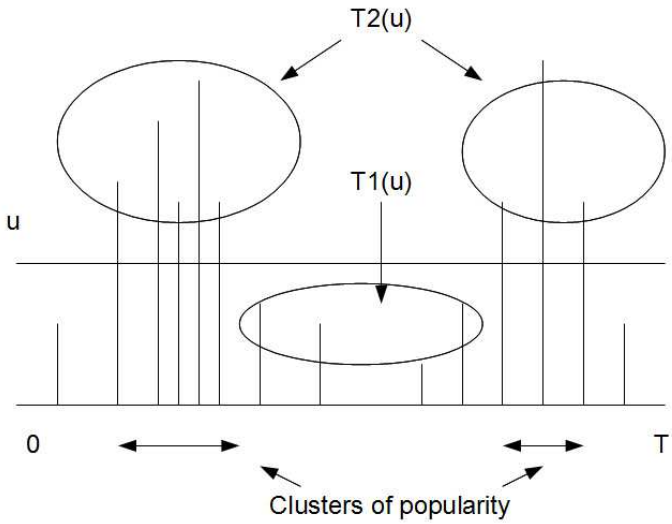


Fig. 1: Clusters of exceedances where the popularity of requested documents exceeds the threshold u within the time frame $[0, T]$; $T_1(u)$ and $T_2(u)$ denote inter-cluster and cluster sizes for the threshold u, respectively.

We obtain the cache utilization and occupancy by means of clusters and explain the impact of correlated requests on the fault probability from perspectives of clustering. The CCR allows to recommend the cache size.
We focus on clusters of exceedances of the popularity process. The popularity may change over time. When the next cluster of the most popular documents is requested, it is sent to the cache. The former cache content is evicted. If some popular content is requested repeatedly then it remains in the cache without replacement. In case, the cache size is smaller than the cluster size, the threshold u may be increased until the cluster size becomes smaller or equal to the cache size. The cache content may be changed completely by a new cluster. In our approach, the document size can be variable, i.e. a random variable (r.v.).

We concern only probabilistic aspects of fitting the cache avoiding statistical and combinatorial methods.
The paper is organizing as follows. In Section 2 necessary results from the extreme value theory are mentioned. Our achievements are given in Section 3. We finalize our exposition with conclusions.

2. Facts from extreme value theory

We focus on clusters of exceedances of an underlying process over a sufficiently high threshold as a source for caching. The cluster contains a set of consecutive exceedances of the process between two consecutive non-exceedances, [11]. The mean cluster size (i.e. the mean number of exceedances per cluster) is approximated by the reciprocal of the extremal index θ. θ is the dependence measure of extremes since it allows to represent the distribution of maxima of n dependent r.v.s.

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

$$\lim_{n\to\infty} n(1-F(u_n)) = \tau \qquad \textit{and} \tag{1}$$

$$\lim_{n\to\infty} P\{M_n \leq u_n\} = e^{-\tau\theta}$$

hold ([12], p.53).

For an iid sequence $\theta = 1$ holds. $\theta = 0$ indicates a total dependence. Clusters form a compound Poisson process with the rate $\theta\tau$, [13].

2.1. Clusters of exceedances. Let us consider the inter-cluster size as

$$T_1(u) \quad = \quad \min\{j \geq 1 : M_{1,j} \leq u, X_{j+1} > u | X_1 > u\}$$

and the cluster size as

$$T_2(u) \quad = \quad \min\{j \geq 1 : L_{1,j} > u, X_{j+1} \leq u | X_1 \leq u\},$$

where $M_{1,j} = \max\{X_2, ..., X_j\}$, $M_{1,1} = -\infty$, $L_{1,j} = \min\{X_2, ..., X_j\}$, $L_{1,1} = +\infty$, Fig. 1.
In [11] asymptotically equivalent distributions of $T_1(u_n)$ and $T_2(u_n)$ are derived for sequence of increasing thresholds $\{u_n\}$. It was proposed to take sufficiently high quantiles x_{ρ_n} of the level $(1-\rho_n)$ ($\rho_n \to 0$ as $n \to \infty$[2]) of the process X_t as u_n. If specific mixing conditions are fulfilled uniformly in $j \in [a, n]$, $a = a(n) \to \infty$, $n \to \infty$ then geometric-like models

$$P\{T_1(x_{\rho_n}) = j\} \sim \rho_n(1-\rho_n)^{(j-1)\theta}, \qquad P\{T_2(x_{\rho_n}) = j\} \sim q_n\,(1-q_n)^{(j-1)\theta} \tag{2}$$

[2]This follows from (1) since $\rho_n = 1 - F(x_{\rho_n}) \sim \tau/n$.

hold as $n \to \infty$, where $q_n = 1 - \rho_n$ and n is a sample size.[3] The required mixing conditions are equivalent to the condition

$$|P\{X_{j+1} \le x_{\rho_n} | X_1 \le x_{\rho_n}\} - P\{X_{j+1} \le x_{\rho_n}\}| = o(1), \qquad n \to \infty. \quad (3)$$

In case $j = 1$ the random sequence $\{X_i\}$ must be independent with $\theta = 1$ to satisfy (3). (3) is valid, for example, for the wide class of m-dependent processes for $j > m$ and Markov chains for $j \ge 2$.
The ssymptotic expectation of $T_2(x_{\rho_n})$ was obtained in [11]. It was derived that if for some $\varepsilon > 0$

$$\sup_n E(T_2^{1+\varepsilon}(x_{\rho_n}))/\Lambda_{n,2} \quad < \quad \infty, \qquad \Lambda_{n,2} = q_n/(1-(1-q_n)^\theta)^2$$

holds, then it follows

$$\lim_{n\to\infty} E(T_2(x_{\rho_n}))/\Lambda_{n,2} = 1. \quad (4)$$

In [14] it was derived that if for some $\varepsilon > 0$

$$\sup_n E(T_1^{1+\varepsilon}(x_{\rho_n}))/\Lambda_{n,1} < \infty, \qquad \Lambda_{n,1} = 1/(1-(1-\rho_n)^\theta)^2$$

holds, then it follows

$$\lim_{n\to\infty} E(T_1(x_{\rho_n}))/(\Lambda_{n,1}\rho_n) = 1. \quad (5)$$

It is remarkable that models (2) are valid for any distribution of the underlying process $\{X_t\}$ if the dependence structure (3) holds.

3. Main results

Let $\{X_i\}_{1\le i\le M}$ be a stationary process of the content popularity with marginal cumulative distribution function $F(x)$ and $M_n = \max\{X_1, ..., X_n\}$. We assume $P\{X_1 = x_F\} = 0$, where $x_F = \sup\{x : F(x) < 1\}$ is the right end-point of $F(x)$. This indicates that x_F is infinite and $F(x)$ is heavy-tailed. Let $\{Y_i\}$ be iid sizes of documents from the catalog with finite mean $EY_i = \alpha_1$. Let $\{\tau_n\}_{n\ge1}$ and $\{\tau_{i,n}\}_{n\ge1}$ be stationary processes of inter-arrival times (IATs) of documents and IRTs of the ith document, $i = 1, .., M$, respectively. The frequency of requests changes over time. We determine the popularity of the ith document at its jth request time $T_{i,j} = \sum_{n=1}^{j} \tau_{i,n}$ by

$$X_i = j/N_{T_{i,j}}, \qquad i = 1, ..., M, \qquad j \ge 1,$$

where $N_{T_{i,j}}$ is the total number of requests in time interval $[0, T_{i,j}]$.

[3] $f_n \sim g_n$ implies $\lim_{n\to\infty} f_n/g_n = 1$.

3.1. Cluster Caching Rule. Let us introduce the CCR. Only highly requested documents corresponding to clusters of exceedances of the popularity process may hit the cache. The new cluster content renews the cache apart of objects that could be found both in previous and (or) in present clusters and thus, they could hit the cache earlier. Frequently requested documents may appear in the popularity cluster several times. We denote the number of exceedances in the cluster corresponding to different objects as $T_2^*(u)$ such as $T_2^*(u) \leq T_2(u)$. $T_1^*(u)$ denotes the number of non-exceedances in the inter-cluster corresponding to different objects ($T_1^*(u) \leq T_1(u)$). The cluster content or more exactly $T_2^*(u)$ may be smaller or larger than the cache size C. This may lead to a not complete cache utilization or to loss of documents exceeding the cache in size, respectively. One can decrease the cluster content by increasing of u or vice versa. In case u is fixed, one can send the excess documents to the next cache if there is a line or a tree of caches. We focus here on the single-cache control.

The CCR is similar to the TTL and the LFU rules, respectively. Really, LFU caches only the most popular documents and TTL caches only frequent documents with sufficiently small IRTs. In contrast to the latter rules, the CCR may lead to a long random cache occupancy by popular objects if they are found in several consecutive clusters. By the LRU rule and the TTL rules any requested documents may hit the cache including one time requested ones. It leads to a un-effective use of the cache. The CCR does not cache rarely requested objects. The mean cluster size $ET_2(u) \approx 1/\theta$ may be proposed as a cache size.

3.2. Cache utilization. Let us consider the cache utilization for fixed and random content sizes. Let Y_i be fixed. The utilization may be determined by the ratio $T_2^*(x_{\rho_n})/C$ of the cluster and the cache sizes. Clusters may contain changing popularity of the same documents. Since $ET_2(u) \approx 1/\theta$ holds we get

$$E(T_2^*(x_{\rho_n})/C) \quad \leq \quad E(T_2(x_{\rho_n})/C) \approx 1/(\theta C). \tag{6}$$

Hence, the mean utilization may reach 100% if $C = 1/\theta$.

Let Y_i be random. The cluster content and the inter-cluster duration are determined respectively by

$$S_{T_2(u)} \quad = \quad \sum_{i=1}^{T_2(u)} Y_i, \qquad S_{T_1(u)} = \sum_{i=1}^{T_1(u)} \tau_i. \tag{7}$$

From independence of Y_i and $T_2(u)$ and from Wald's equation we get

$$E(S_{T_2^*(u)}/C) \quad \leq \quad E(S_{T_2(u)}/C) = E(T_2(x_{\rho_n}))E(Y_i)/C \approx \alpha_1/(\theta C) \tag{8}$$

Thus, the mean utilization may reach 100% if $C = \alpha_1/\theta$. The cache size requires a statistical estimation of the extremal index θ. Since θ is a function

of the threshold u, one can use (4) to determine C. (6) can be rewritten as

$$E(T_2(x_{\rho_n})/C) \sim q_n/((1-(1-q_n)^\theta)^2 C)$$

for $q_n \sim 1-\tau/n$.[4] Then $C = q_n/(1-(1-q_n)^\theta)^2$ can be selected. If the content popularity $\{X_i\}$ are iid and $\theta = 1$ holds then $C = 1/q_n$ may be selected.

3.3. Cache occupancy. The cache occupancy shows how long an object stays in cache within some time interval. Let $C(t)$ denote a binary process that shows the presence of the ith document in the cache. Denoting independent renewal processes of IRTs of individual objects and their TTLs as $\{\tau_{i,k}\}_{k\geq 1}$ and $\{T_{i,k}\}_{k\geq 1}$, respectively, in [1] the following formula of the cache occupancy

$$\lim_{n\to\infty} \int_0^{t_n} C(s)ds/t_n = \lim_{n\to\infty} \sum_{k=1}^{n} \min\{\tau_{i,k}, T_{i,k}\} / \sum_{k=1}^{n} \tau_{i,k} \tag{9}$$

was proposed. Here, $t_n = \sum_{k=1}^{n} \tau_{i,k}$ is the time of the nth request of the object and $C(s)$ is a binary process indicating the presence of the object in cache. (9) corresponds to the TTL rule with *R-policy* when the stopping time to be in cache is equal to $\min\{t : \tau_t > T_t\}$.

The CCR-approach is different from (9). Once the object hits the cache, it will stay there at least the time equal to the inter-cluster duration $S_{T_1(u)}$. Thus, assuming that the IATs $\{\tau_i\}$ of the documents are iid r.v.s with $\beta_1 = E(\tau_i) < \infty$ and the mutual independence of τ_i and $T_1(u)$, we obtain from Wald's equation and (5), (7)

$$E(S_{T_1(x_{\rho_n})}) = E(T_1(x_{\rho_n}))E(\tau_i) \sim \beta_1/(1-(1-\rho_n)^\theta)^2$$

with $\rho_n \sim \tau/n$. More precisely, the cache occupancy of the ith object may be longer or equal to

$$O_i \geq \sum_{j=1}^{L_{i,T}} \left(\sum_{k=1}^{T_{1,j}(u)} \tau_k + \sum_{k=k^*}^{T_{2,j}(u)} \tau_k \right),$$

where $k^* = \min\{t : X_t \geq u\}$ is the first hitting time of the cache by the ith object located in the cluster, $L_{i,T}$ is the number of consecutive clusters containing the ith document and, $T_{1,j}(u)$ and $T_{2,j}(u)$ are inter- and cluster sizes, respectively, at time interval $[0, T]$.

3.4. Dependent and independent requests. If one uses clusters of the popularity exceedances over the threshold to fit the cache, then for any stationary popularity distribution (2) is valid. A consistent statistical estimation of θ is also required.

[4]This follows from (1) since $1 - F(x_{\rho_n}) = \rho_n = 1 - q_n \sim \tau/n$ as $n \to \infty$.

Using (2) one can approximate the probability to hit the whole cache by one cluster as

$$P\{T_2(x_{\rho_n}) = C\} = q_n\,(1-q_n)^{(C-1)\theta} + o(1), \qquad n \to \infty. \tag{10}$$

Looking on the right-hand side probability of (10) one may conclude that $(1-q_n)^{(C-1)\theta}$ tends to 0 as $q_n \to 1$ and q_n does not influence much if $C \to \infty$. Since the dependence is accumulated in the extremal index θ, this explains why correlations of IRTs of some document do not impact on the fault probability asymptotically as $C \to \infty$. The latter conclusion was done in [5], [6] regarding the LRU and the LFU rules. It was concluded that 1) the LRU and the LFU caching work the same both for iid and correlated requests, hence one should not take the dependence into account; 2) the fault probability is generated by the distribution tail of the requests. We can explain these features by the clustering and (2). One cannot exclude θ as the dependence measure from the consideration for a fixed C. The iid requests imply $\theta = 1$. Furthermore, clusters of exceedances are generated by rare events, namely, by exceedances over a sufficiently high threshold. The latter relate to the tail of the popularity or IRT distributions. The higher the threshold the more chance to have statistically independent clusters. Such clusters contain likely new observations that cannot be found in the cache. Cluster contents cause missing objects in the cache and the fault probability.

4. Conclusions and future work

A new caching rule *CCR* based on clusters of exceedances by an underlying process over a sufficiently large threshold is proposed. Considering the content popularity as underlying process, its clusters of exceedances indicate popular frequently requested documents . The latter are to be located in the cache. The CCR generalizes the TTL and LFU rules. The CCR is formulated for a single cache. The content size may be random. The IRTs may be correlated. Future work will concern further investigation of the CCR and its extension to lines and trees of caches.

Acknowledgments. The author was partly supported by the Russian Foundation for Basic Research, grant 13-08-00744 A

REFERENCES

1. Berger D. S., Gland P., Singla S., Ciucu F. Exact Analysis of TTL Cache Networks: The Case of Caching Policies Driven by Stopping Times // The 2014 ACM International Conference on Measurement and Modeling of Computer Systems. SIGMETRICS '14. 2014. P. 595-596.
2. Foback N. C., Nain P., Neglia G., Towsley D. Analysis of ttl-based cache networks //IEEE VALUETOOLS. 2012. P. 1-10.
3. Friecker C., Robert P., Roberts J. A versatile and accurate approximation for LRU cache performance // In Proceedings of ITC. 2012. P. 1-8.

4. Che H., Tung Y., Wang Z. Hierarchical web caching systems: modeling, design and experimental results // IEEE JSAC. 2002. V. 20(7). P. 1305–1314.
5. Jelenković P. R., Radovanović A. Least-Recently-Used Caching with Dependent Requests // Theoretical Computer Science. 2004. V. 326(1-3). P. 293 - 327.
6. Jelenković P. R., Radovanović A. Asymptotic optimality of the static frequency caching in the presence of correlated requests // Operations Research Letters. 2009. V. 37(5). P. 307 - 311.
7. Clauset A., Shalizi C. R., Newman M. E. J. Power-Law Distributions in Empirical Data // SIAM Rev. 2009. V. 51(4). P. 661–703.
8. Newman M. E. J. Power laws, Pareto distributions and Zipf's law // arXiv:cond-mat/0412004v3 [cond-mat.stat-mech]. 2006.
9. Imbrenda C., Muscariello L., Rossi D. Analyzing Cacheable Traffic in ISP Access Networks for Micro CDN Applications via Content-Centric Networking // ACM SIGCOMM Information Centric Networks (ICN). 2014. 09/2014.
10. Lee D., Choi J., Kim J.-H., Noh S. H., Min S. L., Cho Y., Kim C. S.. LRFU: A Spectrum of Policies that Subsumes the Least Recently Used and Least Frequently Used Policies // IEEE Transactions on computers. 2001. V. 50 (12). P. 1352-1362.
11. Markovich N. M. Modeling clusters of extreme values // Extremes. 2014. V. 17(1). P. 97-125.
12. Leadbetter M.R., Lingren G., Rootzén H. Extremes and Related Properties of Random Sequence and Processes. Springer, New York, 1983.
13. Beirlant J., Goegebeur Y., Teugels J., Segers J. Statistics of Extremes: Theory and Applications. Wiley, Chichester, West Sussex, 2004.
14. Markovich N. M. (2015). Clusters of extremes: models, estimation and applications (submitted).

RESEARCH OF THE PROCESS OF TRAFFIC FLOWS CONTROL BY MEANS OF SIMULATION

M.A. Rachinskaya, M.A. Fedotkin
Lobachevsky State University of Nizhni Novgorod,
Nizhni Novgorod, Russian Federation

The present paper investigates the model of a crossroad controlled by a cyclic traffic light. Two traffic flows approaching the intersection are supposed to have different probabilistic structure. The current research conducted by means of simulation is based on a mathematical model constructed previously in the form of a queueing system with variable structure. The paper includes the following investigations: 1) a time moment when transient period is elapsed is determined; 2) an estimate of crossroad load is proposed; 3) the estimates for mathematical expectation and dispersion of time an arbitrary vehicle is waiting for service and the size of a certain vehicular queue approaching the crossroad before the start of service signal for the corresponding flow are presented; 4) the optimal values of service duration for the traffic flows are determined.

ИССЛЕДОВАНИЕ ПРОЦЕССА УПРАВЛЕНИЯ ПОТОКАМИ С ПОМОЩЬЮ ИМИТАЦИОННОГО МОДЕЛИРОВАНИЯ

М.А. Рачинская, М.А. Федоткин
Нижегородский государственный университет им. Н.И. Лобачевского,
г. Нижний Новгород, Российская Федерация
rachinskaya.maria@gmail.com, fma5@rambler.ru

Аннотация

В работе исследуется модель перекрестка на пересечении транспортных магистральных потоков, управляемых циклически автоматом-светофором. Входные потоки машин имеют различную вероятностную структуру. На основе ранее построенной математической модели в виде системы массового обслуживания с переменной структурой средствами имитационного моделирования определяются следующие характеристики: 1) момент завершения переходных процессов в системе; 2) оценка загрузки перекрестка по отдельным потокам и системы по всем потокам; 3) значения оценок математического ожидания и дисперсии основных показателей качества работы системы: времени ожидания начала обслуживания заявки некоторого потока и произвольной заявки системы, размера очереди по некоторому потоку перед началом обслуживания; 4) полигон частот числа машин,

обслуженных за цикл работы системы для каждого потока; 5) оптимальные значения длительностей интервалов обслуживания потоков.

Ключевые слова: Система массового обслуживания, поток пачек, имитационное моделирование, загрузка системы

1. Постановка задачи

Объект исследований данной работы — перекресток, на который с крупных магистральных дорог поступает 2 транспортных потока Π_1 и Π_2. При моделировании перекресток рассматривается как система массового обслуживания с переменной структурой. Так, потоки машины воспринимаются как входные потоки заявок, перекресток с устройством, регулирующим движение машин, рассматривается как обслуживающее устройство, под обслуживанием понимается переезд машин через перекресток. Управление переездом осуществляется автоматом-светофором, в котором выделено 4 следующих состояния (фазы). В фазах $\Gamma^{(1)}$ и $\Gamma^{(3)}$ зеленого света по перекрестку осуществляется переезд только машин потоков Π_1 и Π_2 соответственно. Фазы $\Gamma^{(2)}$ и $\Gamma^{(4)}$ желтого света выделены для переналадки обслуживающего устройства — при этом машины потоков Π_1 и Π_2 соответственно, начавшие переезд в фазу зеленого света, завершают его, новые машины на обслуживание не поступают. Для управления потоками выбран циклический алгоритм переключения состояний светофора: $\Gamma^{(1)} \to \Gamma^{(2)} \to \Gamma^{(3)} \to \Gamma^{(4)} \to \Gamma^{(1)} \to \ldots$. Для любого $k \in \overline{1,4}$ длительность фазы $\Gamma^{(k)}$ обозначена через T_k. Полная смена фаз светофора происходит за цикл длительностью $T = \sum_{k=1}^{4} T_k$. Устанавливаются интенсивности μ_j (здесь и далее $j \in \{1,2\}$) обслуживания соответствующих потоков. Выбранная экстремальная стратегия обслуживания предполагает, что из очереди машин по потоку Π_j, ожидающих возможности переезда, в фазу $\Gamma^{(2j-1)}$ обслуживания соответствующего потока выбирается как можно большее число машин, но не превышающее естественной пропускной способности $l_j = [\mu_j T_{2j-1}]$. Наблюдать за изменением состояния указанной системы, функционирующей в непрерывном времени, будем в дискретные случайные моменты $\tau_0 > 0, \tau_1, \tau_2, \ldots$ последовательного переключения фазы светофора.

В работах [1, 2, 3] была построена и изучена вероятностная модель указанного перекрестка в случае, когда число транспортных потоков больше 2. Однако, в силу обширности изучаемого объекта, получение некоторых результатов аналитически не представляется возможным. В связи с этим на основе результатов исследования математической модели перекрестка была построена компьютерная имитационная модель, позволяющая проведить численные исследования.

Входные транспортные потоки могли быть сформированы под влиянием различных факторов. Например, исследования [2] показали, что при плохих дорожных и затрудняющих свободное движение погодных услови-

ях, в случае низкой плотности дорожного трафика хорошей аппроксимацией любого из входных потоков Π_j является неординарный поток пачек, параметрами которого являются интенсивность λ_j поступления пачек и вероятности p_j, q_j и $s_j = 1 - p_j - q_j$ поступления пачек из одной, двух и трех машин соответственно. Имитационная модель построена для случая таких входных потоков. Отметим, что варьируя значения параметров, из потоков указанной вероятностной структуры можно получить, например, пуассоновские потоки ($p_j = 1$) или потоки Бартлетта ($s_j = 0$). Такая модель позволяет отражать влияние внешней среды (переменчивой погоды, участков различного качества дорог и пр.) на образование потоков.

2. Стационарный режим в системе

Согласно результатам работы [3], критерий существования в системе стационарного режима заключается в одновременном выполнении неравенств $\lambda_j T(2s_j + q_j + 1) < l_j, j \in \{1, 2\}$. Наблюдения за реальными перекрестками показывают, что до момента окончания всех переходных процессов часто проходит достаточно длительный интервал времени. Для определения момента достижения стационарного режима в системе по потоку Π_j предлагается следующий метод. Будем имитировать функционирование перекрестка при отсутствии машин в очереди по потоку Π_j в момент τ_0, а также при наличии некоторого количества $[\lambda_j T]$ машин в начальный момент. Пусть величины $\gamma_{j,v}$ и $\gamma^+_{j,v}$ есть случайные времена ожидания начала обслуживания машины с номером v потока Π_j в системе при отсутствии и наличии машин в очереди в начальный момент соответственно. Будем считать значения величин $\widehat{\gamma}_{j,u} = \frac{1}{u}\sum_{v=1}^{u}\gamma_{j,v}$ и $\widehat{\gamma}^+_{j,u} = \frac{1}{u}\sum_{v=1}^{u}\gamma^+_{j,v}$. Введем следующие параметры метода: d, N — натуральные числа и $0 < \delta < 1$. Пусть случайная величина θ_j определяет номер u машины потока Π_j, для которой первый раз произошло d-кратное по u выполнение неравенства

$$|\widehat{\gamma}^+_{j,u} - \widehat{\gamma}_{j,u}| < \delta\widehat{\gamma}_{j,u}. \tag{1}$$

Это означает, что траектории процессов обслуживания машин в системе при отсутствии и наличии машин в момент τ_0 сблизились d раз. В свою очередь величина ζ_j определяет момент начала обслуживания машины с номером θ_j. Эту процедуру повторим N раз с независимыми реализациями потока Π_j, сгенерированными при различных начальных значениях датчика псевдослучайных чисел. Для каждой из реализаций с номером n (здесь и далее $1 \le n \le N$) имеем копии введенных величин: θ^n_j и ζ^n_j. Введем также случайные величины θ^*_j и η^*_j, определяемые из равенств $\theta^*_j = \max_{1\le n\le N}\theta^n_j$ и $\theta^*_j = \theta^{\eta^*_j}_j$. Заметим, что η^*_j определяет номер реализации, на которой достигнут максимум величин вида θ^n_j. Тогда длительность переходного процесса по потоку Π_j можно определять величинами θ^*_j или $\zeta^{\eta^*_j}_j$. Считаем, что по истечении времени $\zeta^{\eta^*_j}_j$ стационарный режим по потоку Π_j достигнут во всех N реализациях. Длительность переходного процесса в системе

в целом есть величина $\zeta^* = \max_{j\in\{1,2\}} \zeta_j^{\eta_j^*}$. Работа алгоритма для потока Π_1 при $N = 10$ реализациях и значениях параметров $\lambda_1 = 0.19, p_1 = 0.55$, $q_1 = 0.35, T_1 = 19, T_2 = T_4 = 3, T_3 = 17, \mu_1 = 0.65, d = 3, \delta = 0.07$ проиллюстрирован на рисунке 1. Для каждой реализации с номером n точки с абсциссой u (номер машины) и ординатой $|\widehat{\gamma}_{1,u}^{+} - \widehat{\gamma}_{1,u}| - \delta\widehat{\gamma}_{1,u}$ (разность, сформированная согласно (1)) соединены для наглядности ломаной. Справа от графика для каждой n-ой реализации фиксированы значение u величины θ_1^n и значение t в минутах величины ζ_1^n, когда впервые d точек ломаной окажется ниже оси абсцисс. Длительностью переходного процесса считаем время t, соответствующее реализации, в которой номер u машины максимален. В указанном примере стационарный режим достигнут после 266 машины. Длительность переходного процесса 16.1 мин.

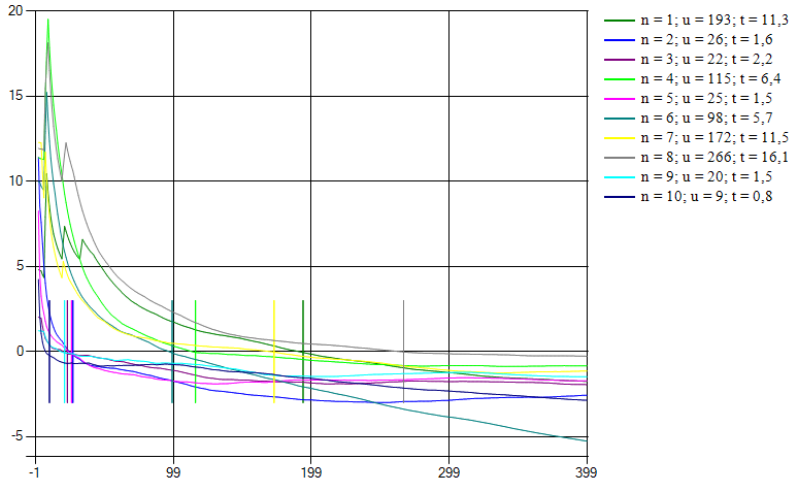


Рис. 1: Динамика сближения траекторий потока Π_1 для $N = 10$ реализаций

3. Загрузка перекрестка

Под загрузкой перекрестка по потоку Π_j понимаем отношение времени, которое тратит система на обслуживание машин потока Π_j, ко всему времени функционирования перекрестка. В качестве оценки загрузки (квазизагрузки) $\widetilde{\rho}_j$ исследуемой системы по потоку Π_j предлагается использовать соотношение $\widetilde{\rho}_j = \frac{\lambda_j(1+q_j+2s_j)}{\mu_j}$. Отметим, что в числителе данной дроби стоит оценка интенсивности поступления машин на перекресток — произведение интенсивности поступления пачек потока Π_j на математическое ожидание числа машин в пачке, а в знаменателе — интенсивность обслуживания машин данного потока в зеленую фазу светофора. Квазизагрузка $\widetilde{\rho}$ системы по всем потокам считается по формуле $\widetilde{\rho} = \widetilde{\rho}_1 + \widetilde{\rho}_2$.

Обратим внимание на то, что предложенная оценка загрузки по потоку Π_j удовлетворяет следующим естественным условиям:

- $0 < \widetilde{\rho_j} < 1$ (в стационарном режиме $\lambda_j T(1 + q_j + 2s_j) < [\mu_j T_{2j-1}]$ и, следовательно, $\frac{\lambda_j(1+q_j+2s_j)}{\mu_j} < \frac{T_{2j-1}}{T} < 1$);
- с ростом интенсивности машин входного потока (с увеличением значения параметра λ_j и/или среднего числа машин в пачке) значение квазизагрузки $\widetilde{\rho_j}$ растет, в то время как с ростом интенсивности μ_j обслуживания машин потока Π_j значение квазизагрузки убывает;
- если при некотором изменении параметров λ_j, q_j, μ_j значение величины $\lambda_j T(2s_j + q_j + 1) - [\mu_j T_{2j-1}]$ стремится к нулю, то значение квазизагрузки по потоку стремится к величине $\frac{T_{2j-1}}{T}$.

Аналогичным условиям удовлетворяет и оценка $\widetilde{\rho}$ загрузки перекрестка по всем потокам. Согласованность таких оценок с реальными данными может быть подтверждена следующим образом. Пусть процесс имитации останавливается для каждой реализации с номером n, когда обслужены все машины хотя бы одного из предварительно сгенерированных входных потоков с некоторым заданным максимально возможным количеством пачек. Тогда введем следующие случайные величины: $\widehat{\kappa_j} = [\zeta_j^{\eta_j^*}/T]$ — номер цикла функционирования системы, на котором был достигнут стационарный режим для потока Π_j, κ^n — число циклов процесса имитации для n-ой реализации, $\kappa^* = \min_{1 \le n \le N} \kappa^n$. Пусть также $\alpha_{j,k}^n$ — случайное число машин потока Π_j, которое было обслужено в течение k-го цикла в реализации с номером n. Тогда по окончанию процесса имитации случайная величина

$$\rho_j = \frac{1}{N} \sum_{n=1}^{N} \frac{1}{(\kappa^* - \widehat{\kappa_j})T} \sum_{k=\widehat{\kappa_j}}^{\kappa^*} \frac{\alpha_{j,k}^n}{\mu_j} \tag{2}$$

определяет загрузку перекрестка по потоку Π_j. Общая загрузка системы по потокам вычисляется как $\rho = \rho_1 + \rho_2$. Эксперименты показывают близость значений ρ_1, ρ_2 и ρ, полученных путем имитации при достаточной длительности T_{st} стационарного режима, к значениям квазизагрузок $\widetilde{\rho_1}, \widetilde{\rho_2}$ и $\widetilde{\rho}$. Различные строки таблицы 1 соответствуют различным наборам параметров входных потоков и перекрестка.

T_{st} (час.)	ρ_1	$\widetilde{\rho_1}$	ρ_2	$\widetilde{\rho_2}$	ρ	$\widetilde{\rho}$
2.98	0.402	0.403	0.411	0.41	0.813	0.813
3.98	0.452	0.453	0.385	0.388	0.838	0.841
6.04	0.333	0.333	0.352	0.353	0.686	0.687
9.73	0.189	0.189	0.225	0.225	0.414	0.414

Таблица 1: Сравнение загрузок и квазизагрузок системы.

4. Оценки показателей качества функционирования перекрестка

Известными показателями работы систем массового обслуживания являются среднее время ожидания начала обслуживания заявки системы, средняя длина очереди на обслуживание и т. д. К сожалению, для изучаемой системы не удается найти аналитические формулы для аналогичных показателей, поэтому будем оценивать их средствами имитации. Пусть случайные величины $\beta_{j,k}^n$ считают число машин потока Π_j, находящихся в очереди перед началом обслуживания потока Π_j в k-ом цикле в n-ой реализации. С применением трех статистических критериев [4] (фазово-частотный критерий Валлиса-Мура, инверсионный критерий и фазово-частотный критерий с учетом длин фаз) было показано, что времена ожидания $\gamma_{j,\theta_j^*}^n, 1 \le n \le N$, начала обслуживания первой заявки в стационарном режиме во всех реализациях являются независимыми и одинаково распределенными величинами. То же справедливо и для количества $\beta_{j,\widehat{\kappa_j}}^n, 1 \le n \le N$, машин потока Π_j в очереди перед началом обслуживания в первом цикле стационарного режима во всех реализациях. В то же время относительно величин последовательности $\{\gamma_{j,u}^n; u \ge \theta_j^*\}$, состоящей из времен ожидания начала обслуживания последовательно поступивших машин потока Π_j некоторой реализации с номером n, такого вывода сделать нельзя. Аналогичное утверждение верно и для величин последовательности $\{\beta_{j,k}^n; k \ge \widehat{\kappa_j}\}$. Тогда исходя из представленных рассуждений предлагаются следующие выражения для подсчета оценок $\widetilde{\mathbf{M}}\gamma_{j,\theta_j^*}$, $\widetilde{\mathbf{D}}\gamma_{j,\theta_j^*}$, $\widetilde{\mathbf{M}}\gamma$, $\widetilde{\mathbf{M}}\beta_{j,\widehat{\kappa_j}}$, $\widetilde{\mathbf{D}}\beta_{j,\widehat{\kappa_j}}$ и $\widetilde{\mathbf{M}}\beta$:

- $\widetilde{\mathbf{M}}\gamma_{j,\theta_j^*} = \frac{1}{N}\sum_{n=1}^N \gamma_{j,\theta_j^*}^n$ и $\widetilde{\mathbf{D}}\gamma_{j,\theta_j^*} = \frac{1}{N-1}\sum_{n=1}^N (\gamma_{j,\theta_j^*}^n - \widetilde{\mathbf{M}}\gamma_{j,\theta_j^*})^2$ — оценки математического ожидания и дисперсии времени ожидания начала обслуживания заявки с номером θ_j^* потока Π_j;
- $\widetilde{\mathbf{M}}\gamma = \frac{\sum_{j=1}^2 \lambda_j(2s_j+q_j+1)\widetilde{\mathbf{M}}\gamma_{j,\theta_j^*}}{\sum_{j=1}^2 \lambda_j(2s_j+q_j+1)}$ — оценка для среднего взвешенного времени ожидания начала обслуживания произвольной заявки в системе, находящейся в стационарном режиме;
- $\widetilde{\mathbf{M}}\beta_{j,\widehat{\kappa_j}} = \frac{1}{N}\sum_{n=1}^N \beta_{j,\widehat{\kappa_j}}^n$ и $\widetilde{\mathbf{D}}\beta_{j,\widehat{\kappa_j}} = \frac{1}{N-1}\sum_{n=1}^N (\beta_{j,\widehat{\kappa_j}}^n - \widetilde{\mathbf{M}}\beta_{j,\widehat{\kappa_j}})^2$ — оценки математического ожидания и дисперсии очереди перед началом обслуживания машин потока Π_j в цикле с номером $\widehat{\kappa_j}$;
- $\widetilde{\mathbf{M}}\beta = \frac{\sum_{j=1}^2 \lambda_j(2s_j+q_j+1)\widetilde{\mathbf{M}}\beta_{j,\widehat{\kappa_j}}}{\sum_{j=1}^2 \lambda_j(2s_j+q_j+1)}$ — оценка для средней взвешенной очереди перед началом обслуживания в произвольном цикле стационарного режима.

На рисунках 2 и 3 отражена динамика изменения значений указанных оценок в стационарном режиме системы для следующих значений параметров: $\lambda_1 = 0.19$, $p_1 = 0.55$, $q_1 = 0.35$, $\lambda_2 = 0.17$, $p_2 = 0.55$, $q_2 = 0.3$, $T_1 = 18$, $T_2 = T_4 = 2$, $T_3 = 17$, $\mu_1 = 0.9$, $\mu_2 = 0.92$, $d = 3$, $\delta = 0.05$, $N = 200$. Здесь и на графиках время указано в секундах, а очередь — в машинах.

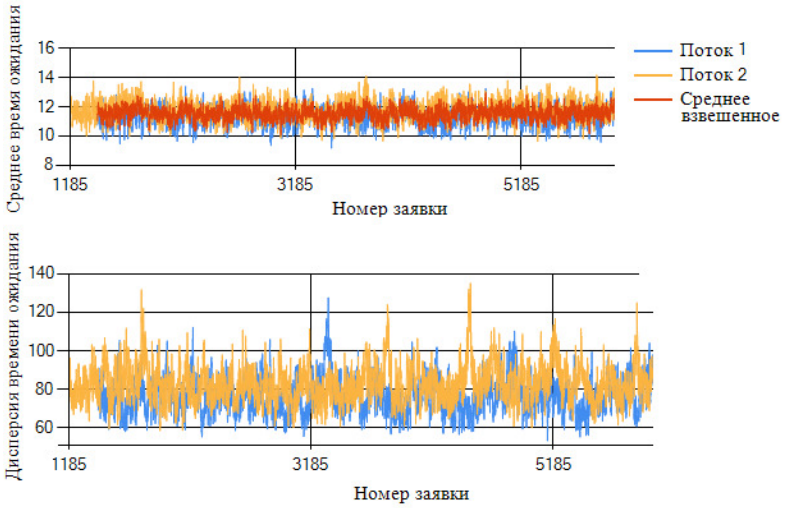


Рис. 2: Динамика изменения значений оценок с ростом номера заявки

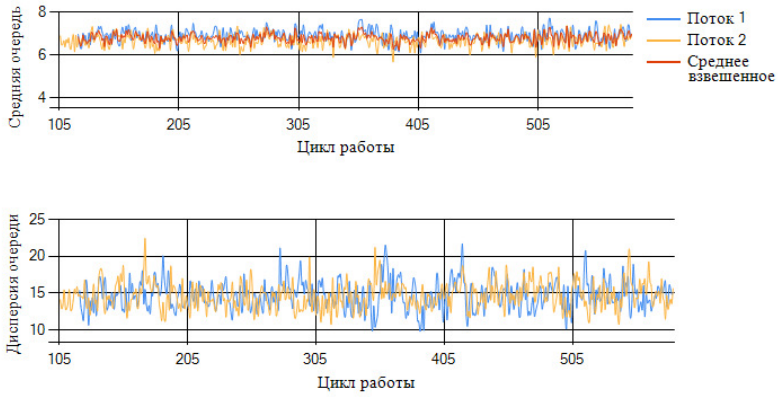


Рис. 3: Динамика изменения значений оценок с ростом номера цикла

Заметим, что для времени ожидания и для размера очереди значения на графиках колеблются в небольшом диапазоне вблизи полученных значений $\widetilde{\mathbf{M}}\gamma_{1,\theta_1^*} = 11.24$, $\widetilde{\mathbf{D}}\gamma_{1,\theta_1^*} = 77.79$, $\widetilde{\mathbf{M}}\gamma_{2,\theta_2^*} = 11.42$, $\widetilde{\mathbf{D}}\gamma_{2,\theta_2^*} = 77.03$, $\widetilde{\mathbf{M}}\gamma = 11.33$, $\widetilde{\mathbf{M}}\beta_{1,\widehat{\kappa_1}} = 6.72$, $\widetilde{\mathbf{D}}\beta_{1,\widehat{\kappa_1}} = 15.09$, $\widetilde{\mathbf{M}}\beta_{2,\widehat{\kappa_2}} = 6.73$, $\widetilde{\mathbf{D}}\beta_{2,\widehat{\kappa_2}} = 14.93$ и $\widetilde{\mathbf{M}}\beta = 6.72$.

На рисунке 4 приведен полигон частот числа обслуженных машин отдельных потоков за цикл функционирования перекрестка для тех же зна-

чений параметров, но при $N = 700$. Отметим, что при построении имитационной модели было допущено, что машина некоторого потока Π_j еще может начать обслуживаться непосредственно в момент, когда светофор переходит в желтую фазу $\Gamma^{(2j)}$ дообслуживания данного потока. В связи с этим предположением максимальное число машин потока Π_j, которое может пересечь перекресток за цикл, получается равным $l_j + 1$.

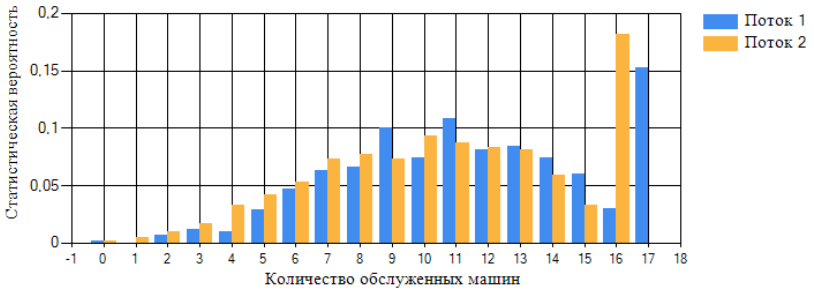


Рис. 4: Полигон частот обслуженных машин за цикл работы системы

В работе [4] указан алгоритм определения оптимальных значений длительностей T_1 и T_3 зеленых фаз светофора, при которых достигается минимальное значение оценки $\widetilde{\mathbf{M}}\gamma$. Для указанных параметров алгоритм оптимизации получает $T_1 = 6$ и $T_3 = 13$ при $T_2 = T_4 = 2$ и среднем взвешенном времени ожидания начала обслуживания произвольной заявки системы $\widetilde{\mathbf{M}}\gamma = 3,965$ сек.

ЛИТЕРАТУРА

1. Федоткин М. А., Кудрявцев Е. В., Рачинская М. А. О корректности вероятностных моделей динамики транспортных потоков на автомагистрали // Proceedings of International Workshop "Distributed computer and communication networks"(DCCN-2010). Moscow. 2010. P. 86-93.
2. Fedotkin M., Rachinskaya M. Parameters Estimator of the Probabilistic Model of Moving Batches Traffic Flow // Distributed Computer and Communication Networks. Springer International Publishing. 2014. V.279. P. 154-168.
3. Рачинская М. А., Федоткин М. А. Построение и исследование вероятностной модели циклического управления потоками малой интенсивности // Н.Новгород: Вестник ННГУ. 4(1). 2014. С. 370-376.
4. Рачинская М. А., Федоткин М. А. Численное исследование и синтез дискретных управляющих систем обслуживания // IX Международная конференция "Дискретные модели в теории управляющих систем": Москва и Подмосковье, 20-22 мая 2015 г.: Труды. М.: МАКС Пресс. 2015. С. 200-202.

QUICKEST MULTIDECISION ABRUPT CHANGE DETECTION WITH SOME APPLICATIONS TO NETWORK MONITORING

I. Nikiforov

Université de Technologie de Troyes, UTT/ICD/LM2S, UMR 6281, CNRS
12, rue Marie Curie, CS 42060 10004 TROYES CEDEX - FRANCE
nikiforov@utt.fr

Abstract

The quickest change detection/isolation (multidecision) problem is of importance for a variety of applications. Efficient statistical decision tools are needed for detecting and isolating abrupt changes in the properties of stochastic signals and dynamical systems, ranging from on-line fault diagnosis in complex technical systems (like networks) to detection/classification in radar, infrared, and sonar signal processing.

Keywords: sequential change detection/isolation, multidecision problems, anomaly detection, network monitoring

1. Introduction

The quickest change detection/isolation (multidecision) problem is of importance for a variety of applications. Efficient statistical decision tools are needed for detecting and isolating abrupt changes in the properties of stochastic signals and dynamical systems, ranging from on-line fault diagnosis in complex technical systems (like networks) to detection/classification in radar, infrared, and sonar signal processing. The early on-line fault diagnosis (detection/isolation) in industrial processes (SCADA systems) helps in preventing these processes from more catastrophic failures.

The quickest multidecision detection/isolation problem is the generalization of the quickest changepoint detection problem to the case of $K-1$ post-change hypotheses. It is necessary to detect the change in distribution as soon as possible and indicate which hypothesis is true after a change occurs. Both the rate of false alarms and the misidentification (misisolation) rate should be controlled by given levels.

2. Problem statement

Let $X_1, X_2, \ldots$ denote the series of observations, and let ν be the serial number of the *last pre-change* observation. In the case of multiple hypothesis, there are several possible post-change hypotheses $\mathcal{H}_j$, $j = 1, 2, \ldots, K-1$. Let $\mathbb{P}_k^j$ and $\mathbb{E}_k^j$ denote the probability measure and the expectation when $\nu = k$ and $\mathcal{H}_j$ is the true post-change hypothesis, and let $\mathbb{P}_\infty$ and $\mathbb{E}_\infty = \mathbb{E}_0^0$ denote the same when $\nu = \infty$, i.e., there is no change. Let (see [1] for details)

$$\mathbb{C}_\gamma = \left\{ \delta = (T, d) \colon \min_{0 \le \ell \le K-1} \min_{1 \le j \ne \ell \le K-1} \mathbb{E}_0^\ell \left(\inf_{r \ge 1} \{ T_r \,:\, d_r = j \} \right) \ge \gamma \right\}, \tag{1}$$

where T is the stopping time, d is the final decision (the number of post-change hypotheses) and the event $\{d_r = j\}$ denotes the first false alarm of the j-th type, be the class of detection and isolation procedures for which the average run length (ARL) to false alarm and false isolation is at least $\gamma > 1$. In the case of detection–isolation procedures, the risk associated with the detection delay is defined analogously to Lorden's worst-worst-case and it is given by [1]

$$\mathsf{ESADD}(\delta) = \max_{1\le j\le K-1} \sup_{0\le\nu<\infty} \left\{\operatorname{esssup} \mathbb{E}_\nu^j[(T-\nu)^+|\mathcal{F}_\nu]\right\}. \tag{2}$$

Hence, the minimax optimization problem seeks to

$$\text{Find } \delta_{\text{opt}} \in \mathbb{C}_\gamma \text{ such that } \mathsf{ESADD}(\delta_{\text{opt}}) = \inf_{\delta\in\mathbb{C}_\gamma} \mathsf{ESADD}(\delta) \text{ for every } \gamma > 1, \tag{3}$$

where $\mathbb{C}_\gamma$ is the class of detection and isolation procedures with the lower bound γ on the ARL to false alarm and false isolation defined in (1).

Another minimax approach to change detection and isolation is as follows [2, 3]; unlike the definition of the class $\mathbb{C}_\gamma$ in (1), where we fixed *a priori* the changepoint $\nu = 0$ in the definition of false isolation to simplify theoretical analysis, the false isolation rate is now expressed by the maximal probability of false isolation $\sup_{\nu\ge0} \mathbb{P}_\nu^\ell(d = j \ne \ell|T > \nu)$. As usual, we measure the level of false alarms by the ARL to false alarm $\mathbb{E}_\infty T$. Hence, define the class

$$\mathbb{C}_{\gamma,\beta} = \left\{\delta = (T,d)\colon \mathbb{E}_\infty T \ge \gamma,\ \max_{1\le\ell\le K-1} \max_{1\le j\ne\ell\le K-1} \sup_{\nu\ge0} \mathbb{P}_\nu^\ell(d = j|T > \nu) \le \beta\right\}. \tag{4}$$

Sometimes Lorden's worst-worst-case ADD is too conservative, especially for recursive change detection and isolation procedures, and another measure of the detection speed, namely the maximal conditional average delay to detection $\mathsf{SADD}(T) = \sup_\nu \mathbb{E}_\nu(T-\nu|T>\nu)$, is better suited for practical purposes. In the case of change detection and isolation, the SADD is given by

$$\mathsf{SADD}(\delta) = \max_{1\le j\le K-1} \sup_{0\le\nu<\infty} \mathbb{E}_\nu^j(T-\nu|T>\nu). \tag{5}$$

We require that the $\mathsf{SADD}(\delta)$ should be *as small as possible* subject to the constraints on the ARL to false alarm and the maximum probability of false isolation. Therefore, this version of the minimax optimization problem seeks to

$$\text{Find } \delta_{\text{opt}} \in \mathbb{C}_{\gamma,\beta} \text{ such that } \mathsf{SADD}(\delta_{\text{opt}}) = \inf_{\delta\in\mathbb{C}_{\gamma,\beta}} \mathsf{SADD}(\delta) \text{for every } \gamma > 1 \text{ and } \beta \in (0,1). \tag{6}$$

A detailed description of the developed theory and some practical examples can be found in the recently published book [4].

3. Efficient procedures of quickest change detection/isolation

Asymptotic theory. In this paragraph we recall a lower bound for the worst mean detection/isolation delay over the class $\mathbb{C}_\gamma$ of sequential change detection/isolation

tests proposed in [1]. First, we start with a technical result on sequential multiple hypotheses tests and then we give an asymptotic lower bound for $\mathsf{ESADD}(\delta)$.

Lemma 1. *Let $(X_k)_{k\geq 1}$ be a sequence of i.i.d. random variables. Let $\mathcal{H}_0, \ldots, \mathcal{H}_{K-1}$ be $K \geq 2$ hypotheses, where $\mathcal{H}_i$ is the hypothesis that X has density f_i with respect to some probability measure μ, for $i = 0, \ldots, K-1$ and assume the inequality*

$$0 < \rho_{ij} \stackrel{def.}{=} \int f_i \log \frac{f_i}{f_j} d\mu < \infty, \quad 0 \leq i \neq j \leq K-1,$$

to be true.

Let $\mathbb{E}_i(N)$ be the average sample number (ASN) in a sequential test (N, δ) which chooses one of the K hypotheses subject to a $K \times K$ error matrix $A = \left[a_{ij}\right]$, where $a_{ij} = \mathbb{P}_i(\text{accepting } \mathcal{H}_j),\ i, j = 0, \ldots, K-1$.

Let us reparameterize the matrix A in the following manner :

$$\begin{pmatrix} 1-\sum_{\ell=1}^{K-1}\alpha_\ell & \alpha_1 & \ldots & \alpha_{K-1} \\ \gamma_1 & 1-\sum_{\ell=2}^{K-1}\beta_{1,\ell}-\gamma_1 & \ldots & \beta_{1,K-1} \\ \gamma_2 & \beta_{2,1} & \ldots & \beta_{2,K-1} \\ \ldots & \ldots & \ldots & \ldots \\ \gamma_i & \beta_{i,1} & \ldots & \beta_{i,K-1} \\ \ldots & \ldots & \ldots & \ldots \\ \gamma_{K-1} & \beta_{K-1,1} & \ldots & 1-\sum_{\ell=1}^{K-2}\beta_{K-1,l}-\gamma_{K-1} \end{pmatrix}.$$

Then a lower bound for the ASN $\mathbb{E}_i(N)$ is given by the following formula :

$$\mathbb{E}_i(N) \ \geq \ \max\left\{\frac{(1-\tilde{\gamma}_i)\ln\left(\sum_{\ell=1}^{K-1}\alpha_l\right)^{-1} - \log 2}{\rho_{i0}}, \max_{1\leq j\neq i\leq K-1}\left\{\frac{(1-\tilde{\gamma}_i)\ln\beta_{ji}^{-1} - \log 2}{\rho_{ij}}\right\}\right\}$$

for $i = 1, \ldots, K-1$, where

$$\tilde{\gamma}_i = \gamma_i + \sum_{\ell=1,\ell\neq i}^{K-1}\beta_{i,\ell}.$$

Theorem 1. *Let $(Y_k)_{k\geq 1}$ be an independent random sequence observed* sequentially *:*

$$\mathcal{L}(Y_k) = \begin{cases} P_0 & \text{if} \quad k \leq \nu \\ P_\ell & \text{if} \quad k \geq \nu+1 \end{cases}, \quad \nu = 0, 1, 2, \ldots, \quad \text{for} \ \ 1 \leq \ell \leq K-1$$

The distribution P_ℓ has density f_ℓ, $\ell = 0, \ldots, K-1$. An asymptotic lower bound for ESADD(δ)*, which extends the result of Lorden [5] to multiple hypotheses case, is :*

$$\mathsf{ESADD}(T;\gamma) \gtrsim \frac{\log\gamma}{\rho^*} \quad as \quad \gamma \to \infty,$$

where

$$\rho^* \stackrel{def.}{=} \min_{1\le\ell\le K-1} \min_{0\le j\neq\ell\le K-1} \rho_{\ell,j} \quad and \quad 0 < \rho_{\ell,j} \stackrel{def.}{=} \mathbb{E}_1^l\left(\log\frac{f_l(Y_i)}{f_j(Y_i)}\right) < \infty$$

is the K-L information.

Generalized CUSUM test. The generalized CUSUM (non recursive) test asymptotically attains the above mentioned lower bound [1]. Let us introduce the following stopping time and final decision

$$\tilde{N} = \min\{\tilde{N}^1, \ldots, \tilde{N}^{K-1}\}; \quad \tilde{d} = \operatorname{argmin}\{\tilde{N}^1, \ldots, \tilde{N}^{K-1}\}$$

of the detection/isolation algorithm. The stopping time $\tilde{N}^\ell$ is responsible for the detection of hypothesis $\mathcal{H}_\ell$:

$$\begin{aligned}
\tilde{N}^\ell &= \inf_{k\ge1} \tilde{N}^\ell(k),\ \tilde{N}^\ell(k) = \inf\left\{n \ge k : \min_{0\le j\neq\ell\le K-1} S_k^n(\ell, j) \ge h\right\} \\
\tilde{N}^\ell &= \inf\left\{n \ge 1\ : \max_{1\le k\le n} \min_{0\le j\neq\ell\le K-1} S_k^n(\ell, j) \ge h\right\},\ \ S_k^n(\ell, j) = \sum_{i=k}^{n} \log\frac{f_\ell(Y_i)}{f_j(Y_i)}.
\end{aligned}$$

The generalized matrix recursive CUSUM test, which also attains the asymptotic lower bound, has been considered in [6, 7]. Let us introduce the following stopping time and final decision

$$\widehat{N}_r = \min\{\widehat{N}^1, \ldots, \widehat{N}^{K-1}\}; \quad \tilde{d}_r = \operatorname{argmin}\{\widehat{N}^1, \ldots, \widehat{N}^{K-1}\}$$

of the detection/isolation algorithm. The stopping time $\widehat{N}^\ell$ is responsible for the detection of hypothesis $\mathcal{H}_l$:

$$\begin{aligned}
\widehat{N}^\ell &= \inf\left\{n \ge 1 : \min_{0\le k\neq j\le K-1} Q_n(\ell, j) \ge h\right\}, \\
Q_n(\ell, j) &= (Q_{n-1}(\ell, j) + Z_n(\ell, j))^+, \ \ Z_n(\ell, j) = \log\frac{f_\ell(Y_n)}{f_j(Y_n)}
\end{aligned}$$

For some safety critical applications, a more tractable criterion consists in minimizing the maximum detection/isolation delay:

$$\mathsf{SADD}(\delta) = \max_{1\le j\le K-1} \sup_{0\le\nu<\infty} \mathbb{E}_\nu^j(T-\nu|T>\nu). \tag{7}$$

subject to :

$$\mathbb{C}_{\gamma,\beta} = \left\{ \delta \ : \mathbb{E}_\infty T \geq \gamma, \ \max_{1\leq \ell \leq K-1} \max_{1\leq j\neq \ell \leq K-1} \sup_{\nu \geq 0} \mathbb{P}_\nu^\ell (d = j | T > \nu) \leq \beta \right\}.$$

for $1 \leq \ell, j \neq \ell \leq K-1$. An asymptotic lower bound in this case is given by the following theorem [3].

Theorem 2. *Let $(Y_k)_{k\geq 1}$ be an independent random sequence observed* sequentially *:*

$$\mathcal{L}(Y_k) = \begin{cases} P_0 & \text{if} \quad k \leq \nu \\ P_\ell & \text{if} \quad k \geq \nu + 1 \end{cases}, \quad \nu = 0, 1, 2, \ldots, \ \text{for} \ \ 1 \leq \ell \leq K-1$$

Then

$$\mathsf{SADD}(N; \gamma, \beta) \gtrsim \max \left\{ \frac{\log \gamma}{\rho_{\mathrm{d}}^*}, \ \frac{\log \beta^{-1}}{\rho_{\mathrm{i}}^*} \right\} \ \ \text{as} \ \ \min\{\gamma, \beta^{-1}\} \to \infty,$$

where $\rho_{\mathrm{d}}^* = \min_{1\leq j\leq K-1} \rho_{j,0}$ *and* $\rho_{\mathrm{i}}^* = \min_{1\leq \ell \leq K-1} \min_{1\leq j\neq \ell \leq K-1} \rho_{\ell,j}$.

Vector recursive CUSUM test. If $\gamma \to \infty$, $\beta \to 0$ and $\log \gamma \geq \log \beta^{-1}(1 + o(1))$, then the above mentioned lower bound can be realized by using the following recursive change detection/isolation algorithm [2] :

$$N_r = \min_{1\leq \ell \leq K-1} \{N_r(\ell)\}, \quad d_r = \arg \min_{1\leq \ell \leq K-1} \{N_r(\ell)\},$$

where $N_r(\ell) = \inf \left\{ n \geq 1 : \min_{0\leq j\neq \ell \leq K-1} \left[S_n(\ell, j) - h_{\ell,j} \right] \geq 0 \right\}$,

$$S_n(\ell, j) = g_n(\ell, 0) - g_n(j, 0), \ g_n(\ell, 0) = (g_{n-1}(\ell, 0) + Z_n(\ell, 0))^+ ,$$

with $Z_n(\ell, 0) = \log \frac{f_\ell(Y_n)}{f_0(Y_n)}$, $g_0(\ell, 0) = 0$ for every $1 \leq \ell \leq K-1$ and $g_n(0, 0) \equiv 0$,

$$h_{\ell,j} = \begin{cases} h_d & \text{if} \quad 1 \leq \ell \leq K-1 \quad \text{and} \quad j = 0 \\ h_i & \text{if} \quad 1 \leq j, \ell \leq K-1 \quad \text{and} \quad j \neq \ell \end{cases}.$$

4. Applications to network monitoring

In this section the above mentioned theoretical results are illustrated by application of the proposed detection/isolation procedures to the problem of network monitoring.

Let us consider a network composed of r nodes and n mono-directional links, where y_ℓ denotes the volume of traffic on the link ℓ at discrete time k (see details in [8, 9]). For the sake of simplicity, the subscript k denoting the time is omitted now. Let $x_{i,j}$ be the Origin-Destination (OD) traffic demand from node i to node j at time k. The traffic matrix $X = \{x_{i,j}\}$ is reordered in the lexicographical order as a column vector $X = \left[(x_{(1)}, \ldots, x_{(m)}) \right]^T$, where $m = r^2$ is the number of OD flows.

Let us define an $n \times m$ routing matrix $A = [a_{\ell,k}]$ where $0 \leq a_{\ell,k} \leq 1$ represents the fraction of OD flow k volume that is routed through link ℓ. This leads to the linear model

$$Y = A\, X,$$

where $Y = (y_1, \ldots, y_n)^T$ is the Simple Network Management Protocol (SNMP) measurements. Without loss of generality, the known matrix A is assumed to be of full row rank, i.e., rank $A = n$.

The problem consists in detecting and isolating a significant volume anomaly in an OD flow $x_{i,j}$ by using only SNMP measurements $y_1, \ldots, y_n$. In fact, the main problem with the SNMP measurements is that $n \ll m$. To overcome this difficulty a parsimonious linear model of non-anomalous traffic has been developed in the following papers [10, 11, 12, 13, 14, 15, 16, 17].

The derivation of this model includes two steps: *i)* description of the ambient traffic by using a spatial stationary model and *ii)* linear approximation of the model by using piecewise polynomial splines.

The idea of the spline model is that the non-anomalous (ambient) traffic at each time k can be represented by using a known family of basis functions superimposed with unknown coefficients, i.e., it is assumed that

$$X_k \approx B\mu_k, \quad k = 1, 2, \ldots,$$

where the $m \times q$ matrix B is assumed to be known and $\mu_t \in \mathbb{R}^q$ is a vector of unknown coefficients such that $q < n$. Finally, it is assumed that the model residuals together with the natural variability of the OD flows follow a Gaussian distribution, which leads to the following equation:

$$X_k = B\mu_k + \xi_k \tag{8}$$

where $\xi_k \sim \mathcal{N}(0, \Sigma)$ is Gaussian noise, with the $m \times m$ diagonal covariance matrix $\Sigma = \mathrm{diag}\,(\sigma_1^2, \ldots, \sigma_m^2)$. The advantages of the detection algorithm based on a parametric model of ambient traffic and its comparison to a non-parametric approach are discussed in [11, 14], (see also [18] for PCA based approach). Hence, the link load measurement model is given by the following linear equation :

$$Y_k = A\,B\mu_k + A\xi_k = H\mu_k + \zeta_k + [\theta_\ell], \tag{9}$$

where $Y_k = (y_1, \ldots, y_n)_k{}^T$ and $\zeta_k \sim \mathcal{N}(0, A\Sigma A^T)$. Without any loss of generality, the resulting matrix $H = A\,B$ is assumed to be of full column rank. Typically, when an anomaly occurs on OD flow ℓ at time $\nu + 1$ (change-point), the vector θ_ℓ has the form $\theta_\ell = \varepsilon\, a(\ell)$, where $a(\ell)$ is the ℓ-th normalized column of A and ε is the intensity of the anomaly. The goal is to detect/isolate the presence of an anomalous vector θ_ℓ, which cannot be explained by the ambient traffic model $X_k \approx B\mu_k$.

Therefore, after the de-correlation transformation, the change detection/isolation problem is based on the following model with nuisance parameter X_k :

$$Y_k = HX_k + \xi_k + \theta(k, \nu), \quad \xi_k \sim \mathcal{N}(0, \sigma^2 I_n), \quad k = 1, 2, \ldots, \tag{10}$$

where H is a full rank matrices of size $n \times q$, $n > q$, and $\theta(k, \nu)$ is a change occurring at time $\nu + 1$, namely :

$$\theta(k, \nu) = \begin{cases} 0 & \text{if} \quad k \neq \nu \\ \theta_\ell & \text{if} \quad k \geq \nu + 1 \end{cases}, \; 1 \leq \ell \leq K - 1.$$

This problem is invariant under the group $G = \{Y \to g(Y) = Y + HX\}$ (see details in [19]). The invariant test is based on maximal invariant statistics. The solution is the projection of Y on the orthogonal complement $R(H)^{\perp}$ of the column space $R(H)$ of the matrix H. The parity vector $Z = WY$ is a maximal invariant to the group G.

$$WH = 0, \;\; W^T W = P_H = I_r - H(H^T H)^{-} H^T, \;\; WW^T = I_{n-q}.$$

Transformation by W removes the interference of the nuisance parameter X

$$Z = WY = W\xi \; (+W\theta).$$

Hence, the sequential change detection/isolation problem can be re-written as

$$Z_k = WY_k = W\xi_k + \; W\theta(k, \nu), \;\; \xi_k \sim \mathcal{N}(0, \sigma^2 I_{n-q}), \;\; k = 1, 2, \ldots.$$

Theorem 3. *Let $(Y_k)_{k\geq 1}$ be the output of the model given by* (10) *observed* sequentially. *Then the generalized CUSUM or matrix recursive CUSUM tests attain the lower bound corresponding to the minimax setup :*

$$\mathsf{ESADD}(N; \gamma) \;\gtrsim\; \frac{\log \gamma}{\overline{\rho^*}} \;\; as \;\; \gamma \to \infty, \;\; \overline{\rho^*} \stackrel{def.}{=} \inf_{X^\ell, X^j} \min_{1\leq \ell \leq K-1} \; \min_{0 \leq j \neq \ell \leq K-1} \rho_{\ell,j}(X^\ell, X^j)$$

where X^ℓ (resp. X^j) corresponds to the hypothesis $\mathcal{H}_\ell$ (resp. $\mathcal{H}_j$). The vector recursive CUSUM test attains the lower bound

$$\mathsf{SADD}(N; \gamma, \beta) \gtrsim \max\left\{ \frac{\log \gamma}{\overline{\rho^*}_{\mathrm{d}}}, \;\; \frac{\log \beta^{-1}}{\overline{\rho^*}_{\mathrm{i}}} \right\} \; as \; \gamma \to \infty, \; \beta \to 0, \log \gamma \geq \log \beta^{-1}(1 + o(1)),$$

where

$$\overline{\rho^*}_{\mathrm{d}} = \inf_{X^j, X^0} \min_{1\leq j \leq K-1} \rho_{j,0}(X^j, X^0) \;\; and \;\; \overline{\rho^*}_{\mathrm{i}} = \inf_{X^\ell, X^j} \min_{1\leq \ell \leq K-1} \; \min_{1 \leq j \neq \ell \leq K-1} \rho_{\ell,j}(X^\ell, X^j).$$

5. Acknowledgement

This work was partially supported by the French National Research Agency (ANR) through ANR CSOSG Program (Project ANR-11-SECU-0005).

REFERENCES

1. Nikiforov, I.V., A generalized change detection problem. *IEEE Transactions on Information Theory*, 41(1) : 171-187, 1995.
2. Nikiforov, I.V., A simple recursive algorithm for diagnosis of abrupt changes in random signals. *IEEE Transactions on Information Theory*, 46 (7) :2740-2746, 2000.
3. Nikiforov, I.V. A lower bound for the detection/isolation delay in a class of sequential tests, *IEEE Transactions on Information Theory*, 49 (11), p. 3037-3047, 2003.

4. A. Tartakovsky, I. Nikiforov and M. Basseville (2014) Sequential Analysis : Hypothesis Testing and Changepoint Detection, CRC Press, Taylor & Francis Group.
5. Lorden, G. Procedures for reacting to a change in distribution, *Annals Math. Statistics,* vol.42, pp. 1897-1908, 1971.
6. Oskiper, T. and Poor, H. V. Online activity detection in a multiuser environment using the matrix CUSUM algorithm, *IEEE Transactions on Information Theory*, 48 (2), p. 477–493, 2002
7. Tartakovsky, A. G. Multidecision Quickest Change-Point Detection: Previous Achievements and Open Problems, *Sequential Analysis*, 27, p. 201–231, 2008
8. Lakhina, A. *et al.*, Diagnosing network-wide traffic anomalies, in *SIGCOMM*, 2004.
9. Zhang, Y. *et al.* (2005): Network anomography, in *IMC'05*.
10. Fillatre, L., Nikiforov, I., Vaton, S. Sequential Non-Bayesian Change Detection-Isolation and Its Application to the Network Traffic Flows Anomaly Detection, in Proceedings of the 56th Session of ISI, Lisboa, 22-29 August 2007, pp. 1-4, (special session).
11. Casas, P., Fillatre, L., Vaton, S., Nikiforov, I. Volume Anomaly Detection in Data Networks: an Optimal Detection Algorithm vs. the PCA Approach. International Workshop on Traffic Management and Traffic Engineering for the Future Internet, FITraMEn'08, 11-12 Decembre 2008, Porto. 8 p.
12. Fillatre, L., Nikiforov, I., Vaton, S., Casas, P. Network Traffic Flows Anomaly Detection and Isolation. 4th edition of the International Workshop on Applied Probability, IWAP 2008, 7-10 July 2008, Compiègne. 1-6 p, (invited paper).
13. Fillatre, L., Nikiforov, I., Casas, P., Vaton, S. Optimal volume anomaly detection in network traffic flows. 16th European Signal Processing Conference (EUSIPCO'08), 25-29 August 2008, Lausanne. 5 p.
14. Casas, P., Fillatre, L., Vaton S., Nikiforov I. Volume anomaly detection in data networks : an optimal detection algorithm vs the PCA approach. *Lecture Notes in Computer Science*, 2009, vol. 5464, n. 96, pp. 96-113
15. Casas, P., Vaton, S., Fillatre, L., Nikiforov, I. V. Optimal Volume Anomaly Detection and Isolation in Large-Scale IP Networks Using Coarse-Grained Measurements, *Computer Networks*, vol. 54, pp. 1750-1766, 2010.
16. Casas, P., Fillatre, L., Vaton, S., Nikiforov, I. Reactive Robust Routing: Anomaly Localization and Routing Reconfiguration for Dynamic Networks. Journal of Network and Systems Management, vol.19, n. 1, p. 58-83, 2010.
17. Fillatre, L., Nikiforov, I. Asymptotically Uniformly Minimax Detection and Isolation in Network Monitoring. *IEEE Transactions on Signal Processing*, vol. 60, no. 7, July, pp. 3357-3371, 2012.
18. Ringberg, H. *et al.* (2007) Sensitivy of PCA for traffic anomaly detection, in *SIGMETRICS*.
19. M. Fouladirad and I. Nikiforov, "Optimal statistical fault detection with nuisance parameters," *Automatica*, vol. 41, no. 7, pp. 1157–1171, July 2005.

HEURISTIC SOLUTION FOR THE OPTIMAL THRESHOLDS IN A CONTROLLABLE MULTI-SERVER HETEROGENEOUS QUEUEING SYSTEM WITHOUT PREEMPTION[1]

Dmitry Efrosinin[1 2], *Vladimir Rykov*[3]

[1]Johannes Kepler University Linz, Linz, Austria
[2]V.A. Trapeznikov Institute of Control Sciences, RAS, Moscow, Russia
[3]Gubkin Russian State University of Oil and Gas, Moscow, Russia
dmitry.efrosinin@jku.at, vladimir_rykov@mail.ru

Abstract

As it known the optimal policy which minimizes the long-run average cost per unit of time in a multi-server queueing system with heterogeneous servers without preemption has a threshold structure. It means that the slower server must be activated whenever all faster servers are busy and the number of customers in the queue exceeds some specified for this server threshold level. The optimal thresholds can be evaluated using the Howard iteration algorithm or by minimizing the function of the average cost which can be obtained in closed form as a function of unknown threshold levels. The both cases have sufficient restrictions on dimensionality of the model. In present paper we provide a heuristic method to derive expressions for the optimal threshold levels in explicit form as functions of system parameters like service intensities, usage and holding costs for an arbitrary number of servers. The proposed method is based on the fitting of the boundary planes between the areas where the optimal threshold takes a certain value.

Keywords: Controllable queueing system, heterogeneous servers, long-run average cost, threshold policy, optimal allocation

1. Introduction

Controllable queueing systems with heterogeneous servers and single queue find wide application in various fields of human activity including resource allocation in telecommunication and computer networks, control problems in production lines and so on. For detailed review of the literature on heterogeneous queues the reader is referred to [2]. The queues with preemption or priority interruption of service, i.e. when the customer can change the server during the service process, are well studied, see e.g [6]. The optimal allocation mechanism for customers in such systems is defined normally through a simple Fastest Free Server policy, when a new customer is transfered to the fastest

[1]This work was funded by the Russian Foundation for Basic Research (RFBR), Project №15-08-08677-a.

available server even if it is already under service on some slower server. In this case the system can be exhaustively studied independently of the number of servers in the system. In contrast for the queueing systems without preemption the Fastest Free Server allocation policy occurs to be not the best one. As it was shown in papers [8, 9] the optimal policy, which minimizes the long-run average cost per unit of time, is of threshold type, i.e. for any server exists some threshold level, which specifies the number of customers in the queue when the server must be activated. There are a number of methods how to calculate the optimal threshold levels of the control policy. They can be evaluated numerically by means of the Howard iteration algorithm [5] or by numerical minimization of the average cost function evaluated in closed form, see e.g. [3, 4]. In both cases we have problems with dimensionality of the model if the number of servers is relative high. Therefore a natural question arises: Is there a possibility to get some heuristic solution which will be valid for an arbitrary number of servers.

In framework of the paper we solve such a problem and construct formulas for the optimal threshold levels which seems to be quite appropriate for the studied systems. The kernel element of the proposed heuristic method consists in evaluation of the average cost function in closed form as a function of unknown threshold levels. It is performed by the method of the difference equations. Then we analyze the boundaries between the optimality regions where the fixed threshold level takes a certain value. It can be done at least for two and three server models but obtained results can be generalized to the case of an arbitrary number of servers.

The rest of the paper is organized as follows. Section 2 describes the mathematical model based on a homogeneous Markov process under fixed threshold policy. In Section 3 the stationary state probabilities are evaluated by means of the difference equations approach. Section 4 deals with optimization problem. Heuristic method for the optimal threshold level estimation is presented in Section 5. Finally, some numerical examples are illustrated in Section 6.

In further sections we will use the notations $\boldsymbol{e}_j$ for the vector with 1 in the j-th (beginning from 0-th) position and 0 elsewhere, $1_{\{A\}}$ for the indicator function, where $1_{\{A\}} = 1$ if the condition A holds, and 0 otherwise, $\overline{a, b}$ for the elements of the set $[a, b] \cap \mathbb{N}_0$.

2. Mathematical model

Consider a queueing system where the customers arrive according to the Poisson process with intensity λ and K heterogeneous servers have exponentially distributed service times with intensities $\mu_1, \mu_2, \ldots, \mu_K$. The service of the customers is assumed to be without preemption, i.e. the customer being served on some server can not change this server. The interarrival and service times are assumed to be mutually independent. Define the control policy

$$f = (q_1, q_2, \ldots, q_K),$$

which prescribes how to allocate the customers between the servers and is of threshold type defined as a sequence of threshold levels

$$1 = q_1 \le q_2 \le \cdots \le q_K < \infty.$$

According to this policy the first k servers must be occupied whenever there are q customers in the queue and $q = \overline{q_k, q_{k+1} - 1}$. The cost structure consists of the following components:
c_0 – the holding cost per unit of time for any waiting customer in the queue,
c_j – the usage cost per unit of time for any busy server j.
The servers are enumerated in such a way that

$$\begin{aligned} &0 < c_1\mu_1^{-1} \le c_2\mu_2^{-1} \le \cdots \le c_K\mu_K^{-1}, \\ &0 < \mu_1^{-1} \le \mu_2^{-1} \le \cdots \le \mu_K^{-1}, \end{aligned} \tag{1}$$

where the ratio $c_j\mu_j^{-1}$ stands for the average usage cost of the jth server. The states of the system at time t are described by the vector $\{Q(t), D(t)\}_{t\ge 0}$, where $Q(t)$ – the number of customers in the queue and $D(t) = \{D_1(t), \ldots, D_K(t)\}$ – the states of the servers,

$$D_j(t) = \begin{cases} 0, & \text{the server } j \text{ is idle,} \\ 1, & \text{the server } j \text{ is busy} \end{cases}, \quad j = \overline{1, K}.$$

The multidimensional random process

$$\{X(t)\}_{t\ge 0} = \{Q(t), D(t)\}_{t\ge 0} \tag{2}$$

is a homogeneous Markov process. The state space of the process $\{X(t)\}_{t\ge 0}$ is

$$E = \{x = (q, d); q \in \mathbb{N}_0,\, d \in D(q)\}. \tag{3}$$

Note that the state space of the servers $D(q)$ depends on the queue length q,

$$D(q) = \left\{ d;\ \begin{array}{l} d_j \in \{0,1\}, j = \overline{1,K},\ q = 0 \\ d_j = 1, d_i \in \{0,1\},\ 1 \le j \le k \le i-1 \le K-1,\ q = \overline{q_k, q_{k+1}-1} \\ d_j = 1, j = \overline{1,K},\ q \ge q_K \end{array} \right\},$$

and the number of states $|D(q)|$ is

$$|D(q)| = \begin{cases} 2^K, & q = 0, \\ 2^{K-k}, & q = \overline{q_k, q_{k+1} - 1},\ k = \overline{1, K-1}, \\ 1, & q \ge q_K. \end{cases}$$

Denote by $A_{q_k}(x)$ and $\overline{A}_{q_k}(x)$, $k = \overline{1,K}$, the following events and their complements,

$$\begin{aligned} A_{q_k}(x) &= \{q(x) \ge q_k,\, d_i(x) = 1,\, i = \overline{1, k-1},\, d_k(x) = 0\}, \\ \overline{A}_{q_k}(x) &= \{q(x) \le q_k - 1,\, d_i(x) = 1,\, i = \overline{1, k-1},\, d_k(x) = 0\}. \end{aligned} \tag{4}$$

Analyzing the transitions of the Markov process $\{X(t)\}_{t\geq 0}$ and using the notations (4) we get the system of balance equations for the stationary state probabilities π_x, $x=(q,d)\in E$,

$$\pi_x = \lim_{t\to\infty} \mathbb{P}[X(t)=x]$$

in the form

$$\begin{aligned}
&\Big(\lambda + \sum_{k=1}^{K} d_k(x)\mu_k\Big)\pi_x = \qquad (5)\\
&= \lambda\Big(\sum_{k=1}^{K} \pi_{x-\boldsymbol{e}_k} 1_{\{A_{q_k-1}(x-\boldsymbol{e}_k)\}} + \pi_{x-\boldsymbol{e}_0} 1_{\{\overline{A}_{q_k-1}(x-\boldsymbol{e}_0)\}}\Big) +\\
&+ \sum_{k=1}^{K} d_k(x)\mu_k \pi_{x+\boldsymbol{e}_0} 1_{\{A_{q_k}(x+\boldsymbol{e}_0)\}} + \sum_{k=1}^{K}(1-d_k(x))\mu_k \pi_{x+\boldsymbol{e}_k} 1_{\{\overline{A}_{q_k}(x+\boldsymbol{e}_k)\}}.
\end{aligned}$$

To get a solution of this system for the fixed threshold policy in closed form we intend to apply the method of the difference equations.

3. Evaluation of the stationary state probabilities

Theorem 1. *For the two-server queueing system $M/M/2$ with a threshold policy $f=(1,q_2)$ the stationary state probabilities satisfy the relations,*

$$\begin{aligned}
&\pi_{(0,0,0)} = [R_{q_2+1}^{-1} B_{q_2+1} - 1]\pi_{(0,0,1)}, \qquad (6)\\
&\pi_{(q,1,0)} = [R_{k-q_2} B_{q_2+1} - B_{q+1}]\pi_{(0,0,1)},\ k=\overline{0,q_2-1},\\
&\pi_{(q,1,1)} = B_{q+1}\pi_{(0,0,1)},\ q=\overline{0,q_2-1},\\
&\pi_{(q,1,1)} = N_{q-q_2+1}\pi_{(q_2-1,1,1)},\ q\geq q_2\\
&\pi_{(0,0,1)} = \left[R_{q_2+1}B_{q_2+1} + \frac{R_{q_2-1}B_{q_2+1}}{R_{q_2+1}-R_{q_2}} + \frac{N_1 B_{q_2+1}}{1-N_1}\right]^{-1},
\end{aligned}$$

where

$$\begin{aligned}
&R_n = \Big(\frac{\lambda}{\mu_1}\Big)^n,\ N_n = \Big(\frac{\lambda}{\mu_1+\mu_2}\Big)^n,\ B_n = b_1\beta_1^n + b_2\beta_2^n,\\
&\beta_{1,2} = \frac{(\lambda+\mu_1+\mu_2) \pm \sqrt{(\lambda+\mu_1+\mu_2)^2 - 4\lambda\mu_1}}{2\mu_1},\\
&b_1 = \frac{1-\beta_1}{\beta_2-\beta_1},\ b_2 = \frac{\beta_2-1}{\beta_2-\beta_1}.
\end{aligned}$$

Proof. Due to threshold structure of the control policy the system of balance equations (5) are divided into separate subsystems which can be treated as

homogeneous difference equations.
Subsystem 1.

$$\begin{aligned}
&(\lambda+\mu_2)\pi_{(0,0,1)} = \mu_1\pi_{(0,1,1)}, \\
&(\lambda+\mu_1+\mu_2)\pi_{(0,1,1)} = \lambda\pi_{(0,0,1)} + \mu_1\pi_{(1,1,1)}, \\
&(\lambda+\mu_1+\mu_2)\pi_{(q,1,1)} = \lambda\pi_{(q-1,1,1)} + \mu_1\pi_{(q+1,1,1)},\ q = \overline{1, q_2-2}.
\end{aligned} \tag{7}$$

For this system the solution is assumed to be in form

$$\pi_{(q,1,1)} = \beta^{q+1}\pi_{(0,0,1)},\ q = \overline{0, q_2-1}. \tag{8}$$

The substitution of this solution to the equation of the previous system for the arbitrary q and subsequent division by β^{q-1} lead to the equation

$$\mu_1\beta^2 - (\lambda+\mu_1+\mu_2)\beta + \lambda = 0.$$

This quadratic equation has two real roots,

$$\beta_{1,2} = \frac{(\lambda+\mu_1+\mu_2) \pm \sqrt{(\lambda+\mu_1+\mu_2)^2 - 4\mu_1}}{2\mu_1}.$$

Therefore the probabilities $\pi_{(k,1,1)}$ as solutions of the difference equations must satisfy the relation

$$\pi_{(q,1,1)} = B_{q+1}\pi_{(0,0,1)},\ q = \overline{0, q_2-1}, \tag{9}$$

where

$$B_n = (b_1\beta_1^n + b_2\beta_2^n).$$

From the first equation of the subsystem (6) for the probability $\pi_{(0,1,1)}$ we get

$$\begin{aligned}
&b_1\beta_1 + b_2\beta_2 = \beta_1 + \beta_2 - 1, \\
&b_1 + b_2 = 1,
\end{aligned}$$

whence it appears the relations for the constants $b_1 \;\; b_2$.
Subsystem 2.
For the states $x \in E$, where the queue length exceeds the level q_2, we have

$$(\lambda+\mu_1+\mu_2)\pi_{(q,1,1)} = \lambda\pi_{(q-1,1,1)} + (\mu_1+\mu_2)\pi_{(q+1,1,1)},\ q \geq q_2. \tag{10}$$

Assuming the solution of these difference equations in the form

$$\pi_{(q,1,1)} = \eta^{q-q_2+1}\pi_{(q_2-1,1,1)}$$

and substituting of this result to the equality (10), it is easy to show, that there is exist the only one non-trivial solution of the quadratic equation, $\eta = \frac{\lambda}{\mu_1+\mu_2}$, i.e.

$$\pi_{(q,1,1)} = N_{q-q_2+1}\pi_{(q_2-1,1,1)},\ q \geq q_2, \tag{11}$$

where

$$N_n = \eta^n = \Big(\frac{\lambda}{\mu_1+\mu_2}\Big)^n.$$

Subsystem 3.
Consider the following subsystem of the system of balance equations,

$$\begin{aligned}
&(\lambda+\mu_1)\pi_{(0,1,0)} = \lambda\pi_{(0,0,0)} + \mu_1\pi_{(1,1,0)} + \mu_2\pi_{(0,1,1)}, \\
&(\lambda+\mu_1)\pi_{(q,1,0)} = \lambda\pi_{(q-1,1,0)} + \mu_1\pi_{(q+1,1,0)} + \mu_2\pi_{(q,1,1)},\ q=\overline{1,q_2-2}, \\
&(\lambda+\mu_1)\pi_{(q_2-1,1,0)} = \lambda\pi_{(q_2-2,1,0)} + \mu_2\pi_{(q_2-1,1,1)}, \\
&(\lambda+\mu_1+\mu_2)\pi_{(q_2-1,1,1)} = \lambda\pi_{(q_2-2,1,1)} + \lambda\pi_{(q_2-1,1,0)} + (\mu_1+\mu_2)\pi_{(q_2,1,1)}.
\end{aligned} \tag{12}$$

To solve the system the corresponding equations are summed up for the probabilities $\pi_{(q,1,0)}$ and $\pi_{(q,1,1)}$. As a result we obtain the following equations

$$\pi_{(q,1,1)} + \pi_{(q,1,0)} = R_{q+1-q_2}(\pi_{(q_2-1,1,1)} + \pi_{(q_2-1,1,0)}),\ q=\overline{0,q_2-1},$$

where $R_n = \Big(\frac{\lambda}{\mu_1}\Big)^n$. Now we rewrite the last equation of the subsystem (12),

$$\lambda(\pi_{(q_2-1,1,1)} + \pi_{(q_2-1,1,0)}) = (\lambda+\mu_1+\mu_2)\pi_{(q_2-1,1,1)} - \lambda\pi_{(q_2-2,1,1)}.$$

But from the system (7) it follows,

$$(\lambda+\mu_1+\mu_2)\pi_{(q_2-1,1,1)} - \pi_{(q_2-2,1,1)} = \frac{\mu_1}{\lambda}(b_1\beta_1^{q_2+1} + b_2\beta_2^{q_2+1}),$$

that in turn leads to the relation

$$\begin{aligned}
\pi_{(q,1,0)} &= [R_{q-q_2}(b_1\beta_1^{q_2+1} + b_2\beta_2^{q_2+1}) - (b_1\beta_1^{q+1} + b_2\beta_2^{q+1})]\pi_{(0,0,1)} = \\
&= [R_{q-q_2}B_{q_2+1} - B_{q+1}]\pi_{(0,0,1)},\ q=\overline{0,q_2-1}.
\end{aligned} \tag{13}$$

Finally from

$$\lambda\pi_{(0,0,0)} = \mu_1\pi_{(0,1,0)} + \mu_2\pi_{(0,0,1)}$$

we get

$$\begin{aligned}
\pi_{(0,0,0)} &= [\Big(\frac{\mu_1}{\lambda}\Big)^{q_2+1}(b_1\beta_1^{q_2+1} + b_2\beta_2^{q_2+1}) - 1]\pi_{(0,0,1)} \\
&= [R_{q_2+1}^{-1}B_{q_2+1} - 1]\pi_{(0,0,1)}.
\end{aligned} \tag{14}$$

■

Theorem 2. *For the three-server queueing system $M/M/3$ with a threshold policy $f = (1, q_2, q_3)$ the stationary state probabilities satisfy the following relations,*

$$\pi_{(0,0,1,1)} + \pi_{(0,0,1,0)} = \Big(\frac{(R_{q_2+1} - 1)B_{q_2+1}}{R_{q_2+2} - R_{q_2+1}} + \frac{(N_{q_3-q_2+1} - N_1)B_{q_2}}{N_1 - 1} + \\ + \frac{\Phi_1 H_{q_3} N_{q_3-q_2+1} B_{q_2}}{(1-\Phi_1)H_{q_3+1}} \Big)^{-1}, \tag{15}$$

$$\pi_{(0,0,1,1)} = \frac{N_{q_3-q_2+1}B_{q_2}}{H_{q_3+1}A_{q_2}}[\pi_{(0,0,1,1)} + \pi_{(0,0,1,0)}],$$

$$\pi_{(0,0,0,1)} = (G_{q_2}^{-1}[A_{q_2} + R_1^{-1}A_{q_2+1} - N_1^{-1}H_{q_2+1}A_{q_2}] - 1)\pi_{(0,0,1,1)},$$

$$\pi_{(0,0,0,0)} = (R_{1+q_2}^{-1}B_{q_2+1} - 1)[\pi_{(0,0,1,1)} + \pi_{(0,0,1,0)}] - \pi_{(0,0,0,1)}$$

$$\pi_{(q,1,0,0)} = (R_{q_2-q}^{-1}B_{q_2+1} - B_{q+1})[\pi_{(0,0,1,1)} + \pi_{(0,0,1,0)}] \\ - G_{q+1}[\pi_{(0,0,1,1)} + \pi_{(0,0,0,1)}] + A_{q+1}\pi_{(0,0,1,1)}, \quad q = \overline{0, q_2 - 1},$$

$$\pi_{(q,1,0,1)} = G_{q+1}[\pi_{(0,0,1,1)} + \pi_{(0,0,0,1)}] - A_{q+1}\pi_{(0,0,1,1)}, \quad q = \overline{0, q_2 - 1},$$

$$\pi_{(q,1,1,0)} = B_{q+1}[\pi_{(0,0,1,1)} + \pi_{(0,0,1,0)}] - A_{q+1}\pi_{(0,0,1,1)}, \quad q = \overline{0, q_2 - 1},$$

$$\pi_{(q,1,1,0)} = N_{q+1-q_2}B_{q_2}\pi_{[}\pi_{(0,0,1,1)} + \pi_{(0,0,1,0)}] - H_{q+1}A_{q_2}\pi_{(0,0,1,1)}, \quad q = \overline{q_2, q_3 - 1},$$

$$\pi_{(q,1,1,1)} = A_{q+1}\pi_{(0,0,1,1)}, \quad q = \overline{0, q_2 - 1},$$

$$\pi_{(q,1,1,1)} = H_{q+1}A_{q_2}\pi_{(0,0,1,1)}, \quad q = \overline{q_2, q_3 - 1},$$

$$\pi_{(q,1,1,1)} = \Phi_{q-q_3+1}H_{q_3}A_{q_2}\pi_{(0,0,1,1)}, \quad q \geq q_3,$$

where

$$R_n = \Big(\frac{\lambda}{\mu_1}\Big)^n,\ N_n = \Big(\frac{\lambda}{\mu_1 + \mu_2}\Big)^n,\ \Phi_n = \Big(\frac{\lambda}{\mu_1 + \mu_2 + \mu_3}\Big)^n,$$

$$A_n = a_1\alpha_1^n + a_2\alpha_2^n,\ B_n = b_1\beta_1^n + b_2\beta_2^n,\ G_n = g_1\gamma_1^n + g_2\gamma_2^n,\ H_n = h_1\xi_1^n + h_2\xi_2^n,$$

$$\alpha_{1,2} = \frac{(\lambda + \mu_1 + \mu_2 + \mu_3) \pm \sqrt{(\lambda + \mu_1 + \mu_2 + \mu_3)^2 - 4\lambda\mu_1}}{2\mu_1},$$

$$\beta_{1,2} = \frac{(\lambda + \mu_1 + \mu_2) \pm \sqrt{(\lambda + \mu_1 + \mu_2)^2 - 4\lambda\mu_1}}{2\mu_1},$$

$$\gamma_{1,2} = \frac{(\lambda + \mu_1 + \mu_3) \pm \sqrt{(\lambda + \mu_1 + \mu_3)^2 - 4\lambda\mu_1}}{2\mu_1},$$

$$\xi_{1,2} = \frac{(\lambda + \mu_1 + \mu_2 + \mu_3) \pm \sqrt{(\lambda + \mu_1 + \mu_2 + \mu_3)^2 - 4\lambda(\mu_1 + \mu_2)}}{2(\mu_1 + \mu_2)},$$

$$a_1 = \frac{1 - \alpha_1}{\alpha_2 - \alpha_1},\ a_2 = \frac{\alpha_2 - 1}{\alpha_2 - \beta_1},\ b_1 = \frac{1 - \beta_1}{\beta_2 - \beta_1},\ b_2 = \frac{\beta_2 - 1}{\beta_2 - \beta_1},$$

$$g_1 = \frac{1 - \gamma_1}{\gamma_2 - \gamma_1},\ g_2 = \frac{\gamma_2 - 1}{\gamma_2 - \gamma_1},\ h_1 = \frac{L_2 - L_1}{K_1 - K_2},\ h_2 = 1 - h_1,$$

$$K_1 = \Big[\Big(\frac{1}{\xi_2} - 1\Big)\xi_1^{q_3-q_2} - \Big(\frac{1}{\xi_1} - 1\Big)\xi_2^{q_3-q_2}\Big]A_{q_2},$$

$$L_1 = \Big(\frac{1}{\xi_1} - 1\Big)\xi_2^{q_3-q_2}A_{q_2},$$

$$K_2 = \frac{\mu_3}{\lambda}A_{q_2}\Big(\frac{\xi_1 - \xi_1^{q_3-q_2+1}}{1-\xi_1} - \frac{\xi_2 - \xi_2^{q_3-q_2+1}}{1-\xi_2} - G_{q_2}^{-1}\frac{\gamma_2^{q_2+1} - \gamma_1^{q_2+1}}{\gamma_2 - \gamma_1}\Big(\frac{1}{\xi_1} - \frac{1}{\xi_2}\Big)\Big),$$

$$L_2 = \frac{\mu_3}{\lambda}G_{q_2}^{-1}\frac{\gamma_2^{q_2+1} - \gamma_1^{q_2+1}}{\gamma_2 - \gamma_1}\Big(\Big(1 - \frac{1}{\xi_1}\Big)A_{q_2} + R_1^{-1}A_{q_2+1}\Big) + \frac{\mu_3}{\lambda}\frac{\xi_2 - \xi_2^{q_3-q_2+1}}{1-\xi_2}A_{q_2}.$$

Proof. The proof can be performed in a similar way as before by dividing the system of balance equations into subsystems, which are solved as homogeneous difference equations. ■

4. Optimization problem for performance characteristics

For every fixed threshold policy f we wish to guarantee that the process $\{X(t)\}_{t\geq 0}$ with a state space E is an irreducible, positive recurrent Markov process defined through its infinitesimal matrix $\Lambda = [\lambda_{xy}(q_2, \ldots, q_K)]$, which depends on the threshold policy. The immediate cost $c(x)$ of the specified controllable Markov model is defined as

$$c(x) = c_0 q_0(x) + \sum_{j=1}^{K} c_j d_j(x),$$

where the notations $q(x)$ and $d_j(x)$ stand for the elements of the vector state $x \in E$. As it is known [11], for ergodic Markov process with costs the long-run average cost per unit of time for the policy f coincides with corresponding assemble average,

$$g^f = \lim_{t\to\infty}\frac{1}{t}V^f(x,t) = \sum_{y\in E} c(y)\pi_y^f, \tag{16}$$

where

$$V^f(x,t) = \int_0^t \sum_{y\in E} \mathbb{P}^f[X(u) = y | X(0) = x]c(y)du \tag{17}$$

denotes the total average cost up to time t when the process starts in state x and $\pi_y^f = \mathbb{P}^f[X(t) = y]$ denotes a stationary probability of the process given policy f. The policy f^* is said to be optimal when for any available policy f

$$g^{f^*} = \min_f g^f. \tag{18}$$

For existence of optimal policies we refer to Aviv and Federgruen [1], Puterman [7], Sennott [10]. Obviously the optimal policy exists if $\sum_{y\in E}|c(y)\pi_y^f| < \infty$.

For the model under study it is always the case, since the infinite part of this sum is definitely finite,

$$\sum_{q=0}^{\infty}\Big[c_0(q+q_K-1)+\sum_{j=1}^{K}c_j\Big]\rho^q\pi_{(q_{K-1},d)}=$$

$$c_0\frac{q_K-1-\rho(q_K-2)}{(1-\rho)^2}+\sum_{j=1}^{K}c_j\frac{1}{1-\rho}<\infty,$$

if

$$\rho=\frac{\lambda}{\sum_{j=1}^{K}\mu_j}<1,$$

which coincides with a stability condition of the system. Hence we can obtain the explicit solution for the function $g^f := g(q_2,\dots,q_K)$ at least up to the case of three servers. These formulas are used to get a heuristic solution.

Remark 1. The method based on a solution of difference equations can be applied theoretically to the system with a higher number of servers as well, although it requires to solve a larger number of subsystems. The number of rows in the system of balance equations is equal to $\sum_{k=0}^{K}(q_{k+1}-q_k)2^{K-k}$, where $q_0=0$ and $q_{K+1}=N, N$ - maximal possible number of customers in the system. It is clear that the calculation of the stationary state probabilities and minimization of the function g over K parameters is computationally not feasible for large K, what really motivates us to look for some heuristic solution.

5. Heuristic solution for the optimal thresholds

As it was shown in [2] the optimal thresholds $q_k, k=\overline{2,K}$, can be calculated by means of the Howard iteration algorithm. But it has sufficient restrictions on dimensionality of the model and number of states. In this section we propose a heuristic method for optimal thresholds estimation, which provide us with explicit formulas depending on the system parameters. Two types of model are discussed: With $\lambda=0$ and $\lambda>0$. In the first case the model reduces to the equivalent scheduling problem. It is assumed here that there are a number of customers in the system and no new customers join the system. The optimization problem consists in allocation of the customers between the heterogeneous servers with the aim to minimize the total average cost function until the system becomes empty. For the scheduling problem the optimal thresholds q_k are evaluated exactly. For the original problem with new arrivals the optimal thresholds can be only estimated. Since we are not able to give the rigorous proofs to all our results, they are summarized in the following conjecture. The details and main principles of the used approach are given afterwards.

Conjecture 1. *The optimal thresholds q_k, $k = \overline{2, K}$, are defined by*

$$q_k \approx \hat{q}_k = \left\lfloor \frac{1}{c_0} \left[\frac{c_k}{\mu_k} F_k - \sum_{j=1}^{k-1} c_j \right] \right\rfloor, \text{ where} \tag{19}$$

$$F_k = \begin{cases} \sum\limits_{j=1}^{k-1} \mu_j, & \lambda = 0, \\ \frac{1}{2}\Big[\sum\limits_{j=1}^{k-1} \mu_j - \lambda + \sqrt{\Big(\sum\limits_{j=1}^{k-1} \mu_j - \lambda \Big)^2 + 4(k-1)\mu_k \lambda} \Big], & \lambda > 0. \end{cases} \tag{20}$$

Proof. Consider the case $\lambda = 0$. Due to our assumption about threshold structure of the control policy f, we can calculate the total average cost until the system becomes empty recursively. Assuming the known values of thresholds $q_2, \ldots, q_{k-1}$ we obtain the value of q_k. Obviously for the state $x = (0, \ldots, 0)$ we have

$$V(x) = 0,$$
$$V(x + \boldsymbol{e}_1) = \frac{c_1}{\mu_1},$$
$$V(x + \boldsymbol{e}_0 + \boldsymbol{e}_1) = \frac{c_0 + c_1}{\mu_1} + V(x + \boldsymbol{e}_1) = \frac{c_0}{\mu_1} + \frac{2c_1}{\mu_1},$$
$$\ldots$$
$$V(x + (q_2 - 1)\boldsymbol{e}_0 + \boldsymbol{e}_1) = \frac{(q_2 - 1)c_0 + c_1}{\mu_1} + V(x + (q_2 - 2)\boldsymbol{e}_0 + \boldsymbol{e}_1) = \frac{q_2(q_2 - 1)c_0}{2\mu_1} + \frac{q_2 c_1}{\mu_1}.$$

When the queue length has reached the level q_2, it becomes optimal to use the second server. Then we have

$$V(x + (q_2 - 1)\boldsymbol{e}_0 + \boldsymbol{e}_1 + \boldsymbol{e}_2) = \frac{c_2}{\mu_2} + \frac{q_2(q_2 - 1)c_0}{2\mu_1} + \frac{q_2 c_1}{\mu_1}.$$

Repeating the procedure up to the level q_k one can obtain

$$V\Big(x + (q_k - 1)\boldsymbol{e}_0 + \sum_{j=1}^{k-1} \boldsymbol{e}_j\Big) = \frac{(q_k - q_{k-1}) \sum_{j=1}^{k-1} c_j}{\sum_{j=1}^{k-1} \mu_j} + \frac{(q_k - q_{k-1})(q_{k-1} + q_k - 1)c_0}{2 \sum_{j=1}^{k-1} \mu_j}.$$

Again, since q_k is an optimal threshold, the condition

$$V\Big(x + (q_k - 1)\boldsymbol{e}_0 + \sum_{j=1}^{k} \boldsymbol{e}_j\Big) = \frac{c_k}{\mu_k} + V\Big(x + (q_k - 1)\boldsymbol{e}_0 + \sum_{j=1}^{k-1} \boldsymbol{e}_j\Big) < v\Big(x + q_k \boldsymbol{e}_0 + \sum_{j=1}^{k-1} \boldsymbol{e}_j\Big)$$

implies

$$\frac{c_k}{\mu_k}+\frac{(q_k-q_{k-1})\sum_{j=1}^{k-1}c_j}{\sum_{j=1}^{k-1}\mu_j}+\frac{(q_k-q_{k-1})(q_{k-1}+q_k-1)c_0}{2\sum_{j=1}^{k-1}\mu_j}<$$
$$\frac{(q_k-q_{k-1}+1)\sum_{j=1}^{k-1}c_j}{\sum_{j=1}^{k-1}\mu_j}+\frac{(q_k-q_{k-1}+1)(q_{k-1}+q_k)c_0}{2\sum_{j=1}^{k-1}\mu_j}.$$

From the last inequality we get

$$q_k=\frac{1}{c_0}\Big[\frac{c_k}{\mu_k}\sum_{j=1}^{k-1}\mu_j-\sum_{j=1}^{k-1}c_j\Big].$$

■

Remark 2. Threshold levels defined by (19) satisfy the inequalities

$$\frac{1}{c_0}\left[\frac{c_k}{\mu_k}\Big(\sum_{j=1}^{k-1}\mu_j-\lambda\Big)-\sum_{j=1}^{k-1}c_j\right]\le\hat{q}_k^*\le\frac{1}{c_0}\left[\frac{c_k}{\mu_k}\sum_{j=1}^{k-1}\mu_j-\sum_{j=1}^{k-1}c_j\right]. \qquad (21)$$

Proof. The left inequality of (21) follows directly from

$$F_k\ge\frac{1}{2}\Big[\sum_{j=1}^{k-1}\mu_j-\lambda+\sqrt{\Big(\sum_{j=1}^{k-1}\mu_j-\lambda\Big)^2}\Big]=\sum_{j=1}^{k-1}\mu_j-\lambda.$$

To prove the inequality at the right hand side it is necessary to show that

$$F_k\le\sum_{j=1}^{k-1}\mu_j.$$

By solving this inequality using simple algebraic manipulations we get

$$(k-1)\mu_k\le\sum_{j=1}^{k-1}\mu_j,$$

which is true due to the ordering (1). ■

To get the formulas for the case $\lambda>0$ consider first the problem of the mean number of customers minimization, i.e. when $c_j=1, j=\overline{0,K}$. To reduce the number of system parameters in the system (5) we divide the right and the left hand side by λ and introduce the notations $r_j=\frac{\mu_j}{\lambda}, j=\overline{1,K}$. The proposed algorithm 1 is based on a functional estimation of the boundaries between the areas where the optimal threshold $q_k, k=\overline{1,K}$, takes a certain value.

Numerical analysis confirms our expectations that the optimal threshold q_k depends only on parameters of the first k servers and the boundaries between the regions of optimality have a linear structure, i.e. they can be represented as hyperplanes.

Algorithm 1.
Step 0. *Calculation of the optimal thresholds* $q_k, k = \overline{1,K}$, *for all possible values of* $(r_1, \ldots, r_K)$ *by means of a function*

$$\bar{N} := \bar{N}(r_1, \ldots, r_K, q_2^*, \ldots, q_K^*) = \sum_{x \in E} \Big(q(x) + \sum_{j=1}^{K} d_j(x) \Big) \pi_x,$$

derived in closed form as discussed in previous section.
Step 1. *Labeling of the regions, where the optimal thresholds* $(q_2, \ldots, q_K)$ *take certain values.*
Step 2. *Identification of the points laying on the boundaries between the regions, where*

$$q_k^* = i \;\; \textit{and} \;\; q_k = i + 1,\; i \geq q_{k-1}.$$

Step 3. *Application of the least-squares method to estimate the unknown coefficients* a_{kj} *of the hyperplanes for each threshold* $q_k, k = \overline{1,K}$,

$$\sum_{j=1}^{k} a_{kj} r_j + a_{kk+1} = 0,$$

and their representation as functions of q_k, $a_{kj} := a_{kj}(q_k^*)$.
Step 4. *Express* $q_k, k = \overline{2,K}$, *through* $r_1, \ldots, r_k$.

If $K = 2$ and $K = 3$, then the geometric visualization of the algorithm is possible. For the two-server model the regions of optimality of the level q_2 and asymptotic lines for the boundaries are illustrated in figures 1(a,b). One can easily identify the boundary lines between the areas,

$$a_{21} r_1 + a_{22} r_2 + a_{23} = 0,$$

where the unknown coefficients are evaluated through coordinates of the two points (x_1, y_1) and (x_2, y_2): $a_{21} = y_2 - y_1, a_{22} = x_1 - x_2, a_{23} = x_2 y_1 - x_1 y_2$. By selecting appropriate coordinates we get the coefficients summarized in the next table:

i	1	2	3	4	5	6	7	8
x_1	6.47	9.65	12.74	15.79	18.83	21.85	24.87	27.90
y_1	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00
x_2	10.48	18.66	28.74	40.79	54.83	70.87	88.87	108.90
y_2	5.00	6.00	7.00	8.00	9.00	10.00	11.00	12.00
a_{21}	2.00	3.00	4.00	5.00	6.00	7.00	8.00	9.00
a_{22}	-4.01	-9.01	-16.00	-25.00	-36.00	-49.01	-64.00	-81.00
a_{23}	-0.91	-1.92	-2.96	-3.95	-4.98	-5.89	-6.96	-8.09

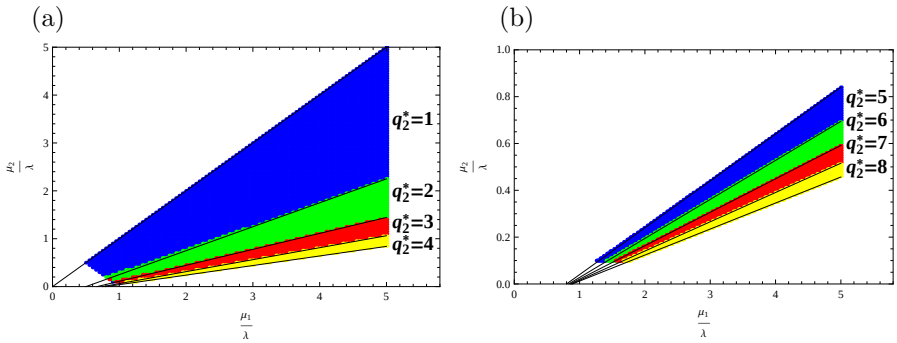


Fig. 1: The regions of optimality for $q_2 = i, i = \overline{1,8}$

From the table one can see that the coefficients of the line can be represented as a function of threshold estimation $\hat{q}_2$,

$$a_{21}(\hat{q}_2) = \hat{q}_2 + 1,\ a_{22}(\hat{q}_2) = -(\hat{q}_2 + 1)^2,\ a_{23}(\hat{q}_2) = -\hat{q}_2.$$

Conjecture 1 follows from solution of the quadratic equation

$$(\hat{q}_2 + 1)r_1 - (\hat{q}_2 + 1)^2 r_2 - \hat{q}_2 = r_1(\hat{q}_2)^2 - (r_1 - 2r_2 - 1)\hat{q}_2 + r_2 - r_1 = 0,$$

with two roots

$$(\hat{q}_2)_{1,2} = \frac{r_1 - 1 \pm \sqrt{(r_1 - 1)^2 + 4r_2}}{2r_2} - 1.$$

Eliminating the negative root due to condition $\hat{q}_2 \geq 1$ and taking into account the notations $r_j, j = \overline{1,K}$ for the integer-valued level q_2 we get an estimation in the form

$$\hat{q}_2 = \left\lfloor \frac{\mu_2 - \lambda + \sqrt{(\mu_2 - \lambda)^2 + 4\mu_2\lambda}}{2\mu_2} - 1 \right\rfloor.$$

To get the formulas for the model with costs we assume that it should have the same structure as (19), where $\lambda = 0$. Hence we get the relation (19) for $\lambda > 0$. In the same way one can derive the formulas for the case $K = 3$. For three server model the regions of optimality for the levels q_2 and q_3 are the planes as illustrated in figures 2. These planes are defined as

$$a_{21}r_1 + a_{22}r_2 + a_{23} = 0,\ a_{31}r_1 + a_{32}r_2 + a_{33}r_3 + a_{34} = 0,$$

where the coefficients are expressed through the estimations $\hat{q}_2$ and $\hat{q}_3$,

$$a_{21}(\hat{q}_2) = \hat{q}_2 + 1,\ a_{22}(\hat{q}_2) = -(\hat{q}_2 + 1)^2,\ a_{23}(\hat{q}_2) = -\hat{q}_2,$$
$$a_{31}(\hat{q}_3) = a_{32}(\hat{q}_3) = \hat{q}_3 + 2,\ a_{33}(\hat{q}_3) = -(\hat{q}_3 + 2)^2,\ a_{34}(\hat{q}_3) = -\hat{q}_3.$$

(a) (b)

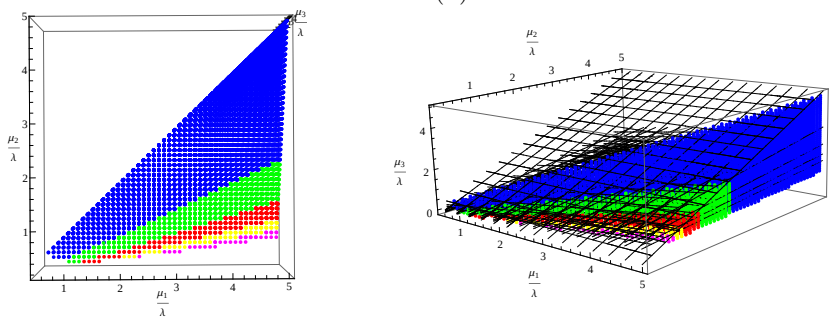


Fig. 2: The regions of optimality for $q_k = i, i = \overline{q_{k-1}, 5}, k = \overline{2, 3}$

As before, the solution of quadratic equations leads to the formulas corresponding to the problem of the mean number of customers minimization. These formulas can be rewritten for the scheduling model ($\lambda = 0$) with costs as in (19). The form of the recently derived formulas for the queues with $K = 2$ and $K = 3$ reveals a possible general expression for the estimations of threshold levels with arbitrary K, as was represented in Conjecture 1.

6. Numerical examples

Example 1. Consider the queueing system $M/M/5$ arrival intensity $\lambda = 0.9$. Other parameters take the following values:

j	0	1	2	3	4	5
c_j	1.50	3.00	2.80	2.60	2.40	2.00
μ_j	-	2.00	0.70	0.50	0.40	0.30
$c_j\mu_j^{-1}$	-	1.50	4.00	5.20	6.00	6.66

In the following table we list the results of calculations for the levels q_k^* and $\hat{q}_k^*, k = \overline{2, K}$ as well as the results for g in case of different control policies and arrival intensity λ

λ	g^{OTP}	$(q_2^*, q_3^*, q_4^*, q_5^*)$	g^{HTP}	$(\hat{q}_2^*, \hat{q}_3^*, \hat{q}_4^*, \hat{q}_5^*)$	g^{FFS}	g^{RSS}	g^{HS}
0.1	0.154	(4,6,7,9)	0.154	(4,6,7,9)	0.162	0.459	0.328
0.9	1.812	(3,4,5,6)	1.812	(3,4,6,7)	2.197	3.707	2.957
1.5	3.687	(2,3,4,5)	3.690	(2,3,5,6)	4.379	5.832	4.972
2.5	8.294	(2,2,3,3)	8.296	(2,2,3,4)	8.917	15.502	8.982
3.5	19.739	(2,2,2,3)	19.857	(2,2,3,3)	19.972	116.274	21.419
3.8	54.689	(2,2,2,2)	54.689	(2,2,2,2)	54.707	586.186	65.852

Here OTP means *Optimal Threshold Policy*, HTP – *Heuristic Threshold Policy*, FFS – *Fastest Free Server* policy, RSS – *Random Server Selection* policy and HS – *Homogeneous System*, where the total service rate $K\mu$ of the homogeneous system is equal to the total service rate $\sum_{j=1}^{K}\mu_j$ of the heterogeneous system.

Example 2. Consider the queueing system $M/M/5$ from the previous example. Let $c_j = 1, j = \overline{0,K}$. The next table includes the results of calculations for the levels q_k^* and $\hat{q}_k^*, k = \overline{2,K}$, as well as the values of the mean number of customers in the system $\bar{N}$ for different control policies and λ.

λ	$\bar{N}^{OTP}$	$(q_2^*, q_3^*, q_4^*, q_5^*)$	$\bar{N}^{HTP}$	$(\hat{q}_2^*, \hat{q}_3^*, \hat{q}_4^*, \hat{q}_5^*)$	$\bar{N}^{FFS}$	$\bar{N}^{RSS}$	$\bar{N}^{HS}$
0.1	0.053	(2,4,5,8)	0.053	(2,4,5,8)	0.055	0.191	0.128
0.9	0.692	(2,3,4,6)	0.692	(2,3,4,7)	0.788	1.509	1.156
1.5	1.459	(2,2,3,5)	1.463	(2,2,4,6)	1.649	2.222	1.955
2.5	3.454	(1,2,2,3)	3.510	(2,2,3,4)	3.674	4.506	3.723
3.5	11.181	(1,1,2,3)	11.185	(2,2,2,3)	11.219	28.944	11.109
3.8	29.815	(1,2,2,2)	29.815	(2,2,2,3)	33.029	144.798	40.458

The results show that the difference in performance between the OTP and HTP does not exceed 1.5% and these policies can be more than 25% superior in performance comparing to the mostly used in practice allocation policy FFS.

7. Conclusion

In this paper we have obtained heuristic relations for the optimal threshold levels which minimize the long-run average cost per unit of time or, in particular, the mean number of customers in the queue system with heterogeneous servers. These formulas give satisfactory values for the levels $q_k, k = \overline{2,K}$, with a relative error which was less than 1.5% in all realized examples. Also it was proved that this policy is superior in performance comparing to some alternative allocation policies. Therefore this policy can be treated at least as a quasi-optimal one. The formulas were verified for different number of servers with quite equal relative errors so we may expect the validity of these expressions to an arbitrary K. These approach was successfully applied to the retrial heterogeneous queueing systems as well, the results will be published shortly.

REFERENCES

1. Aviv, Y., Federgruen, A.: The Value-Iteration Method for Countable State Markov Decision Processes. Operastions Research Letters 24(5), 223–234 (1999).
2. Efrosinin, D.: Controlled queueing systems with heterogeneous servers. Dynamic optimization and monotonicity properties. Saarbrücken: VDM Verlag (2008).

3. Efrosinin, D.: Queueing model of a hybrid channel with faster link subject to partial and complete failures. Annals of Operations Research 202(1), 75–102 (2013).
4. Efrosinin, D., Rykov, V.: On performance characteristics for queueing systems with heterogeneous servers. Automation and Remote Control 69(1), 61–75 (2008).
5. Howard, R.: Dynamic Programming and Markov Processes. Wiley Series, New York (1960).
6. Kumar, B.K., Arivudainambi, D.: Transient solution of an $M/M/c$ queue with heterogeneous servers and balking. Information and Management Sciences 12(3), 15–27 (2001).
7. Puterman, M. L. Markov Decision Process. Wiley series in Probability and Mathematical Statistics (1994).
8. Rykov, V., Efrosinin, D.: Optimal control of queueing systems with heterogeneous servers. Queueing Systems 46, 389–407 (2004).
9. Rykov, V., Efrosinin, D.: On the slow server problem. Automation and Remote Control 70(12), 2013–2013 (2009).
10. Sennott, L.I.: Stochastic Dynamic Programming and the Control of Queueing Systems. John Wihley&Sons, New-York (1999).
11. Tijms, H.C.: Stochastic Models. An Algorithmic Approach. John Wiley&Sons, New-York (1994).

ON POLYNOMIAL CONVERGENCE RATE OF THE AVAILABILITY FACTOR TO ITS STATIONARY VALUE

A. Veretennikov[1], *G. Zverkina*[2]

[1] University of Leeds, Leeds, UK, & National Research University Higher School of Economics, & Institute for Information Transmission Problems, Moscow, Russia

[2] Moscow State University of Railway Engineering, Moscow, Russia

a.veretennikov@maths.leeds.ac.uk, zverkina@gmail.com

Abstract

We establish a computable estimate of the readiness coefficient of a standard binary-state system in the case where working time and repair time distributions are both with heavy tails.

Keywords: readiness coefficient, restorable system, polynomial convergence rate

1. Introduction

Let us consider a restorable system, which may be either in the working state during a random time ξ with a distribution function $F_1(s) \stackrel{\text{def}}{=} \mathbf{P}\{\xi \leqslant s\}$, or it may be broken down and being restored by some service during another random time η with a distribution function $F_2(s) \stackrel{\text{def}}{=} \mathbf{P}\{\eta \leqslant s\}$. All periods of working and repairing are alternate a independent. The readiness coefficient $A(t)$ is defined as the probability that at time t the system is in the working (serviceable) state.

Often in the literature it is accepted that at initial time $t = 0$ the system is serviceable and that it is in the beginning of its working period. We consider a more general case assuming that the activity of the system may have started earlier so that at $t = 0$ the system can be in one of the two states: perfect functionality or complete failure; and further that *before* $t = 0$ the system already spent time x in its current state.

Let us formalize the definition of readiness coefficient. We assume that (ξ_i) are random variables with a common distribution function $F_1(x) = \mathbf{P}\{\xi_i \leqslant x\}$; likewise, (η_i) are random variables with a (another) common distribution function $F_2(x) = \mathbf{P}\{\eta_i \leqslant x\}$; all of them are mutually independent.

If at time $t = 0$ our system is working and its elapsed working time before $t = 0$ equals x, then the *residual time* of this working period is a random variable denoted by $\xi^{(x)}$; its distribution function is denoted by $F_1^{(x)}(s) \stackrel{\text{def}}{=} \mathbf{P}\{\xi^{(x)} \leqslant s\} = \mathbf{P}\{\xi \leqslant s + x | \xi > x\} = 1 - \dfrac{F_1(x+s)}{1-F_1(x)}$. Correspondingly, if at the time $t = 0$ the system is under repair and the duration of this repair before $t = 0$ equals x, then the residual time of this repair period is a random variable denoted by $\eta^{(x)}$ with a distribution function $F_2^{(x)}(s) \stackrel{\text{def}}{=} \mathbf{P}\{\eta^{(x)} \leqslant s\} = \mathbf{P}\{\eta \leqslant s + x | \eta > x\} = 1 - \dfrac{F_2(x+s)}{1-F_2(x)}$.

In the first case we will use notations $t_0 \stackrel{\text{def}}{=} 0$, $t_1 \stackrel{\text{def}}{=} \xi^{(x)} + \eta_1$, $t_i \stackrel{\text{def}}{=} \xi^{(x)} + \eta_1 + \sum_{j=2}^{i}(\xi_j + \eta_j)$, and $t'_0 \stackrel{\text{def}}{=} \xi^{(x)}$, $t'_i \stackrel{\text{def}}{=} \xi^{(x)} + \sum_{j=1}^{i-1}(\eta_j + \xi_{j+1})$. In the second case $t_1 \stackrel{\text{def}}{=} \eta^{(x)}$, $t_i \stackrel{\text{def}}{=} \eta^{(x)} + \sum_{j=2}^{i}(\xi_{j-1} + \eta_j)$, $t'_1 \stackrel{\text{def}}{=} 0$, $t'_1 \stackrel{\text{def}}{=} \eta^{(x)} + \xi_1$, $t'_i \stackrel{\text{def}}{=} \eta^{(x)} + \xi_1 + \sum_{j=2}^{i}(\eta_j + \xi_j)$. In this notation $A(t) \stackrel{\text{def}}{=} \mathbf{P}\left\{t \in \bigcup_i [t_i, t'_i)\right\}$.

It is well known that if distributions of $\xi+\eta$ are non-arithmetical and $\mathbf{E}\,\xi+\mathbf{E}\,\eta < \infty$, then there exists a limiting value $\lim\limits_{t\to\infty} A(t) \stackrel{\text{def}}{=} A = \dfrac{\mathbf{E}\,\xi}{\mathbf{E}\,\xi + \mathbf{E}\,\eta}$. Moreover, if for some $n > 1$, $\mathbf{E}\,\xi^n + \mathbf{E}\,\eta^n < \infty$, then $\limsup\limits_{t\to\infty} |A(t) - A|t^{n-1} < \infty$ (see e.g. [1, Theorem 3, Appendix 1] or [2, Theorem 10.7.4]). In other words, for any $\alpha \in (0, n-1]$ there exists a constant $C(\alpha)$ such that $\forall\ t > 0\ |A(t) - A| \leqslant C(\alpha)(1 + t)^{-\alpha}$. However the general theory does not provide neither the value of $C(\alpha)$, nor any bound for it.

Any knowledge of the value $C(\alpha)$ or its bound is rather important in applications. The goal of this paper is to provide explicit sufficient conditions for $C(\alpha)$ to be finite and to give explicit estimates to this constant.

2. Assumptions and notations

2.1. Assumption. We suppose that for some $\Lambda > K_1 > 3$, $K_2 > 3$,

$$F_1(x) = 1 - e^{-\int_0^x \lambda(s)\,\mathrm{d}\,x} \quad \left(\text{i.e., almost everywhere } \lambda(s) = \frac{F'_1(s)}{1 - F_1(s)}\right),$$

$$\text{and } \Lambda \geqslant \lambda(s) \geqslant \frac{K_1}{1+s} \text{ when } s > 0; \tag{1}$$

$$F_2(s) \geqslant 1 - \frac{1}{(1+s)^{K_2}} \text{ when } s > 0;\ \textit{we do not assume continuity of } F_2(s). \tag{2}$$

Note that from (1) it follows that $F_1(s) \geqslant 1 - \dfrac{1}{(1+s)^{K_1}}$. So, (1) and (2) imply that for all $a \in (0, K_1 - 1)$ and $b \in (k_2 - 1)$ we have, $\dfrac{K_1}{\Lambda} < \mathbf{E}\,\xi^a < \dfrac{a}{K_1 - a - 1} < \infty$ and $\mathbf{E}\,\eta^b < \dfrac{b}{K_2 - b - 1} < \infty$, which suffices for the existence of A.

Notice that $\lambda(s)$ is called *intensity* of failure of the recoverable system, of course, where it is working.

2.2. Notations.

1. Denote $K \stackrel{\text{def}}{=} \min(K_1, K_2)$.

2. The behaviour of the system under consideration may be presented by the random process

$$X_t = (n_t, x_t) = \begin{cases} (1, t - t_i), & \text{if } t \in [t_i, t'_i); \\ (2, t - t'_i), & \text{if } t \in [t'_i, t_{i+1}); \end{cases} \quad n(X_t) \stackrel{\text{def}}{=} n_t,\ x(X_t) \stackrel{\text{def}}{=} x_t.$$

The state space of the process X_t is a set $\mathcal{X} \stackrel{\text{def}}{=} \{\{0,1\} \times \mathbb{R}_+\}$ with a standard σ-algebra. Let $X_0 = (n_0, x_0)$. Denote $\mathcal{S}_j \stackrel{\text{def}}{=} \{(j,x),\ x \in \mathbb{R}_+\} \subset \mathcal{X} \quad (j = \overline{1,2})$.

3. Denote (here $j = \overline{1,2}$):

$$\mathrm{M}_j(k) \stackrel{\text{def}}{=} k \int\limits_0^\infty \frac{s^{k-1}}{(1+s)^{K_j}}\,\mathrm{d}\,s; \quad \mathrm{M}_j^{(x)}(k) \stackrel{\text{def}}{=} \frac{k}{1-F_j(x)} \int\limits_0^\infty \frac{s^{k-1}}{(1+s+x)^{K_j}}\,\mathrm{d}\,s;$$

$$\varkappa(T) \stackrel{\text{def}}{=} \int\limits_0^\infty \frac{K_1 e^{-\Lambda s}}{1+T+s}\,\mathrm{d}\,s; \qquad F_j^{(a)}(s) \stackrel{\text{def}}{=} 1 - \frac{1-F_j(s+a)}{1-F_j(a)}.$$

4. Let us choose

$$R > \Theta_0 \stackrel{\text{def}}{=} \frac{\mathbf{E}(\xi+\eta)^2}{2(\mathbf{E}\,\xi + \mathbf{E}\,\eta)} \left[\leqslant \frac{8\Lambda}{K^2 - 3K}\right], \tag{3}$$

let N be such that $e^{-\Lambda R} > \dfrac{1}{(1+NR)^{K_1}}$, and let

$$q \stackrel{\text{def}}{=} 1 - \left(1 - \frac{\Theta_0}{R}\right)\left(e^{-\Lambda R} - \frac{1}{(1+NR)^{K_1}}\right)\varkappa(NR);$$

$$\Phi(\alpha, X_0) \stackrel{\text{def}}{=} \left(\sum_{i=0}^{\infty}(i+4)^\alpha q^i\right)\Bigg(1 + \mathbf{1}(n_0 = 0)2^{\alpha-1}\left(\mathrm{M}_1^{(x_0)}(\alpha) + \mathrm{M}_2(\alpha)\right) +$$

$$+\mathbf{1}(n_0 = 1)\mathrm{M}_2^{(x_0)}(\alpha) + 2^{\alpha-1}A\left(\frac{\alpha}{(K_1 - \alpha - 1)\mathbf{E}\,\xi} + \mathrm{M}_2(\alpha)\right) +$$

$$+ \frac{(1-A)\alpha}{(K_2 - \alpha - 1)\mathbf{E}\,\eta} + \mathrm{M}_1(\alpha) + \mathrm{M}_2(\alpha)\Bigg).$$

3. Main result

Theorem 1. *Let $K > 3$ and let the conditions (1), (2) be satisfied. Then for the process described earlier with initial state $X_0 = (n_0, x_0)$, for every $\alpha \in (1, K-1)$ there exists a constant $C(\alpha, X_0) \leqslant \Phi(\alpha, X_0)$ such that for all $t \geqslant 0$ the following inequality is true:*

$$|A(t) - A| \leqslant \frac{C(\alpha, X_0)}{(1+t)^\alpha}.$$

In this short paper we only suggest an extended scheme of proof. A complete proof of the Theorem 1 will be given in a subsequent publication.

4. Scheme of proof

4.1. ***Properties of the process X_t.*** The process X_t defined in the Subsection 2.2 (point **2.**) is Markov. Moreover, it possesses a strong Markov property. We skip the standard proof of both claims.

Note that trajectories of the process X_t are right continuous.

4.2. ***On the stationary distribution of*** X_t**.** In terms of [3], [4], the process X_t is a linear-type (piecewise linear) Markov process, and it satisfies the conditions of ergodic theorem from [5, §2.6] (see also [6, Theorem 1]): there exists a stationary distribution $\mathcal{P}$ on $\mathcal{X}$ such that there is a limit $\lim\limits_{t\to\infty} \mathbf{P}\{n_t = j, x_t \leqslant s\} = \mathcal{P}(\{j\} \times [0, s])$ for any initial state X_0 (again and always in the sequel $j = \overline{1,2}$);

$$\mathcal{P}(\{j\} \times (s, \infty)) = \frac{\mathbf{1}\{j = 1\} \int\limits_s^\infty (1 - F_1(s))\,\mathrm{d}\,s + \mathbf{1}\{j = 2\} \int\limits_s^\infty (1 - F_2(s))\,\mathrm{d}\,s}{\mathbf{E}\,\xi + \mathbf{E}\,\eta},$$

and $\mathcal{P}(n_t = 1) = \dfrac{\mathbf{E}\,\xi}{\mathbf{E}\,\xi + \mathbf{E}\,\eta} = A.$

4.3. ***Coupling method.*** To prove the Theorem 1 we will use the *coupling method*, which will be now briefly recalled (for details see [7]).

Suppose some strong Markov process X_t weakly converges to its (unique) stationary regime; denote its marginal distribution by $\mathcal{P}$.

Suppose that on some probability space it is possible to construct two *independent* versions X_t' and X_t'' of this Markov process – i.e., both with the same generator but possibly with different initial distributions – such that the stopping time $\tau(X_0', X_0'') \stackrel{\text{def}}{=} \inf\{t > 0 : \ X_t' = X_t''\}$ has a finite expectation. If, further, we have an estimate $\mathbf{E}\,\psi(\tau(X_0', X_0'')) \leqslant C(X_0', X_0'')$ where $\psi(s) \uparrow$ and $\psi(s) > 0$ as $s > 0$, then we can use a strong Markov property and *coupling inequality*: $\forall\, \mathcal{D} \in \mathcal{B}(\mathcal{X})$

$$\left|\mathbf{P}\{X_t' \in \mathcal{D}\} - \mathbf{P}\{X_t'' \in \mathcal{D}\}\right| \leqslant \mathbf{P}\left\{t \leqslant \tau(X_0', X_0'')\right\} = \mathbf{P}\left\{\psi(t) \leqslant \psi(\tau(X_0', X_0''))\right\}.$$

Hence, due to Markov's inequality,

$$\left|\mathbf{P}\{X_t' \in \mathcal{D}\} - \mathbf{P}\{X_t'' \in \mathcal{D}\}\right| \leqslant \frac{\mathbf{E}\psi(\tau(X_0', X_0''))}{\psi(t)} \leqslant \frac{C(X_0', X_0'')}{\psi(t)}. \tag{4}$$

Once the inequality (4) is estalished for the pair of processes, we may conclude that for the stationary process $\widetilde{X}_t$ with the initial distribution $\mathcal{P}$ and for the process X_t starting from an arbitrary initial state X_0 we get,

$$\left|\mathbf{P}\{X_t \in \mathcal{D}\} - \mathbf{P}\{\widetilde{X}_t \in \mathcal{D}\}\right| = |\mathbf{P}\{X_t \in \mathcal{D}\} - \mathcal{P}(\mathcal{D})\}| \leqslant$$
$$\leqslant \frac{\int\limits_{\mathcal{X}} C(X_0, Y)\mathcal{P}(\,\mathrm{d}\,Y)}{\psi(t)} = \frac{\widetilde{C}(X_0)}{\psi(t)}. \tag{5}$$

Note that since the right hand side here does not depend on $\mathcal{D}$, this inequality, of course, provides an estimate in total variation, that is,

$$\sup_{\mathcal{D}\in\mathcal{X}} |\mathbf{P}\{X_t \in \mathcal{D}\} - \mathcal{P}(\mathcal{D})\}| \leqslant \frac{\widetilde{C}(X_0)}{\psi(t)}.$$

Also, if $A = \{n(X_t) = 0\}$, then $\mathbf{P}\{X_t \in A\} = A(t)$. Hence, in particular, the inequality (5) implies that $|A(t) - A| \leqslant \dfrac{\widetilde{C}(X_0)}{\psi(t)}$.

Now, the goal is to give an estimate of $\widetilde{C}(X_0)$ for the function $\psi(t) = (1+t)^\alpha$.

4.4. ***Coupling, continued.*** We will be using a procedure first suggested in [8]. On some probability space we construct a "paired" Markov process $Z_t = (Z'_t, Z''_t)$ in the state space $\mathcal{X}\times\mathcal{X}$ so that the marginal distributions of the processes Z'_t and Z''_t coincide with the distributions of the processes X'_t and X''_t, respectively:

$$(Z'_t,\, t \geqslant 0) \stackrel{\mathcal{D}}{=} (X'_t,\, t \geqslant 0) \quad \text{and} \quad (Z''_t,\, t \geqslant 0) \stackrel{\mathcal{D}}{=} (X''_t,\, t \geqslant 0)\,; \tag{6}$$

$Z'_0 = X'_0$ and $Z''_0 = X''_0$.

In addition, if at some moment $\widetilde{\tau}$ the random variable Z'_t coincides with Z''_t, i.e. $Z'_{\widetilde{\tau}} = Z''_{\widetilde{\tau}}$, then for all $t \geqslant \widetilde{\tau}$, $Z'_t = Z''_t$. This pair (Z', Z'') is called coupling. Of course, in general, the processes Z'_t and Z''_t will be dependent.

Assuming that the process $Z_t = (Z'_t, Z''_t)$ is already constructed, let us denote $\widetilde{\tau}(X'_0, X''_0)[= \widetilde{\tau}(Z'_0, Z''_0)] \stackrel{\text{def}}{=} \inf\{t > 0 : \ Z'_t = Z''_t\}$. The coupling is called *successful* if $\mathbf{P}\{\widetilde{\tau}(X'_0, X''_0) < \infty\} = 1$. Our coupling constructed below will be successful.

Then, we can use the coupling inequality (4) for the processes Z'_t and Z'_t:

$$|\mathbf{P}\{Z'_t \in \mathcal{D}\} - \mathbf{P}\{Z''_t \in \mathcal{D}\}| \leqslant \mathbf{P}\{t < \widetilde{\tau}(X'_0, X''_0)\}.$$

Due to (6) the same inequality holds true for X'_t and X''_t.

4.5. ***About the process*** Z_t**.** In order to construct a successful coupling, we will use the idea of the "Lemma about three random variables" (see [9]).

The construction of Z_t is based on a sequence of stopping times t_k, at which $\mathbf{1}\left\{n(Z'_{t-0}) \neq n\left(Z'_{t+0}\right)\right\} + \mathbf{1}\left\{n\left(Z''_{t-0}\right) \neq n\left(Z''_{t+0}\right)\right\} > 0$, i.e., of (random) times t_k where one of the processes Z'_t and Z''_t – or both of them – changes its first component.

Let $t_0 = 0$ and denote $m'_t \stackrel{\text{def}}{=} n(Z'_t)$, $m''_t \stackrel{\text{def}}{=} n(Z''_t)$, $z'_t \stackrel{\text{def}}{=} x(Z'_t)$, $z''_t \stackrel{\text{def}}{=} x(Z''_t)$. The sequence (t_k) will be built by induction. Assume that t_k is already determined for some k and consider three cases.

4.5.1. Suppose that $Z'_{t_k} \neq Z''_{t_k}$ and $m'_{t_k} + m''_{t_k} > 2$ (that is, at least one of the processes is in the set $\mathcal{S}_2$). Then on a special probability space $(\Omega, \mathcal{F}, \mathbf{P})$[1] we define independent random variables θ'_k and θ''_k with distribution functions $F^{(z'_{t_k})}_{m'_{t_k}}(s)$ and $F^{(z''_{t_k})}_{m''_{t_k}}(s)$ respectively: they are residual times of stay of the processes Z'_t and Z''_t in the sets $\mathcal{S}_{m'_{t_k}}$ and $\mathcal{S}_{m''_{t_k}}$. Denote $\theta_k \stackrel{\text{def}}{=} \min(\theta'_k, \theta''_k)$ and $t_{k+1} \stackrel{\text{def}}{=} t_k + \theta_k$.

For $t \in [t_k, t_{k+1})$ define,

$$\begin{aligned}
&Z'_t \stackrel{\text{def}}{=} (m'_{t_k}, z'_{t_k} + t - t_k); && Z''_t \stackrel{\text{def}}{=} (m''_{t_k}, z''_{t_k} + t - t_k);\\
&Z'_{t_{k+1}} \stackrel{\text{def}}{=} 1\{\theta'_k = \theta_k\}(m'_{t_k} - (-1)^{m'_{t_k}}, 0) + 1\{\theta'_k \neq \theta_k\}(m'_{t_k}, z'_{t_k} + t_{k+1} - t_k);\\
&Z''_{t_{k+1}} \stackrel{\text{def}}{=} 1\{\theta''_k = \theta_k\}(m''_{t_k} - (-1)^{m''_{t_k}}, 0) + 1\{\theta''_k \neq \theta_k\}(m''_{t_k}, z''_{t_k} + t_{k+1} - t_k).
\end{aligned} \tag{7}$$

[1] Here we skip the details of this probability space $(\Omega, \mathcal{F}, \mathbf{P})$ and the procedures of constructing random variables on it. A more complete presentation will be provided in further publications.

4.5.2. Suppose now that $Z'_{t_k} \neq Z''_{t_k}$ and $m'_{t_k} = m''_{t_k} = 1$. In this case, using the idea of the "Lemma about three random variables" (see [9]) we construct on the space $(\Omega, \mathcal{F}, \mathbf{P})$ the pair of *dependent* random variables (θ'_k, θ''_k) such that:

$$\mathbf{P}\{\theta'_k \leqslant s\} = F_1^{(z'_{t_k})}(s), \quad \mathbf{P}\{\theta''_k \leqslant s\} = F_1^{(z''_{t_k})}(s);$$

$$\mathbf{P}\{\theta'_k = \theta''_k\} = \int_0^\infty \min\left(\left(F_1^{(z'_{t_k})}(s)\right)', \left(F_1^{(z''_{t_k})}(s)\right)'\right) \mathrm{d}\, s \geqslant \int_0^\infty \frac{K_1 e^{-\Lambda s}}{1 + s + \max(z'_{t_k}, z''_{t_k})} \mathrm{d}\, s =$$

$$= \varkappa(\max(z'_{t_k}, z''_{t_k})). \quad (8)$$

Note that clearly $\varkappa(T) \downarrow 0$ if $T \uparrow +\infty$.

Next, we again denote $t_{k+1} \stackrel{\text{def}}{=} t_k + \min(\theta'_k, \theta''_k)$ and apply the same construction given in the formulae (7). This definition and (8) imply the inequality $\mathbf{P}\{Z'_{t_{k+1}} = Z'_{t_{k+1}}\} \geqslant \varkappa(\max(z'_{t_k}, z''_{t_k}))$.

4.5.3. Now, suppose $Z'_{t_k} = Z''_{t_k}$. In this case we construct random variables $\theta'_k = \theta''_k$ (i.e., they are identical) with distribution function $F_{m'_{t_k}}^{(z'_{t_k})}(s)$ on the space $(\Omega, \mathcal{F}, \mathbf{P})$, and $t_{k+1} \stackrel{\text{def}}{=} t_k + \theta'_k$. Here for $t \in [t_k, t_{k+1})$,

$$Z'_t = Z''_t = (m'_{t_k}, z'_{t_k} + t - t_k);\ Z'_{t_{k+1}} = Z''_{t_{k+1}} = (m'_{t_k} - (-1)^{m'_{t_k}}, 0).$$

This construction (4.5.1–4.5.3) gives us the desired pair $Z_t = (Z'_t, Z''_t)$, which satisfies (6) and which is suitable for the successful coupling procedure; we skip the proof of this fact.

4.6. ***Using coupling method.*** Let us fix two initial values $X'_0 \equiv Z'_0 \neq Z''_0 \equiv X''_0$. In this step of the proof we will show the coupling inequality for the process $Z = (Z', Z'')$; hence, the same inequality will be established for the couple (X', X'').

4.6.1. For $t > 0$ denote,

$$\tau'(t) \stackrel{\text{def}}{=} \left(\inf\{s > t :\ Z'_t = (0,0)\}\right), \quad \tau''(t) \stackrel{\text{def}}{=} \left(\inf\{s > t :\ Z''_t = (0,0)\}\right).$$

These are the moments of the beginning of regeneration periods for the processes Z' and Z'' after the nonrandom t. Denote also $\tau(Z'_0, Z''_0) \stackrel{\text{def}}{=} \max(\tau'(0), \tau''(0))$.

At $\tau'(0)$ the regeneration period of the process Z'_t begins. Its length equals $\theta \stackrel{\mathcal{D}}{=} \xi + \eta$ where ξ and η were introduced in the Section 1. After that, the behaviour of Z' does not depend on the initial state Z'_0 (given $\tau'(0)$). The same can be said about the process Z''_t.

Let $t > \tau\left(Z'_0, Z''_0\right)$. Then, there was at least one beginning of the regeneration period of each of the processes Z' and Z'' before t.

Denote $\vartheta'(t) \stackrel{\text{def}}{=} (\tau'(t) - t)$ – the residual time of the last regeneration period of Z'', which started before time t. From the corollary of W. Smith's Key Renewal Theorem (cf. [6, Theorem 2]), the following inequality holds true: if $t > \tau\left(Z'_0, Z''_0\right)$, then

$$\mathbf{E}\left(\vartheta'_t \middle| t > \tau\left(Z'_0, Z''_0\right)\right) \leqslant \frac{\mathbf{E}\,\theta^2}{2\mathbf{E}\,\theta} = \frac{\mathbf{E}\,(\xi + \eta)^2}{2(\mathbf{E}\,\xi + \mathbf{E}\,\eta)}\,[= \Theta_0]\,. \quad (9)$$

The same statement applies to the process Z''.

Note that $\tau \leqslant \tau'(0) + \tau''(0)$, and, by virtue of Jensen's inequality, for all $\alpha \in (1, K-1)$

$$\mathbf{E}\,(\tau_0)^\alpha \leqslant 2^{\alpha-1}\left(\mathbf{E}\,(\tau'(0))^\alpha + \mathbf{E}\,(\tau''(0))^\alpha\right).$$

4.6.2. Without loss of generality we can assume that $\tau = \tau''(0)$. Let $\tau_1 \stackrel{\text{def}}{=} \tau''(0)$, $\tau_{k+1} \stackrel{\text{def}}{=} \min\{\tau''(t),\, t > \tau_{k+1}\}$; $\{\tau_k\}$ is a sequence of beginnings of regeneration periods of Z''.

Denote $\mathcal{E}_k \stackrel{\text{def}}{=} \{\vartheta'(\tau_k) < R \;\&\; (\tau_{k+1} - \tau_k) \in (R, NR)\}$, i.e. at time $\widehat{\tau}_k \stackrel{\text{def}}{=} \tau_k + \vartheta'(\tau_k)$ $Z'_{\widehat{\tau}_k} = (1, 0)$, $Z''_{\widehat{\tau}_k} = (1, z)$, and $z < NR$.

Using (9) and condition (1) by Markov inequality we can estimate $\mathbf{P}\{\mathcal{E}_k\}$:

$$\mathbf{P}\{\mathcal{E}_k\} \geqslant \left(1 - \frac{\Theta_0}{R}\right)\left(e^{-\Lambda R} - \frac{1}{(1+NR)^{K_1}}\right) \stackrel{\text{def}}{=} \pi(R, N). \tag{10}$$

Now, using (8), we have: $\mathbf{P}\{Z'_{\tau_{k+1}} = Z''_{\tau_{k+1}}\} \geqslant \pi(R,N)\varkappa(RN) \stackrel{\text{def}}{=} p$.

4.7. ***Completion of the proof.*** The number of regeneration periods of Z''_t before the processes Z'_t and Z''_t meet each other according to the scheme from the step **4.6.** (that is, any meeting outside this scheme is ignored) is a random variable ν dominated by another one with a geometric distribution with parameter p (ν itself has a more complicated distribution). Denote $q \stackrel{\text{def}}{=} 1 - p$, and $\varsigma\left(Z'_0, Z''_0\right) \stackrel{\text{def}}{=} \inf\{t > 0 :\, Z'_t = Z''_t\}$. Obviously, $\varsigma\left(Z'_0, Z''_0\right) \leqslant \tau_\nu$.

Since we know the distribution of $\tau = \tau\left(Z'_0, Z''_0\right)$ and $\theta \stackrel{\mathcal{D}}{=} \xi + \eta$, we can obtain an estimation of $\mathbf{E}\left(1 + \varsigma\left(Z'_0, Z''_0\right)\right)^\alpha$ for all $\alpha \in (1, K-1)$: by Jensen's inequality we get,

$$\begin{aligned}
\mathbf{E}\left(1 + \varsigma\left(Z'_0, Z''_0\right)\right)^\alpha &\leqslant \mathbf{E}\left(1 + \tau\left(Z'_0, Z''_0\right) + \xi + \sum_{i=1}^{\infty}\left(\mathbf{P}\{\nu = i\}\sum_{k=1}^{i}(\xi_k + \eta_k)\right)\right)^\alpha \leqslant \\
&\leqslant \sum_{i=1}^{\infty} q^{i-1}\mathbf{E}\left(1 + \tau'(0) + \tau''(0) + \xi + \sum_{k=1}^{i}(\xi_k + \eta_k)\right)^\alpha \leqslant \\
&\leqslant \sum_{i=1}^{\infty} q^{i-1}(i+4)^{\alpha-1}\left(1 + \mathbf{E}\left((\tau'(0))^\alpha + \mathbf{E}\,(\tau''(0))^\alpha + (i+1)\mathbf{E}\,\xi^\alpha + \mathbf{E}\,\eta^\alpha\right)\right) \leqslant \\
&\leqslant \sum_{i=1}^{\infty} q^{i-1}(i+4)^{\alpha-1}\left(1 + \mathbf{1}\left(n'_0 = 0\right)2^{\alpha-1}\left(\mathrm{M}_1^{(x'_0)}(\alpha) + \mathrm{M}_2(\alpha)\right) + \mathbf{1}\left(n'_0 = 1\right)\mathrm{M}_2^{(x'_0)}(\alpha) + \right. \\
&\quad + \mathbf{1}\left(n''_0 = 0\right)2^{\alpha-1}\left(\mathrm{M}_1^{(x''_0)}(\alpha) + \mathrm{M}_2(\alpha)\right) + \\
&\quad \left. + \mathbf{1}\left(n''_0 = 1\right)\mathrm{M}_2^{(x''_0)}(\alpha) + (i+1)\mathrm{M}_1(\alpha) + i\mathrm{M}_2(\alpha)\right) = \\
&= C\left(\alpha, Z'_0, Z''_0\right) = C\left(\alpha, X'_0, X''_0\right)
\end{aligned}$$

It is easy to see that

$$\int_X C\left(\alpha, X_0', X_0''\right)\mathcal{P}\left(\mathrm{d}\, X_0''\right) \leqslant \Phi\left(\alpha, X_0'\right), \tag{11}$$

which completes the proof of the theorem.

Remark. The estimate (11) could be improved; moreover, a more careful choice of parameters R and N may provide some enhancement of this bound.

Acknowledgments. The authors are grateful to V. V. Kozlov for very useful consultations. Both authors are supported by the RFBR, project No 14-01-00319 A. For the first author the article was prepared within the framework of a subsidy granted to the HSE by the Government of the Russian Federation for the implementation of the Global Competitiveness Program.

REFERENCES

1. Borovkov A. A. Stochastic processes in queueing theory. Springer-Verlag, 1976.
2. Thorisson H. Coupling, Stationarity, and Regeneration. Springer, 2000.
3. Gnedenko B. V., Kovalenko I. N. Introduction to queueing theory. Birkhauser, Boston, 1989.
4. Kalashnikov V. V. Some properties of piecewise linear Markov processe // Teor. Veroyatnost. i Primenen., 20:3 (1975), 571–583
5. Klimov G. P. Probability theory and mathematical statistics. Mir Publishers, Moscow, 1986.
6. Zverkina G. A. On some limit theorems following from Smith's Theorem (2015). arXiv:1509.06178 [math.PR], to appear in Proceedings of the the 13th International Conference "Kolmogorov Lectures", Jaroslavl, 2015.
7. Lindvall T. Lectures on the Coupling Method. Wiley, New York, 1992.
8. Griffeath D. A maximal coupling for Markov chains // Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete 1975, Volume 31, Issue 2, pp 95-106.
9. Veretennikov A. Coupling method for Markov chains under integral Doeblin type condition // Theory Stoch. Process. 8(24) (2002), no.3-4, 383-390.

ON A CLASS OF QUEUES WITH APPLICATION TO TELECOMMUNICATION[1]

A.Krishnamoorthy[1], *A.S.Manjunath*[2], *V.M.Vishnevsky*[3], *V.C. Narayanan*[4]

[1,2] Cochin University of Science and Technology, Kochi, India

[3] V.A.Trapeznikov Institute of Control Sciences of RAS, Moscow, Russia

[4] Department of Mathematics, Govt. Engineering College, Trissur, India

[1]achyuthacusat@gmail.com, [3]vishn@inbox.ru

Abstract

We consider a queueing system which turns out to be the dual of a queue with second optional service (see [1] and [2]). When a customer is selected for service, the server takes him for service that turns out be different from what is exactly required for him, with probability p and with complementary probability $1-p$, he is taken for the actual service required. A typical case is wrong diagnosis of the ailment of a patient. Once the service at the undesirable stage starts, the customer either completes the service there and goes to the one where he should have been taken for service at the very beginning, or leaves the system (gets absorbed) before completing service in the first part.

In either case a clock that starts ticking the moment the service (un desired one) starts, determines the event to follow: if the clock realizes first (still the undesired service goes on), the customer leaves the system immediately without going for the desired service. On the other hand if the service gets completed first, then he goes for the desired one immediately. The system has only one server to provide service. We derive the system state distribution. Then several system characteristics are analyzed. Next we extent above situation to a case where initially the customer is given service that is (are) not required with probabilities $p_1, \ldots, p_n$, respectively, such that their sum equals p; with complementary probability the customer is chosen for the right service.

Stochastic decomposition of the system state is established.

We give an application of the above described model in telecommunication on highways. An emergency message, at the time when it is ready, may get immediate access for service (with probability $1-p$). With complementary probability p, the server is found to be busy and so has to pass through contention windows in a sequential manner. In this process if the message does not get transmitted within the time frame (a maximum of 100 milliseconds) the message loses its significance. Our objective is to minimize p.

REFERENCES

1. Kailash C. Madan. An M/G/1 queue with second optional service // Queueing Systems, 2000, Vol. 34, Issue 1, pp 37 – 46
2. J.Medhi. A Single Server Poisson Queue with a Second Optional Channel // Queuing Systems, 2002, Vol. 42, Issue 3, pp 239 – 242

[1]The research was supported by the Ministry of Education and Science of the Russian Federation — applied research project №14.613.21.0020 of 22.10.2014 (RFMEFI61314X0020).

ON A RETRIAL QUEUEING MODEL WITH ORBITAL SEARCH OF CUSTOMERS — APPLICATION TO TELECOMMUNICATION ON HIGHWAYS[1]

A.Krishnamoorthy[1], *V.M.Vishnevsky*[2], *T.G.Deepak*[3], *V.C.Joshua*[4]

[1]Cochin University of Science and Technology, Kochi, India
[2]V.A.Trapeznikov Institute of Control Sciences of RAS, Moscow, Russia
[3]Indian Institute of Space Science & Technology, Thiruvananthapuram, India
[4]CMS College, Kottayam, India
[1]achyuthacusat@gmail.com, [2]vishn@inbox.ru, [3]deepak@iist.ac.in,
[4]vcjoshua@yahoo.co.uk

Abstract

Emergency messages originating on the highways need quick transmission, lest they loose significance. However, highways are not equipped with much Road side units (RSU) nor enough of on board units. Thus several emergency messages do not reach the destination in time which result in accidents and other untoward incidents.

In this paper we propose model that enables the smooth passage of maximum number of emergency packets before the expiry time. All messages generated within a 100 millisecond time slot should be either transmitted within that time frame or will loose significance. When an emergency packet originates, if the service facility is idle, the packet is immediately taken for transmission. Else it is sent to an orbit, exclusively for emergency packets. Once the server becomes free, a search mechanism is employed to take the first emergency packet in the orbit for transmission. This is continued, either until all emergency packets are exhausted or the end of the current time slot is arrived at, whichever occurs first. At the end of the time slot all emergency packets from the orbit are flushed out to accommodate new ones for the new time slot. We find a search mechanism that will ensure minimum loss of emergency packets in any time slot, and at the same time revenue of the system is maximized. It may be noted that emergency packets do not produce any revenue to the system and at the same time, the system is duty bound to transmit maximum number of emergency packets. The messages that provide information on facilities in the neighbourhood, entertainments etc. bring high revenue to the system. We look for a trade off between the two.

[1]The research was financially supported by the Ministry of Education and Science of the Russian Federation in the framework of the applied research project №14.613.21.0020 of 22.10.2014 (RFMEFI61314X0020).

SWARM OF PUBLIC UNMANNED AERIAL VEHICLES AS A QUEUING NETWORK

R. Kirichek[1]*, A. Paramonov*[2]*, A. Koucheryavy*[3]
The Bonch-Bruevich Saint-Petersburg State University of Telecommunications,
Saint-Petersburg, Russia
kirichek@sut.ru[1], alex-in-spb@yandex.ru[2], akouch@mail.ru[3]

Abstract

Unmanned aerial vehicles which are used to build flying ubiquitous sensor networks are viewed as a queuing system and their swarm — as a queuing network. It is proved that a sufficiently large number of UAVs swarm can be considered as a network of Jackson. The distribution of the lengths of the shortest paths for the UAVs swarms with a cube and a sphere is determined.

Keywords: public flying ubiquitous sensor network, unmanned aerial vehicle, the queuing system, the queuing network, the length of the shortest path

1. Introduction

One of the most attractive areas of the networks and communication systems has recently been Flying Ad Hoc Networks (FANET) [1, 2, 3]. Initially used mainly for military purposes, UAVs are currently used in civilian applications [4, 5]. By analogy with the division of terrestrial in the Ad Hoc network [6, 7] and ubiquitous or wireless sensor networks [8.9] in the field of Ad Hoc networks there were flying ubiquitous sensor network FUSN [10]. Widespread public unmanned aerial vehicles and related networking features FUSN enable to identify a new class of public communications networks FUSN-P (Public) [11]. One of the main features of FUSN-P is that the UAV is operated usually by nonprofessional users, so that it requires the simplest handling of them during operation. For this purpose, in [11] in the FUSN-P it was proposed to use the UAV flight for the data collection from the sensor fields on a given route. Simultaneous use of multiple UAVs leads both to creation of a swarm and to the possibility of considering it as a swarm of the queuing network. Notable works of UAV swarms as a part of FANET usually pursued the target of cooperation the UAV opportunities for solving military tasks, for search of the target, etc. [12, 13, 14]. We believe that the wide spread of public unmanned aerial vehicles enables to consider a separate UAV as a queuing system [10] and a swarm as a queuing network.

2. UAV as a queuing system

Let sensory nodes FUSN-P which are considered, for example, for the head [15], are located on the UAV, which perform the flight of the sensor field territory (terrestrial network USN) and collect the data from the terrestrial-based sensor nodes. While servicing a plurality of nodes, the UAV can be seen as a queuing system, the input

of which receives the entity (terrestrial sensor nodes in the service area) which can expect the service within the time of their stay in the area of accessibility. The entities (nodes) that have not been serviced during this time are denial of the service. The flow rate is dependent on the radius of the service area, the density of nodes and the speed of the UAV. To serve the UAV terrestrial sensor assembly some time is spent and the node should be in the area of accessibility during the period of service (Figure 1).

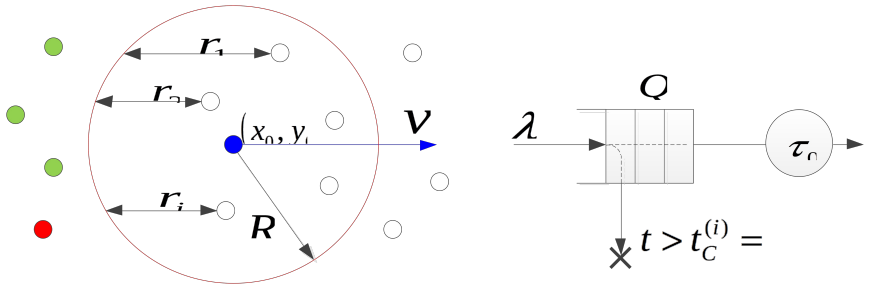


Figure 1: UAV as a queuing system.

If the coordinates of the terrestrial nodes are accidental, the entry system receives the random flow of the entities. The properties of this flow are determined by the properties of the sensor field (publishing sites on the surface), the radius of the service drones and its speed. We will make the following assumptions:

- the sensory field is a Poisson field;
- the UAV is believed to move in a straight line at a constant velocity v;
- the zone service is a circle with a radius R.

Define the distribution function for the incoming flow entities. For this purpose we will examine the service area of the UAV at time 0 and at time t. During t the entities (nodes) which are found in the area that is defined by a shift of the UAV service area for time t will go in the system. According to the properties of the Poisson field, the probability of presence of n points (nodes) in a certain area is determined by the Poisson distribution and depends only on the field area. The probability of presence of z entities (nodes) in the field S is

$$p_z = \frac{a^z}{z!} e^{-a} \tag{1}$$

where a = p * S; p - is the number of points (nodes) in a unit area; S - is the field area.

$$p_z(t) = \frac{(p \cdot S(t))^z}{z!} e^{-p \cdot S(t)} \tag{2}$$

The field area can be defined as

$$S(t) = 2R \cdot vt \tag{3}$$

The flow rate, i.e. the average number of entities per unit of time is equal to

$$\gamma = p \cdot 2R \cdot vt \tag{4}$$

The distribution of the time interval between the entities We will consider the random variable T as the time interval between two successive events in the stream and will find its distribution function.

$$F(t) = P(T < t) \tag{5}$$

Then the probability that z entities will go to the time section of the length t is

$$P(T \geq t) = 1 - F(t) \tag{6}$$

Therefore, the probability can be calculated by the formula

$$P(T \geq t) = p_0(t) = e^{-p \cdot 2 \cdot R \cdot vt} \tag{7}$$

Considering this fact, the distribution function of the time interval between the entities is

$$F(t) = 1 - e^{-p \cdot 2 \cdot R \cdot vt} \tag{8}$$

Thus, the elementary flow will enter the system, the time intervals between the entities, which are distributed exponentially with a mean.

3. Swarm of UAV-P as a queuing network

Taking into consideration the above mentioned facts, the flow of entities (messages), which arrives at the node of each of the UAV has the properties of a simple entity flow. Beside the flow of messages from a particular terrestrial sensor field, the viewed nodes receive the traffic flows from other nodes on the network.

Further we will assume that the output flow of messages from i node with probability r_{ij} is an input to the node j. With probability

$$1 - \sum_{j=1}^{n} r_{ij} \tag{9}$$

the entities will leave the node i and will be sent to the external environment, i.e., to the gateway, Fig. 2.

In the general case, the service time of the messages on the route segment t consists of two main components: the time of sending the message on channel τ and the time-out state of the channel readinessψ, which are generally random.

Changing of the channel status is a random process that occurs under the influence of many independent factors (events), such as the entry and stepping out of communication range due to the random deviations from the desired path of movement, the effect of interference from transmitters located on the other elements of the system

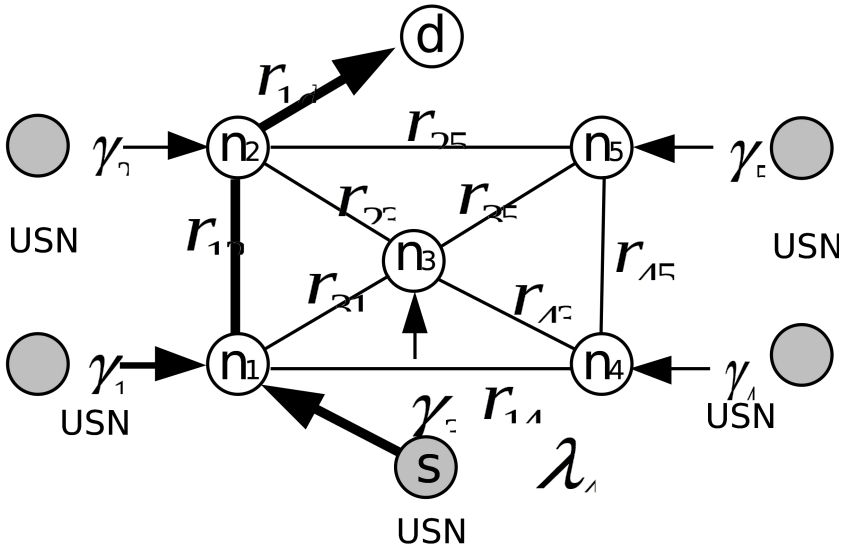


Figure 2: Model of data delivery route between the source (s) and receiver (t).

and others. It is expected that with a sufficiently large number of independent events the channel readiness intervals will have the distribution which is close to exponential distribution, therefore, the state of waiting time readiness ψ will also have the similar distribution.

If the time distribution of the message sending through the communication channel τ is close to an exponential one, then the assumption of exponential distribution of service time t is quite possible.

If we strengthen the above mentioned conditions of the network by the assumption of exponential service time of messages in the nodes, these conditions will coincide with the conditions of the network Jackson [16].

$$\dot{T} = \sum_{j=1}^{M} \frac{\lambda_j}{\gamma} T_j \tag{10}$$

where M - is the number of channels in the network; n - is the number of network nodes; T_j - is the delay in the j-th channel; $\gamma = \sum_{i=1}^{n} \gamma_i$ - is the total traffic network; λ - is the total traffic served in the j-th channel.- is the delay in the j-th channel;

The value $T_j = \frac{1}{\mu_j - \lambda_j}$, where $\mu_j = \frac{1}{t_j}$ - is the service rate in the j-th channel.

Delivery time for a particular route network θ_k can be estimated by using the properties of the Jackson network. It is known that each node of the network can be considered as independent QS M/M/1, and the whole route — as a series of independent QS M/M/1, fig.3.

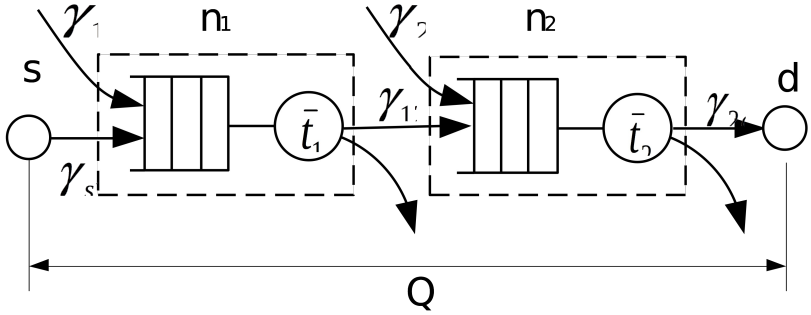


Figure 3: Model of data delivery route between the source (s) and receiver (t).

The distribution function of time to deliver a message in this system can be described by Erlang distribution.

In case of equality of all $\lambda_i = \lambda$, and $\mu_i = \mu$ with the average value $m \cdot t$, which is the average time to deliver a message on the route $\theta_k = m_k \cdot t$ where m_k is the number of channels in the k-route.

$$S(x, m) = \frac{m \cdot \mu \cdot (m \cdot \mu \cdot x)^{m-1}}{(m-1)!} e^{-m \cdot \mu \cdot x} \tag{11}$$

The order m, in this case, corresponds to the number of transits (hops), assuming that the message transmission (service) for each of them is equal.

The more accurate approximation of the viewed network as the Jackson network is, the more n there are and the nearer service time of distribution blitz to exponential distribution is. With a relatively small number of n nodes and a small number of routes network the properties can significantly differ from the properties of the Jackson network. In this case, the route pattern can be described as a multiphase system G/G/1. Getting of the distribution function of delivery time, in this case, can be very difficult. However, an approximate estimate of the average delivery time to the j channel route is possible, as it is shown in [16]

$$\widetilde{T} \approx \frac{p_j \cdot \bar{t}}{2(1-p_j)} \frac{\sigma^2_{a_j} + \sigma^2_{t_j}}{\bar{t}_j^2} \frac{\bar{t}^2_{a_j} + \sigma^2_{t_j}}{a_j^2 + \sigma^2_{t_j}} \tag{12}$$

Where $p_j = \lambda_j \bar{t}_j$; $\sigma^2_{a_j}$ - dispersion of intervals between messages; $\sigma^2_{t_j}$ - dispersion of service time in j channel; $\bar{t}_j$ - Service time in j channel; $a_j = \frac{1}{\lambda_j}$ - the mean value of the interval between messages in j channel.

Then the delivery time on the route will be equal to

$$\theta_k = \sum_{j=1}^{m_k} \widetilde{T}_j \tag{13}$$

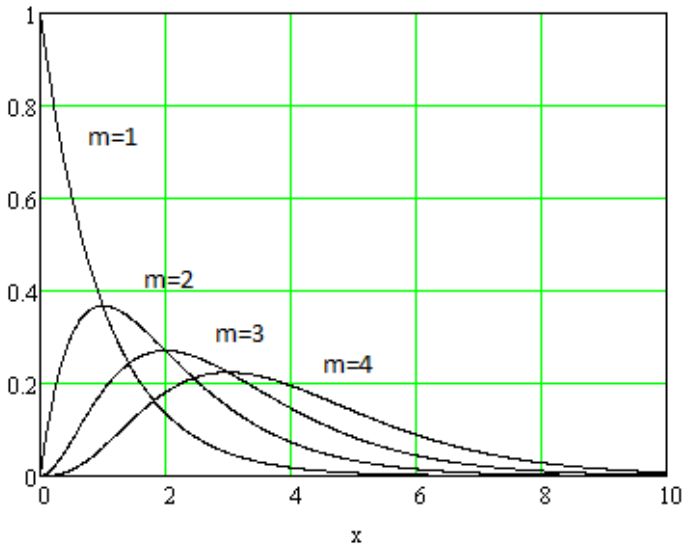


Figure 4: The probability density of the delivery time on the route of m length = 1,2,3,4 hops.

where m_k is the number of channels in the k route.

It should be noted that the more accurate the estimate of the mean delivery time on the route 4 and 5 is, the higher the intensity of the message flow is λ_j and the smaller relation between the service time in the channel and time of messages receipt on the input of each channel.

It is obvious that one of the determining factors of the delivery time is the number of ”hops” (channels) in the route m. This number depends on the used methods of routing. It is logical to assume that the route of the minimum length (with a minimum number of ”hops”) is chosen. Figure 5 shows the implementation of the random distribution of nodes in the space which is defined by a cube 200x200x200 m (a) and by an equal volume of a sphere (b).

Figure 6 shows the distribution of the lengths of the shortest paths in the network which is formed by nodes that are arranged in a cube, with a communication node radius of 50 m.

This distribution was obtained by simulation. The shortest route was chosen by the criterion of a minimum number of hops. The average path length was 4.47 hop. For comparison, the same figure shows a Poisson distribution with a mean of 4.47. The connectivity probability was 0.98.

Figure 7 shows the distribution of the lengths of the shortest paths in the network with a random arrangement of 100 nodes in the area with a communication network node radius of 50 m.

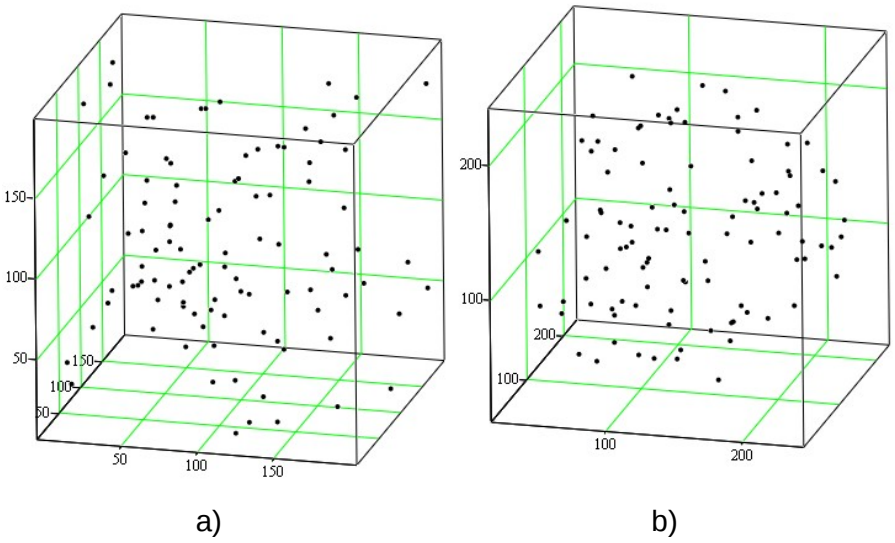


Figure 5: Random placement of 100 nodes in the cube 200x200x200 m (a) and in the sphere of equal volume (b) 14.

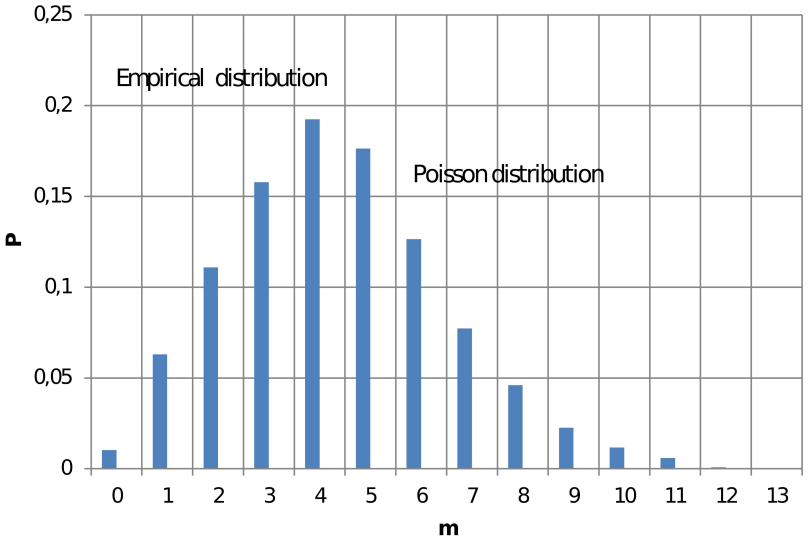


Figure 6: Distribution of the lengths of the shortest paths in the network of 100 nodes in the cube m 200x200x2004.

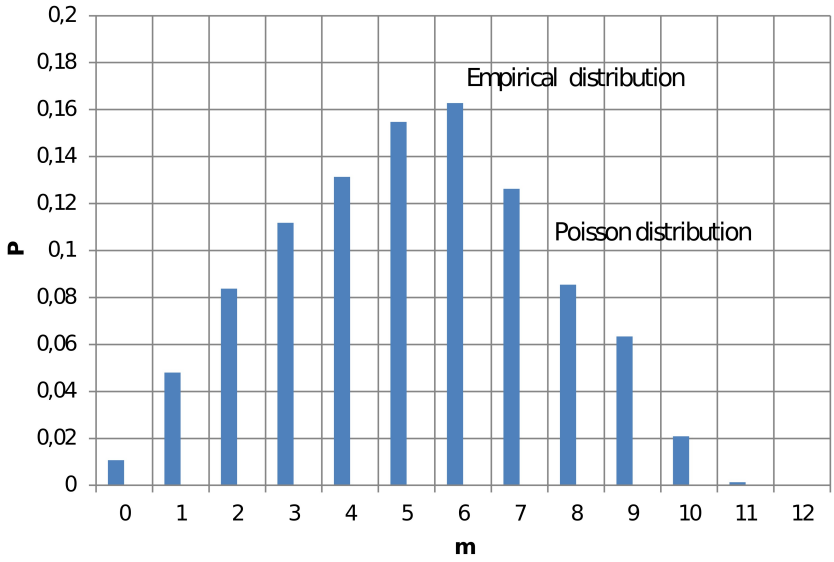


Figure 7: The distribution of the lengths of the shortest paths in the network of 100 nodes in an equal volume area.

The average path length was 5.18 hop. For comparison, the same figure also shows a Poisson distribution with a mean of 5.18. In this case, the network connectivity was 0.94.

The connectivity probability can be defined as the probability of falling into the sphere of a given radius of at least one node.

Out of the properties of the Poisson field, this probability is

$$P \geq 1 = 1 - e^{-a} \tag{14}$$

Where a - is the expected number of points in the field.

$$a = V \cdot p \tag{15}$$

where $V = \frac{4}{3}\pi \cdot x^3$ is the sphere volume of radius x; p - is the sphere volume of radius x;

Then the dependence of the probability density and connectivity of the network node communication radius is equal to

$P = 1 - e^{\frac{4}{3}\pi \cdot x^3 \cdot p}$

For the simulated network, starting from (17), it is equal to 0.999. The values of connectivity which are obtained from the simulation results are within the error due to the finite size of the sample. It should be noted that the expression (17) gives the probability of connectivity for the unlimited Poisson field. In this case, the field is

limited with a certain volume. In the case of restrictions, "edge effect" takes place which considers that the probability of connectivity for the nodes near the border is less than for the nodes that are closer to the center of considered limiting volume of the figure. This is obvious when considering a node which is located strictly at the boundary field.

The adjacent to it node can be located within the area. If the boundary is the plane, the extent to which communication with the neighboring node is possible is less than half for the site located near the center of the examined area (if the communication range is smaller than the area of the node). In this regard, it should be expected that the assessment of the connection probability (17) is the upper bound. Also the closer to the probability the value of the connected network will be (17), the larger the ratio of bounding shape to its surface area is. It is obvious that by increasing of the geometric dimensions the ratio will increase. As it is seen from the given figures in the case of considering the limited space of a cube, the length of the shortest path is well described by a Poisson distribution. In the case when the space is limited by a sphere, the distribution of the lengths of the shortest paths differs from the Poisson distribution to a greater extent. The average lengths of the shortest path (in the race) in the cases of cube and sphere are expected to vary.

4. Conclusion

1) While the organization of interaction with UAVs USN nodes to collect data under the certain conditions, the network connections between the UAV can be seen as a queuing network.
2) When a sufficiently large number of nodes which are located on the UAV model, the network Jackson can be used. In this case, the delivery time of data between the sources and receiver will obey the law of Erlang.
3) With a relatively small number of UAV to estimate the time of the data delivery it is possible to use familiar approximate estimates for systems G/G/1.
4) The number of "hops" in the shortest route between the nodes of the UAV is distributed according to the law which is close to the Poisson law that enables to estimate the length of the routes and the delay of the data delivery.

Acknowledgments. The reported study was supported by RFBR, research project No15 07-09431a "Development of the principles of construction and methods of self-organization for Flying Ubiquitous Sensor Networks".

REFERENCES

1. I.Bekmezci, O.K.Sahingoz, S.Temel. Flying Ad-Hoc Networks: A Survey. Ad Hoc Networks, Elsevier, v.11, issue 3, May 2013.
2. O.K.Sahingoz. Networking Model in Flying Ad Hoc Networks (FANETs): Concepts and Challenges. Journal of Intelligent Robotics Systems. V.74, issue 1-2, Springer, 2014.
3. S.K.Singh. A Comprehensive Survey on Fanet: Challenges and Advancements. International Journal of Computer Science and Information Technologies, v.6 (3), 2015.

4. D.Rosario, Z.Zhao, T.Braun, E.Cerqueira, A.Santos. A Comparative Analysis of Beaconless Opportunistic Routing Protocols for Video Dissemination over Flying Ad-hoc Networks. The 14th International Conference on Internet of Things, Smart Spaces, and Next Generation Networks and Systems. NEW2AN 2014. LNCS 8638, Springer, Heidelberg, — 2014.
5. E. P. de Freitas, T. Heimfarth, I. F. Netto, C. E. Lino, C. E. Pereira, A. M. Ferreira, F. R. Wagner, and T. Larsson. UAV relay network to support WSN connectivity. ICUMT, Proceedings, 2010.
6. Vinel, A., Vishnevsky, V., Koucheryavy, Y. "A simple analytical model for the periodic broadcasting in vehicular ad -hoc networks", 2008 IEEE Globecom Workshops, GLOBECOM 2008.
7. J Jakubiak, Y Koucheryavy. State of the art and research challenges for VANETs. Consumer communications and networking conference, 2008. CCNC 2008. 5th IEEE
8. I.F. Akyildiz, M.C. Vuran, O.B. Akan, W. Su. Wireless Sensor Networks: A Survey revisited. Computer Networks Journal, 2005
9. A. Koucheryavy, A. Salim. Prediction-based Clustering Algorithm for Mobile Wireless Sensor Networks. Proceedings, International Conference on Advanced Communication Technology, 2010. ICACT 2010. Phoenix Park, Korea.
10. R.Kirichek, A.Paramonov, A.Koucheryavy. Flying Ubiquitous Sensor Networks as a Quening System. Proceedings, International Conference on Advanced Communication Technology, 2015. ICACT 2015, Phoenix Park, Korea.
11. A.Koucheryvy, A.Vladyko, R.Kirichek. State of the Art and Research Challenges for Public Flying Ubiquitous Sensor Networks. The 15th International Conference on Internet of Things, Smart Spaces, and Next Generation Networks and Systems. NEW2AN 2015. LNCS, Springer, Heidelberg, 2015 (accepted).
12. Y. Altshuler, V. Yanovsky, I. Wagner, A. Bruckstein, The cooperative hunters-efficient cooperative search for smart targets using UAV swarms, in: Second International Conference on Informatics in Control, Automation and Robotics (ICINCO), the First International Workshop on Multi-Agent Robotic Systems (MARS), 2005.
13. R.R.McCunea, G.R.Madey. Swarm Control of UAVs for Cooperative Hunting with DDDAS. International Conference on Computational Science, ICCS 2013.
14. G. Madey, M. Blake, C. Poellabauer, H. Lu, R. McCune, Y. Wei. Applying DDDAS principles to command, control and mission planning for UAV swarms. Procedia Computer Science, 9, 2012.
15. A.Futahi, A.Koucheryavy, A.Paramonov, A.Prokopiev. Ubiquitous Sensor Networks in the Heterogeneous LTE Network. Proceedings, International Conference on Advanced Communication Technology, 2015. ICACT 2015, Phoenix Park, Korea.
16. L.Kleinrock. Queueing Systems. V.1.Theory. J.Wiley Sons, 1975.
17. L.Kleinrock. Queueing Systems. V.2.Computer Applications. J.Wiley Sons, 1976.

DEVELOPMENT OF WIRELESS CAMERA SENSOR NETWORK MODEL

A. Karpov[1], *L. Voskov*[2], *S. Efremov*[2]
[1] National Research University Higher School of Economics (HSE), Moscow, Russia,
[2] Moscow State Institute of Electronics and Mathematics Higher School of Economics (MIEM HSE), Moscow, Russia

Abstract

This paper describes a model of wireless camera sensor network with autonomous power sources, taking into account the specifics of such networks. We present new approach to definition of the energy efficiency in camera sensor networks based on user requirements. We also discuss the factors which influence on the image recognition process in sensor nodes and distinguish the levels of intelligence of camera sensor network.

РАЗРАБОТКА МОДЕЛИ ФУНКЦИОНИРОВАНИЯ БЕСПРОВОДНОЙ СЕНСОРНОЙ СЕТИ КАМЕР

А.В. Карпов[1], *Л.С. Восков*[2], *С. Г. Ефремов*[2]
[1] Национальный исследовательский университет "Высшая школа экономики", Москва, Россия,
[2] Московский институт электроники и математики Национального исследовательского университета "Высшая школа экономики", Москва, Россия
karpov-av2@narod.ru, voskov@narod.ru, efremov-sg@narod.ru

Аннотация

В данной работе представлена модель функционирования сенсорной сети камер с автономными источниками питания, учитывающая специфику ее работы. Рассматривается подход к определению энергетической эффективности сенсорной сети камер на основе учета требований пользователей. Также определяются факторы, влияющие на распознавание изображений на оконечных узлах, выделяются уровни интеллектуальности сети камер.

Ключевые слова: сенсорная сеть камер, модель, энергетическая эффективность, требования пользователей, уровни интеллектуальности

1. Введение

Сенсорная сеть камер (camera sensor network, wireless image sensor network, visual sensor network, smart cameras network) – беспроводная сенсорная сеть, где в качестве основного сенсора используется маломощная видео- или фотокамера. Цель развертывания сенсорной сети камер состоит в удаленном получении информации об объектах мониторинга и ее передача на центральный узел системы в течение длительного промежутка времени [1, 2, 3, 4].

Модель получения информации камерой по своей природе отличается от модели получения информации любого другого типа сенсора. Как правило, датчик собирает данные из окружающей среды на расстоянии срабатывания. Камера, в свою очередь, характеризуется моделью направленного получения информации - она получает изображения удаленных объектов в определенном направлении, так называемом «поле зрения» (Field of View). Благодаря этому становится возможным бесконтактное измерение видимых характеристик объекта. Например, в некоторых случаях бывает невозможно или невыгодно измерять, фиксировать наступление события каким-либо инструментом, устройством (трещина в стене, аналоговый прибор со стрелкой и т.д.), которое бы прямо, а не косвенно измеряло бы требующийся параметр. В этих случаях возможно использование камеры, которая является универсальным устройством бесконтактного измерения или датчиком, фиксирующим наступление событий.

Одно из основных отличий сенсорных сетей камер от других видов сенсорных сетей состоит в природе того, как камера получает информацию из окружающей среды. Большинство сенсоров проводят измерения одномерных сигналов. Однако матрица камеры состоит из большого числа фоточувствительных ячеек. Одно сделанное камерой измерение обеспечивает получение двумерного массива данных, который мы видим как изображение. В результате дополнительной размерности, набор данных имеет больший объем информации, а также более высокую сложность обработки и анализа по сравнению с обычными сенсорами. В связи с этим, увеличивается количество первоначально получаемой информации, поскольку объем данных, занимаемый изображением, значительно больше, чем скалярных данных (например, показания датчика температуры).

В стандартных сенсорных сетях для измерения физических величин (температуры, давления, влажности и т.д.) используются датчики, которые получают скалярные данные. В этом случае полезность информации, заключенная в самом числе, для пользователя максимальна, поскольку количество информации, занимаемое числом, нельзя сократить, только если использовать методы сжатия, агрегирования данных или изменять частоту их сбора с датчика. При получении изображений объекта камерой полезная информация, которую необходимо извлечь, как бы распределяется по всему объему данных, занимаемому изображением.

Поскольку оконечные устройства в системе являются автономными, то запасы энергетических ресурсов сильно ограничены. Кроме того, объемы изображений значительно больше объемов данных, получаемых с сенсоров температуры, влажности в стандартных сенсорных сетях, в результате этого особо остро встает вопрос эффективного использования энергетических ресурсов.

Целью работы является повышение энергетической эффективности беспроводной сенсорной сети камер с автономными источниками питания.

Объектом исследования является беспроводная стационарная сеть с автономными источниками питания, в которой каждый оконечный узел включает фотокамеру.

Предметом исследования является энергетическая эффективность сенсорной сети камер.

2. Новый подход к определению энергетической эффективности сенсорной сети камер на основе учета требований пользователей

Эффективность функционирования сенсорной сети камер, главным образом, определяется количеством получаемой/передаваемой информации по отношению к затратам энергетических ресурсов[5]. Как правило, энергетическая эффективность сенсорной сети камер рассматривается с точки зрения затрат на передачу одного бита информации. Однако в этом случае не рассматривается, содержит ли переданный бит полезную информацию. Для определения энергетической эффективности сенсорной сети камер предлагается рассматривать работу сети с точки зрения полезности передаваемых пользователю данных.

Предлагается новый подход к определению энергетической эффективности сенсорной сети камер с автономными источниками питания, на основе учета запросов пользователей. Например, пользователю необходимо получить информацию о цвете птицы, другому пользователю узнать ее размеры, третьему – определить вид птицы (это также может быть необходимо одному пользователю, но в разные моменты времени). Чтобы не передавать изображение целиком, можно в зависимости от запросов пользователя отправлять только запрашиваемую информацию, то есть только полезные данные. При этом в некоторых случаях может быть не энергоэффективно распознавать и отправлять данные (например, распознать вид птицы на изображении), тогда возможна передача изображения целиком или его предобработка. Предлагается оконечному узлу самостоятельно решать, передавать изображение целиком или проводить его обработку и затем передавать распознанные данные.

Для повышения эффективности использования энергетических ресурсов сети необходимо максимизировать количество передаваемой полезной информации по сети и минимизировать суммарные затраты энергии узлами сети. Локальная обработка изображений на оконечном узле сокращает

общее количество передаваемых по сети данных. Она может включать простые алгоритмы обработки изображений (вычитание фона для детектирования движения/объектов, детектирования краев), а также более сложные алгоритмы компьютерного зрения, такие как выделение характерных точек (feature extraction), классификация объектов, вплоть до распознавания образов и сцены (scene reasoning).

Факторы, влияющие на распознавание изображений на оконечных узлах:

- сложность объекта наблюдения;
- используемые алгоритмы обработки и распознавания;
- условия съемки (темное время суток, погодные условия и т.д.).

Таким образом, в зависимости от приложения, оконечный узел с камерой может обеспечивать разные уровни интеллектуальности, которые определяются сложностью используемых алгоритмов обработки [6].

Низкоуровневые алгоритмы обработки (такие алгоритмы как вычисление разности фреймов для детектирования движения или детектирования краев) может выполнить оконечное устройство, используя основную информацию об окружении, и помочь решить или передавать изображение другим устройствам или продолжить обработку изображения на более высоком уровне.

Более сложные алгоритмы компьютерного зрения (например, выделение признаков объекта, классификация объектов и т.д.) позволяют, например, обеспечивать базовую классификацию полученного объекта. Кроме того, камеры могут взаимодействовать между собой путем обмена информацией о полученных признаках объектов, что позволяет в дальнейшем коллективно принимать решения. С этой точки зрения сенсорная сеть камер становится независимой от пользователя, интеллектуальной системой распределенных камер, которая обеспечивает получение актуальной информации об объекте мониторинга.

3. Уровни интеллектуальности сети камер

В статье [7] авторы выделяют уровни интеллектуальности видеомониторинга. Под интеллектуальным видео-мониторингом авторы понимают любой мониторинг, в котором обработка видео выполняется непосредственно на стороне камер. Таким образом, проводится распределенная обработка информации, снижаются требования к пропускной способности канала.

В существующих системах видео-мониторинга интеллектуальность сети отсутствует (Рис. 1). В системах видео-мониторинга, в которых обработка изображений проводится на оконечных модулях, можно выделить четыре уровня интеллектуальности:

1) На первом уровне осуществляется детектирование движения (motion detection), таким образом, пользователю передаются только кадры, в которых зафиксировано движение.

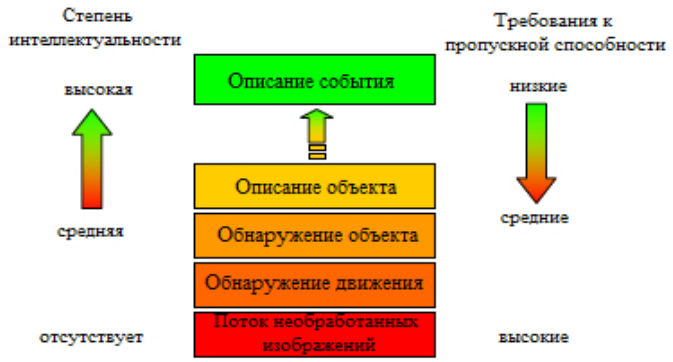


Рис. 1: Уровни интеллектуальности видео-мониторинга

2) На втором уровне камеры могут производить детектирование объекта (object detection), его классификацию (object classification), таким образом, пользователю передаются кадры, на которых зафиксировано какое-либо движение и включающие объект или группу объектов наблюдения.
3) На третьем уровне возможна организация коллективного взаимодействия нескольких камер с целью идентификации объекта и передачи пользователю его текстового описания вместе со снимком.
4) На четвертом уровне сеть интеллектуальных камер лишь уведомляет пользователя о наступлении интересующего его события, отправив пользователю текстово-визуальное или полностью текстовое описание события.

Таким образом, с увеличением уровня интеллектуальности при переносе процесса обработки изображений на оконченные узлы снижаются требования к пропускной способности канала, поскольку передается меньший объем данных, однако повышаются требования к вычислительным ресурсам оконечных узлов, производящих обработку данных.

Для проведения дальнейших исследований необходимо разработать модель беспроводной сенсорной сети камер с автономными источниками питания, учитывающую специфику ее работы.

4. Модель функционирования сенсорной сети камер с автономными источниками питания

В сети можно выделить 3 типа устройств: координатор, оконечное устройство и маршрутизатор. Далее проводится расчет потребляемой энергии каждым устройством в сенсорной сети камер.

Общая формула затрачиваемой энергии при передаче данных от оконечного узла координатору через маршрутизаторы (1):

$$E_{TOTAL} = E_{ED} + E_{ROUTER} \cdot N_{ROUTER} + E_{COORD} \quad (1)$$

, где E_{ED} - энергия, затрачиваемая оконечным устройством на получение, обработку и передачу изображения,
E_{ROUTER} - энергия, затрачиваемая маршрутизатором на передачу данных,
N_{ROUTER} - количество маршрутизаторов,
E_{COORD} - энергия, затрачиваемая координатором на прием изображения.

1) оконечное устройство

Общая энергия, затрачиваемая оконечным устройством с камерой на получение, обработку и передачу данных координатору, составляет (2):

$$E_{ED} = P_{CamGet} \cdot t_{GET} + (P_{CamTX} + P_{active}) \cdot \frac{N_{px} \cdot bpp}{V_{TX}} + P_{active} \cdot t_{proc} + (P_{active} \cdot t_{WAIT} + P_{rx} \cdot t_{CCA} + P_{tx} \cdot t_{DATA} + P_{rx} \cdot t_{ACK}) \cdot N_{frame} \quad (2)$$

, где P_{CamGet} - мощность, затрачиваемая на получение изображения камерой,
t_{GET} - время затрачиваемое на получение изображения камерой,
P_{CamTX} - мощность, затрачиваемая камерой на передачу изображения микроконтроллеру,
P_{active} - мощность, затрачиваемая микроконтроллером при нахождении в активном режиме,
N_{px} - количество пикселей изображения, которое определяется его разрешением,
bpp - глубина цвета изображения (количество бит, отводимое на кодирование одного пикселя изображения),
V_{TX} - скорость передачи данных по интерфейсу, соединяющему память камеры с памятью микроконтроллера,
t_{proc} - время, требующееся на обработку изображения (сжатие, извлечение части изображения, распознавание объектов и т.д.),
t_{WAIT} - время, затраченное на ожидание перед проверкой занятости канала,
P_{rx} - мощность, затрачиваемая микроконтроллером при нахождении в режиме приема данных,
t_{CCA} - время, затраченное на проверку занятости канала,
P_{tx} - мощность, затрачиваемая микроконтроллером при нахождении в режиме передачи данных,
t_{DATA} - время, затраченное на передачу кадра данных,
t_{ACK} - время, затраченное на прием подтверждения,
N_{frame} - количество фреймов, требующих передачи.

2) маршрутизатор, ретранслятор

По алгоритму асинхронного доступа – CSMA-CA без слотов, маршрутизатор всегда должен прослушивать эфир для обнаружения новых устройств и запросов на передачу данных, поэтому он не может переходить в режим пониженного энергопотребления. Маршрутизатор должен периодически прослушивать сеть, принимать данные предназначенные ему для передачи другому устройству сети, отправлять кадр подтверждения о получении данных, передавать данные другому устройству, принимать кадр подтверждения о получении данных другим устройством. Таким образом, формула для расчета потребляемой энергии маршрутизатором при передаче данных имеет вид (3):

$$t_{RXframe} = t_{DATA} + t_{ACK}$$

$$P_{RXframe} = \frac{P_{rx} \cdot t_{DATA} + P_{tx} \cdot t_{ACK}}{t_{RXframe}}$$

$$E_{ROUTER} = (P_{RXframe} \cdot t_{RXframe} + P_{tx} \cdot t_{frame}) \cdot N_{frame} \qquad (3)$$

3) координатор

В функции координатора входит управление сетью, подключение новых устройств к сети, соответственно координатор должен периодически прослушивать эфир с целью трансляции информации о сети другим устройствам, обнаружения запросов на подключение от новых устройств.

Для упрощения расчетов не будем учитывать служебные кадры, генерируемые координатором для создания сети и ее управления. Тогда общая энергия, затрачиваемая координатором при получении данных от оконечного устройства или маршрутизатора, составляет (4):

$$E_{COORD} = (P_{rx} \cdot t_{DATA} + P_{tx} \cdot t_{ACK} + P_{active} \cdot t_{TX}) \cdot N_{frame} \qquad (4)$$

, где P_{active} - мощность координатора, затрачиваемая на передачу данных компьютеру для последующей обработки,
t_{TX} - время, затраченное координатором на передачу данных компьютеру, зависящее от пропускной способности проводного канала и объема передаваемых данных.

Как правило, если сеть является небольшой, то координатор напрямую подключается к более мощному вычислительному устройству (компьютеру) по проводному каналу связи, на котором происходит дальнейшая обработка принятой информации и питается от стационарной электросети. Таким образом, запасы энергии координатора или остаются постоянными, или зависят от запасов энергии более мощного устройства, и вопрос энергоэффективной работы координатора не стоит так остро, как с вариантом автономного устройства.

В дальнейшем, в модели планируется учитывать запросы пользователей, вычислительную сложность алгоритмов обработки изображений.

Энергия, затрачиваемая на получение изображения, анализ ситуации (передавать необработанное изображение или распознавать объект), обработку данных, передачу распознанных данных координатору должна быть меньше энергии, затрачиваемой на получение изображения и передачу «сырого» изображения координатору.

5. Заключение

В работе была представлена модель функционирования сенсорной сети камер с автономными источниками питания. Предложен новый подход к определению энергетической эффективности сенсорной сети камер на основе учета запросов пользователей. Также определены факторы, влияющие на распознавание изображений на оконечных узлах, выделены уровни интеллектуальности сети камер.

Данное научное исследование (исследовательский проект №14-05-0064) выполняется при поддержке Программы «Научный фонд НИУ ВШЭ» в 2014/2015гг.

ЛИТЕРАТУРА

1. J. Lloret, I. Bosch, S. Sendra, A. Serrano A Wireless Sensor Network for Vineyard Monitoring That Uses Image Processing. // Sensors. 2011. Vol.11. Pages 6165-6196.
2. Kays, R., B. Kranstauber, et al. Camera traps as sensor networks for monitoring animal communities. // The 34th IEEE Conference on Local Computer Networks. 2009. Pages 811-818.
3. Bir Bhanu, Chinya V. Ravishankar, Amit K. Roy-Chowdhury, Hamid Aghajan, Demetri Terzopoulos Distributed Video Sensor Networks. // Springer-Verlag London Limited. 2011.
4. Teresa A. Dahlberg, Asis Nasipuri, Craig Taylor Explorebots: A Mobile Network Experimentation Testbed. // SIGCOMM'05 Workshops. August 22-26, 2005.
5. Восков Л. С., Ефремов С. Г. Задача увеличения времени автономной работы беспроводных сенсорных сетей в системах сбора данных и способ ее решения // Датчики и системы. 2013. №4(167). С. 2-9.
6. S. Hengstler, D. Prashanth, S. Fong, and H. Aghajan Mesh-Eye: a hybrid-resolution smart camera mote for applications in distributed intelligent surveillance. //in Proceedings of the 6th International Symposium on Information Processing in Sensor Networks (IPSN'07), pp. 360-369, 2007.
7. S. Hengstler, D. Prashanth, S. Fong, H. Aghajan MeshEye: A Hybrid-Resolution Smart Camera Mote for Applications in Distributed Intelligent Surveillance. // IPSN'07, April 25-27, 2007, Cambridge, Massachusetts, 2007.

AUDIO-DATA TRANSMISSION MODEL FOR WIRELESS SENSOR NETWORKS WITH QoS

I.V. Karpov[1], *L.S. Voskov*[2], *S.G. Efremov*[2]
[1] National Research University Higher School of Economics (HSE), Moscow, Russia, [2] Moscow State Institute of Electronics and Mathematics Higher School of Economics (MIEM HSE), Moscow, Russia

This paper focuses on the audio-data transmission model for Wireless Sensor Networks with Quality of Service support. Since audio sensor networks consist of autonomous nodes with limited energy and hardware resources, transmission quality declines, when multiple audio streams are transferred simultaneously in the network. Unlike existing models, the developed audio-data transmission model takes into account Quality of Service. In conclusion we present our method of data transmission for multiple audio streams in the network.

МОДЕЛЬ ПЕРЕДАЧИ АУДИОДАННЫХ ПО БЕСПРОВОДНЫМ СЕНСОРНЫМ СЕТЯМ С УЧЕТОМ КАЧЕСТВА ПРЕДОСТАВЛЯЕМОГО СЕРВИСА

И.В. Карпов[1], *Л.С. Восков*[2], *С. Г. Ефремов*[2]
[1] Национальный исследовательский университет "Высшая школа экономики", Москва, Россия,[2] Московский институт электроники и математики Национального исследовательского университета "Высшая школа экономики", Москва, Россия
ivakarpov@gmail.com, lvoskov@hse.ru, sefremov@hse.ru

Аннотация

В статье рассматривается модель передачи аудиоданных по беспроводным сенсорным сетям. Поскольку узлы являются автономными и имеют ограниченные ресурсы, то при передаче нескольких потоков аудиоданных появляются задержки, в результате которых качество передачи данных снижается. В отличие от существующих моделей, в разработанной модели учитывается качество предоставляемого сервиса при передаче аудиоданных. Также рассматривается предлагаемый метод передачи нескольких потоков аудиоданных.

Ключевые слова: беспроводная аудио-сенсорная сеть, модель передачи данных, БСС

1. Введение

В данной работе рассматривается модель беспроводной аудио-сенсорной сети, состоящей из случайно распределенных автономных датчиков с одинаковыми техническими характеристиками (объемом памяти, запасами энергии, характеристиками микроконтроллера). Основной задачей сети является организация двусторонней аудио-связи между пользователями сети, то есть передача нескольких потоков аудиоданных между распределенными узлами с заданным качеством обслуживания.

Проведенный обзор литературы выявил небольшое число исследований в области передачи аудиоданных по низкоскоростным сетям, а также отсутствие готовых моделей, которые бы учитывали качество предоставляемого сервиса при передаче данных. Большинство работ посвящено распознаванию полученной аудиоинформации, так как данные сети могут применяться не только в качестве систем аудио-связи [1, 2], но и в качестве систем мониторинга окружающей среды [3, 4, 5, 6].

При передаче аудиоданных по беспроводной аудио-сенсорной сети возникают две основные проблемы, требующие исследования. Во-первых, ставится вопрос о времени работы подобных систем, в связи с ограниченными и не восполняемыми во время работы сети запасами энергии на узлах. Во-вторых, при передаче данных могут возникать задержки и потери пакетов, что критично для аудиоданных. В связи с этим необходимо применять такие протоколы передачи данных и алгоритмы доступа к передающей среде, которые бы позволяли передавать множество потоков данных с минимальными задержками и потерями. Необходимо так построить систему для передачи нескольких потоков аудиоданных, чтобы время ее работы было бы максимальным.

2. Общая модель реконфигурируемой аудио-сенсорной сети

В рассмотренных работах по беспроводным сенсорным сетям для построения общих моделей используют аппарат теории множеств и теории графов [7, 8]. В данной работе за основу взята модель из работы [8].

Сеть представляет собой неориентированный граф со множеством вершин $V_n = \{1, 2, ..., n\}$ и множеством ребер $E_n \subseteq V_n \times V_n$, где вершины представляют собой аудио-сенсорные узлы, а ребра - прямые беспроводные соединения между ними (Рис. 1).

Также имеется множество потоков аудиоданных для передачи: $S = \{s_1, s_2, ..., s_m\}$, где $s_i = (u, v, L_i), u \in V_n, v \in V_n, u \neq v, L_i$ — величина передаваемого трафика (байт/с).

Введем понятие энергетической схемы, важное для оценки времени жизни сети. Энергетическая схема определяет набор мощностей $P = (p_1, p_2, ..., p_n)$, где p_i — мощность, потребляемая i-м узлом сети. Схема определяет связность сети и, следовательно, маршруты передачи потоков данных, таким образом, $E_n = f(w), w \in V_s$, где V_s — множество доступных энергетических схем.

С точки зрения качества обслуживания критичным является смена схем, при которой происходит реконфигурация сети. Введем граф G_s для энергетических схем: $G_s = (V_s, E_s)$. С каждым ребром $e \in E_s$ связан набор коэффициентов $(k_1, k_2, ..., k_m)$ потери качества передаваемых потоков при соответствующей смене схем. Расчет данных коэффициентов является отдельной задачей, не рассматриваемой в настоящей работе.

Таким образом, в общем виде работу аудио-сенсорной сети можно представить в виде четверки:

$$N = (G_s, \Gamma_n, S, \Pi), \tag{1}$$

где G_s — граф энергетических схем (2),
Γ_n — множество сетевых графов. Каждый граф определяется используемой энергетической схемой (3),
S - множество передаваемых по сети аудио-потоков
Π - последовательность применения энергетических схем (4).

$$G_s = (V_s, E_s) \tag{2}$$

$$\Gamma_n = \{G_n(k), k \in V_s\} \tag{3}$$

$$\Pi = (< \pi_1, t_1 >, < \pi_2, t_2 >, ..., < \pi_q, t_q >), \tag{4}$$

$\pi_i \in V_s, t_i$ — время использования сетью схемы π_i

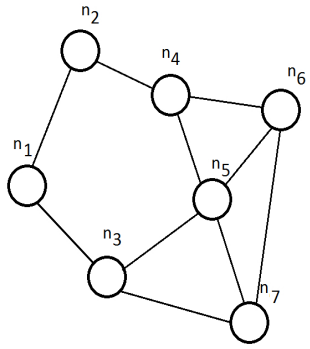


Рис. 1: Беспроводная аудио-сенсорная сеть

Мощность, необходимая для корректной передачи данных от узла i к j должна удовлетворять следующему неравенству:

$$\frac{P_i}{R_{i,j}^{\alpha}} \geq \beta \tag{5}$$

где P_i — мощность, необходимая для успешной передачи данных, $R_{i,j}^{\alpha}$ — расстояние между узлами i, j, α — показатель затухания сигнала, $\beta \geq 1$

— коэффициент, учитывающий качество передачи. Обычно, коэффициент $\beta = 1$, в то время как α зависит от окружающей среды. В идеальной среде коэффициент $\alpha = 2$, но в реальности в интервале [2, 6]. Распространение сигнала в идеальном случае представляют в виде круга с радиусом r, однако в реальных условиях в связи с неоднородностью среды (из-за неравномерного затухания сигнала по всей плоскости распространения), данное упрощение не наблюдается. Тем не менее, оно применимо в работах по беспроводным сенсорным сетям. Также, в реальных условиях, мощности, необходимые для передачи данных от узла i к j и от узла j к i со временем могут изменяться и быть несимметричными. В работе было принято, что в течение времени работы сети, условия окружающей среды не изменялись и оставались постоянными, а мощности, требуемые для передачи - одинаковыми. Также в работе не учитывался физический уровень передачи данных, когда на принимающей стороне возникали ошибки в результате интерференции радиоволн от объектов окружающей среды, а только учитывались факторы, влияющие на передачу данных от элементов сети.

Поскольку узел имеет несколько уровней мощности приемопередатчика, то каждый сенсор, во время образования и работы сети, может выбрать свою мощность, отличающуюся от остальных соседних узлов. Соответственно, при минимальной мощности дальность связи минимальная и энергопотребление низкое, а при максимальной мощности наоборот. Поэтому выбор первоначальных и последующих значений является важным этапом в работе сети. За счет регулирования дальности связи не только освобождается канал передачи данных, но и увеличивается время работы сети из-за сокращения энергопотребления при равных объемах переданной информации.

Возможны следующие варианты работы сети:

1) Сеть работает с предустановленным уровнем мощности на узлах (без смены мощности во время работы)
   (a) Все узлы выбирают одинаковый уровень мощности, отличный от максимального и минимального
   (b) Все узлы выбирают максимальный уровень мощности
   (c) Все узлы выбирают минимальный уровень мощности
   (d) Каждый узел выбирает свой уровень мощности
2) Сеть работает со сменой уровня мощности на узлах во время работы
   (a) Все узлы меняют мощность одновременно на одинаковую
   (b) Все узлы меняют мощность ондовременно, но каждый выбирает свою
   (c) Каждый узел сам выбирает время изменения мощности приемопередатчика на свое значение.

3. Метод пространственно-повторного разделения канала

Обычно, при работе беспроводной сенсорной сети, мощность приемопередатчика устанавливается автоматически при начале работы системы, ли-

бозадается вручную при ее построении и не изменяется в процессе работы сети. К преимуществу данного подхода, с одной стороны, можно отнести предсказуемость маршрутов передачи данных, поскольку дальность связи, как и соседи, постоянны. Дальность связи, в свою очередь, может быть выбрана максимальной, когда передача данных другому узлу происходит с минимальным количеством ретрансляций, то есть передача происходит как можно ближе к узлу получателю. Минимальная дальность связи подразумевает передачу данных самому ближнему узлу, при этом количество ретрансляций может увеличиться многократно, что влечет к дополнительным временным задержкам.

При рассмотрении следующей сети преимущество неизменной дальности связи превращается в недостаток. Если в сети имеется единая точка сбора данных, то ближайшие к ней узлы истощаются быстрее других, образуются так называемые „энергетические дырки“, время работы сети при этом сокращается в разы. Для решения данной проблемы было предложено изменять дальность связи соседних узлов. Таким образом, энергетические затраты распределяются между соседними узлами, в результате чего, увеличивается общее время работы сети. Пример показан на рисунке 2.

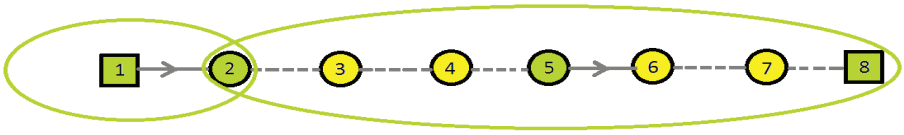


Рис. 2: Пример пространственно-повторного разделения канала

В отличие от классических БСС, когда по сети передаются телеметрические данные (температура, давление и т.п.), нагрузка на сеть при передаче аудиоданных значительно выше, как и требования к их передаче. Для распределения энергозатрат вдоль всего маршрута передачи данных каждого аудиопотока, по аналогии с решением проблемы „энергетических дырок“ около узла сбора данных, предлагается по достижению минимального энергетического порога переназначать дальности связи на узлах, участвующих в передаче. Установленные на узлах мощности называются энергетической схемой. Изменение мощности на узлах предполагает изменение энергетической схемы. Последовательность смены энергетических схем, в дальнейшем, необходимо рассчитать на основе предложенной модели. На рисунке видно, что узел 1 имеет небольшую дальность передачи, в отличие от узла 5. Со временем энергия на узле 5 будет наименьшей по пути передачи данных, что может привести к нарушению работы системы. Согласно предлагаемому методу по достижению минимального уровня энергии на маршруте система изменит энергетическую схему и узел 5 окажется с ми-

нимальной мощностью приемопередатчика, а узел 1 или 2, например, с максимальной.

4. Заключение

В работе была представлена модель беспроводной аудио-сенсорной сети с заданным качеством обслуживания. За счет регулировки мощности приемопередатчиков на узлах, то есть применения энергетических схем можно увеличить время автономной работы сети. Применение динамического изменения дальности связи на узлах является необходимостью.

В дальнейшем планируется провести имитационное моделирование с применением описанного метода регулирования мощности с большим количеством узлов и конкурирующих потоков.

«Данное научное исследование (исследовательский проект №14-05-0064) выполняется при поддержке Программы «Научный фонд НИУ ВШЭ» в 2014/2015гг.»

ЛИТЕРАТУРА

1. Brunelli D., Maggiorotti M. [et al.], Analysis of audio streaming capability of ZigBee networks // Wireless Sensor Networks, Lecture Notes in Computer Science, 2008, vol. 4913, 2008. - P.189-204.
2. Mangharam R., Rowe A. [et al.], Voice over sensor networks// Real-Time Systems Symposium, 2006. RTSS '06.27th IEEE International, 2006. - P.291-302.
3. Werner-Allen G. et al. Fidelity and yield in a volcano monitoring sensor network // Proceedings of the 7th symposium on Operating systems design and implementation. 2006. P. 381-396.
4. Luis E. Palafox, J. Antonio Garcia-Macias, Wireless sensor networks for voice capture in ubiquitous home environments // Wireless Pervasive Computing, 2009. ISWPC 2009.4th International Symposium, 2009. - P.1-5.
5. Hu W., The design and evaluation of a hybrid sensor network for cane-toad monitoring // Information Processing in Sensor Networks, 2005. IPSN 2005. Fourth International Symposium, 2005. - P.503-508.
6. Bidong C., Audio recognition with distributed wireless sensor networks: thes.; University of Victoria - Canada, 2010. - P.59.
7. Базенков Н.И., «Теоретико-игровые алгоритмы формирования децентрализованных беспроводных сетей» : дис. канд. технич. наук. Москва, 2014
8. Ефремов С.Г., «Моделирование времени жизни динамически реконфигурируемых сенсорных сетей с мобильным стоком» : дис. канд. технич. наук. Москва, 2013

HETEROGENEOUS MULTI-PACKET MESSAGE DELAY IN HETEROGENEOUS DATA TRANSMISSION PATH

V. Kokshenev[1], *P. Mikheev*[1], *S. Suschenko*[1], *R. Tkachev*[1]
[1] Tomsk State University, Tomsk, Russia

Abstract

Analysis of heterogeneous message delay in heterogeneous data transmission path is proposed. The data transmission process is modeled by heterogeneous pipeline. The relationship for message delay with analytical equations are defined.

О ЗАДЕРЖКЕ НЕОДНОРОДНОГО МУЛЬТИПАКЕТНОГО СООБЩЕНИЯ В НЕОДНОРОДНОМ ТРАКТЕ ПЕРЕДАЧИ ДАННЫХ

В.В. Кокшенев[1], *П.А. Михеев*[2], *С.П. Сущенко*[3], *Р.В. Ткачев*[4]
[1,2,3,4] Томский Государственный Университет, Томск, Россия,
[1]vladimir_finf@mail.ru, [2]doka.patrick@gmail.com, [3]ssp.inf.tsu@gmail.com,
[4]tkachevrv@mail.ru

Аннотация

Предложен анализ времени передачи неоднородного по длинам пакетов сообщения в тракте передачи данных с различным быстродействием отдельных участков переприема. Процесс передачи моделируется неоднородным по скорости выполнения отдельных фаз конвейером. Получены соотношения позволяющие выполнять расчет времени передачи сообщения

Ключевые слова: задержка сообщения, неоднородное транспортное соединение, конвейерный эффект, неоднородный входной поток пакетов.

1. Введение

Для ускорения передачи и обработки данных в современных компьютерных сетях и вычислительных системах широко применяются методы конвейеризации [1–5]. Конвейерная обработка позволяет одновременно обрабатывать на различных фазах конвейера различные экземпляры потока заявок. Известные результаты исследований быстродействия конвейерной

обработки [1–9] получены только для случаев, когда свойствами неоднородности обладают либо входной поток заявок [3, 6], либо конвейер [7–9]. В реальных системах, с одной стороны трудоемкость предложенных к обработке заявок является существенно неоднородной, а с другой стороны исполнительные элементы этапов конвейера имеют различное быстродействие. В данной работе предложено аналитическое исследование задержки мультипакетного неоднородного по длинам пакетов сообщения в детерминированном неоднородном по быстродействию отдельных участков переприема тракте передачи данных. Важным фактором, определяющим общее время задержки сообщения, является конвейерный эффект проявляющийся в параллелизме передачи различных пакетов сообщения на различных этапах пути.

2. Математическое моделирование

Рассмотрим процесс передачи последовательности из N неоднородных пакетов от источника до адресата по неоднородному транспортному соединению, состоящему из D участков переприема с различным быстродействием. Считаем, что время передачи пакета с номером $n = \overline{1, N}$ на d-м этапе пути ($d = \overline{1, D}$) определяется длительностью τ_{dn}. Дополнительно к этому полагаем, что отправка пакета в очередное звено тракта может начаться только после завершения его передачи на предыдущем этапе пути. При переносе в d-м звене n-го пакета в неоднородном по скорости передачи в отдельных звеньях тракта образуются интервалы ожидания завершения обработки на предыдущем этапе пути следующего ($(n+1)$-го) пакета, обусловленные неоднородностью потока данных:

$$e_{dn} = G\left(\sum_{i=1}^{n} e_{d-1i} + \sum_{i=2}^{n+1} \tau_{d-1i} - \sum_{i=1}^{n-1} e_{di}\right),$$
$$e_{1n} = 0, \quad d = \overline{2, D}, \quad G(x) = \begin{cases} x, & x \geq 0 \\ 0, & x < 0. \end{cases} \tag{1}$$

Определим задержку сообщения $T(D, N)$. Поскольку перенос всех N пакетов в d-м звене транспортного соединения совмещен во времени с передачей $n = \overline{2, N}$ пакетов на предыдущей ($(d-1)$-й) фазе пути, то общее время передачи N пакетов определится соотношением:

$$T(D, N) = \sum_{n=1}^{N} \tau_{1n} + \sum_{d=2}^{D} \left\{ \sum_{n=1}^{N} \tau_{dn} + \sum_{n=1}^{N-1} e_{dn} - \sum_{n=2}^{N} \tau_{d-1n} - \sum_{n=1}^{N-1} e_{d-1n} \right\} =$$
$$= \sum_{d=1}^{D} \tau_{d1} + \sum_{n=2}^{N} \tau_{Dn} + \sum_{n=1}^{N-1} e_{Dn}. \tag{2}$$

Отсюда нетрудно видеть, что для времени передачи однородной последовательности пакетов в однородном тракте ($\tau_{dn} = \tau$, $d = \overline{1, D}$, $n = \overline{1, N-1}$)

справедливо: $T(D,N)=\tau(D+N-1)$. При переносе однородной последовательности пакетов в неоднородном транспортном соединении ($\tau_{dn}=\tau_d$, $n=\overline{1,N}$) получаем:

$$T(D,N)=\sum_{d=1,d\neq M}^{D}\tau_d+N\tau_M,\quad \tau_M=\max_{d=\overline{1,D}}\tau_d.$$

Если же в однородном тракте переносится неоднородная последовательность пакетов сообщения ($\tau_{dn}=\tau_n$, $d=\overline{1,D}$), то общее время обработки составит:

$$T(D,N)=\sum_{n=1,n\neq M}^{N}\tau_n+D\tau_M,\quad \tau_M=\max_{n=\overline{1,N}}\tau_n.$$

3. Анализ передачи упорядоченных последовательностей пакетов в регулярных структурах транспортного соединения

В самом общем случае произвольного расположения пакетов по длинам друг относительно друга и этапов пути найти время передачи последовательности не удается. Однако для регулярных (упорядоченных) последовательностей пакетов и структур тракта передачи данных удается найти общую задержку $T(D,N)$. Проанализируем время (2) при различных соотношениях между τ_{dn}. Если набор пакетов образует невозрастающую последовательность $\tau_{dn}\geq\tau_{dn+1}$, $n=\overline{1,N-1}$, $d=\overline{1,D}$, а звенья тракта — неубывающий по времени передачи на каждой фазе пути порядок $\tau_{dn}\leq\tau_{d+1n}$, $n=\overline{1,N}$, $d=\overline{1,D-1}$, то из (1) нетрудно видеть, что интервалы ожидания имеют нулевую длительность ($e_{Dn}=0$, $n=\overline{1,N}$) и показатель (2) принимает вид:

$$T(D,N)=\sum_{d=1}^{D-1}\tau_{d1}+\sum_{n=1}^{N}\tau_{Dn}. \tag{3}$$

В том случае, когда величины τ_{dn} удовлетворяют невозрастающему порядку в пространстве этапов конвейерного маршрута ($\tau_{dn}\geq\tau_{d+1n}, n=\overline{1,N}$, $d=\overline{1,D-1}$) и неубывающей по трудоемкости передачи последовательности пакетов во входном потоке ($\tau_{dn}\leq\tau_{dn+1}$, $n=\overline{1,N-1}$, $d=\overline{1,D}$), интервалы ожидания e_{dn} задаются выражением

$$e_{dn}=\sum_{i=1}^{d-1}\tau_{in+1}-\sum_{i=2}^{d}\tau_{in},\ n=\overline{1,N-1},\ d=\overline{1,D} \tag{4}$$

и время передачи последовательности становится равным

$$T(D,N)=\sum_{d=1}^{D}\tau_{dN}+\sum_{n=1}^{N-1}\tau_{1n}. \tag{5}$$

Нетрудно видеть, что соотношения (3) и (5) для рассмотренных ограничений на структуру тракта и набор передаваемых пакетов обобщаются выражением:

$$T(D,N)=\sum_{n=1}^{N}\tau_{Mn}+\sum_{d=1,d\neq M}^{D}\tau_{dM},\ \tau_{Mn}=\max_{d=\overline{1,D}}\tau_{dn},\ \tau_{dM}=\max_{n=\overline{1,N}}\tau_{dn}. \quad (6)$$

Пусть τ_{dn} связаны следующими неравенствами: $\tau_{dn}\geq\tau_{d+1n}$, $n=\overline{1,N}$, $d=\overline{1,D-1}$; $\tau_{dn}\geq\tau_{dn+1}$, $n=\overline{1,N-1}$, $d=\overline{1,D}$. Тогда при выполнении дополнительного условия

$$\tau_{dn+1}\geq\tau_{d+1n},\ n=\overline{1,N-1},\ d=\overline{1,D-1}, \quad (7)$$

интервалы e_{dn} определяются соотношением (4), а общая задержка — выражением (5). Поскольку из исходных связей между τ_{dn} следует, что $\tau_{dN}=\min_{n=\overline{1,N}}\tau_{dn}$, то соотношение (5) для общего времени передачи сообщения можно записать в виде:

$$T(D,N)=\sum_{n=1}^{N}\tau_{Mn}+\sum_{d=1,d\neq M}^{D}\tau_{dm},\ \tau_{Mn}=\max_{d=\overline{1,D}}\tau_{dn},\ \tau_{dm}=\min_{n=\overline{1,N}}\tau_{dn}. \quad (8)$$

Если же дополнительное ограничение имеет обратный вид

$$\tau_{dn+1}\leq\tau_{d+1n},\ n=\overline{1,N-1},\ d=\overline{1,D-1}, \quad (9)$$

то интервалы ожидания $e_{Dn}=0$, $n=\overline{1,N}$ и время передачи всего набора пакетов выражается зависимостью (3), интерпретируемой при указанных связях между τ_{dn} соотношением

$$T(D,N)=\sum_{n=1,n\neq M}^{N}\tau_{mn}+\sum_{d=1}^{D}\tau_{dM},\ \tau_{mn}=\min_{d=\overline{1,D}}\tau_{dn},\ \tau_{dM}=\max_{n=\overline{1,N}}\tau_{dn}. \quad (10)$$

В том случае, когда τ_{dn} связаны неравенствами $\tau_{dn}\leq\tau_{d+1n}$, $n=\overline{1,N}$, $d=\overline{1,D-1}$; $\tau_{dn}\leq\tau_{dn+1}$, $n=\overline{1,N-1}$, $d=\overline{1,D}$, при выполнении дополнительного условия (7) e_{dn} вычисляются из (4) и $T(D,N)$ принимает вид (5), эквивалентный обобщающей зависимости (10). При справедливости (9) $e_{Dn}=0$, $n=\overline{1,N}$, а задержка последовательности заявок задается соотношением (3), интерпретируемым выражением (8).

Продолжим изучение интегрального времени передачи набора из N пакетов. Рассмотрим транспортное соединение длины D_1+D_2, в котором первые D_1 последовательных этапов имеют невозрастающее/неубывающее быстродействие, а следующие за ними D_2 этапов имеют обратную зависимость скорости передачи данных от номера этапа. Тогда такой конвейер

имеет два варианта регулярности, не сводящихся к уже рассмотренным случаям:

$$\tau_{dn} \geq \tau_{d+1n}, n=\overline{1,N}, d=\overline{1,D_1-1};\ \tau_{dn} \leq \tau_{d+1n}, n=\overline{1,N}, d=\overline{D_1,D_2-1}; \quad (11)$$

$$\tau_{dn} \leq \tau_{d+1n}, n=\overline{1,N}, d=\overline{1,D_1-1};\ \tau_{dn} \geq \tau_{d+1n}, n=\overline{1,N}, d=\overline{D_1,D_2-1}. \quad (12)$$

С учетом возможных вариантов соотношений между временами передачи набора последовательных пакетов, обусловленных различием их длин:

$$\tau_{dn} \leq \tau_{dn+1},\ n=\overline{1,N-1},\ d=\overline{1,D_1+D_2}; \quad (13)$$

$$\tau_{dn} \geq \tau_{dn+1},\ n=\overline{1,N-1},\ d=\overline{1,D_1+D_2}; \quad (14)$$

найдем общую задержку неоднородного сообщения в транспортном соединении. При выполнении связей (11) и (13) интервалы ожидания между последовательными пакетами до D_1-го этапа пути будут нарастать по закону (4). Поскольку на следующих D_2 этапах маршрута задержка пакетов, по крайней мере, не убывает, то при выполнении неравенства

$$\sum_{i=1}^{D_1-1} \tau_{in+1} - \sum_{i=2}^{D_1} \tau_{in} \geq \sum_{i=D_1}^{D_1+D_2-1} \tau_{in+1} - \sum_{i=D_1+1}^{D_1+D_2} \tau_{in},\ n=\overline{1,N-1} \quad (15)$$

первое звено тракта имеет минимальную скорость передачи пакетов, и общая задержка на маршруте составит

$$T(D_1+D_2,N) = \sum_{n=1}^{N} \tau_{1n} + \sum_{d=2}^{D_1+D_2} \tau_{dN}. \quad (16)$$

Данное выражение, исходя из ограничений (11), (13) и (15), обобщается зависимостью

$$T(D_1+D_2,N) = \sum_{n=1}^{N} \tau_{Mn} + \sum_{d=1,d\neq M}^{D_1+D_2} \tau_{dM},\ \tau_{Mn} = \max_{d=\overline{1,D_1+D_2}} \tau_{dn},\ \tau_{dM} = \max_{n=\overline{1,N}} \tau_{dn}. \quad (17)$$

При справедливости неравенства, обратного к (15), самое узкое место транспортного соединения соответствует последнему (D_1+D_2) звену, и искомое время передачи сообщения примет вид

$$T(D_1+D_2,N) = \sum_{n=1}^{N} \tau_{D_1+D_2n} + \sum_{d=1}^{D_1+D_2-1} \tau_{d1}, \quad (18)$$

которое с учетом указанных связей элементарных задержек пакетов на отдельных фазах пути τ_{dn} переписывается как

$$T(D_1+D_2,N) = \sum_{n=1}^{N} \tau_{Mn} + \sum_{d=1,d\neq M}^{D_1+D_2} \tau_{dm},\ \tau_{Mn} = \max_{d=\overline{1,D_1+D_2}} \tau_{dn},\ \tau_{dm} = \min_{n=\overline{1,N}} \tau_{dn}. \quad (19)$$

Для τ_{dn}, удовлетворяющих неравенствам (11), (14) и дополнительному условию (9), справедливому на этапах пути с первого по D_1-й, задержка сообщения в тракте определится соотношением (18) или его обобщением в виде (17). При справедливости (11), (14), условия (7) для первых D_1 этапов маршрута и неравенства (15) задержка сообщения определится соотношением (16), обобщающимся до (19). Если же удовлетворяется условие, обратное к неравенству (15), то время передачи сообщения находится из (18), или обобщающей зависимости (17). При выполнении (12) задержка в транспортном соединении аналогично вычисляется из соотношений (17) и (19).

4. Заключение

Из анализа рассмотренного ряда зависимостей показателя $T(D, N)$ от длины пути D, количества пакетов в сообщении N и времени передачи каждого пакета в отдельных звеньях маршрута τ_{dn} при различных связях между τ_{dn} можно сделать вывод о том, что хорошим приближением задержки сообщения с неоднородным по длинам набором пакетов в неоднородном тракте является выражение (6), мажорирующее задержку сообщения в транспортном соединении при произвольных соотношениях между τ_{dn}.

ЛИТЕРАТУРА

1. Воеводин В.В. Параллельные вычисления / В.В. Воеводин, Вл.В. Воеводин. СПб.: БХВ-Петербург. 2002.
2. Головкин Б.А. Вычислительные системы с большим числом процессоров. М.: Радио и связь. 1995.
3. Головкин Б.А. Анализ факторов, влияющих на скорость вычисления в параллельных и конвейерных системах // Автоматика и телемеханика, 1988, №5. С. 152–164.
4. Вишневский В.М. Теоретические основы проектирования компьютерных сетей. М.: Техносфера. 2003.
5. Shorten, R. H-TCP: TCP for high-speed and long-distance networks / R. Shorten, D. Leith // Proceedings of the PFLDnet, 2004.
6. Сущенко С.П. Параметрическая оптимизация сети пакетной коммутации // Автоматика и вычислительная техника, 1985, №2. С. 43–49.
7. Сущенко С.П. Анализ сквозной задержки сообщения в многозвенном виртуальном канале // Автоматика и вычислительная техника, 1989, №3. С. 52–64.
8. Сущенко С.П. Влияние длительности сквозного тайм-аута на задержку данных в виртуальном канале // Автоматика и вычислительная техника, 1991, №6. С. 36–40.
9. Сущенко С.П. Анализ влияния длительности сквозного тайм-аута на операционные характеристики виртуального канала // Автоматика и вычислительная техника, 1995, №4. С. 43–66

ON INITIAL WIDTH OF CONTENTION WINDOW INFLUENCE ON WIRELESS NETWORK STATION IEEE 802.11 CHARACTERISTICS

P. Mikheev[1], *S. Suschenko*[1]
[1] Tomsk State University
doka.patrick@gmail.com, ssp@inf.tsu.ru

A mathematical model of access method "carrier sense multiple access with collision avoidance" for two active stations was proposed. The effect of carrier capture and unimodal dependence of the operating characteristics from the initial width of the contention window was detected. Measures of preventing the effect of carrier capture, based on the modifications of the standard protocol were proposed.

1. Random multiple access method in 802.11 wireless networks

Let us analyze the wireless local area network (LAN) based on the IEEE 802.11 standard. The fundamental access method of such LANs is called DCF (Distributed Coordination Function) [1, 2] known as *carrier sense multiple access with collision avoidance* (CSMA/CA) [2, 3, 4].

This mechanism is based upon the fact that the transmitting station checks whether the carrier signal is present in the medium, and, before starting transmission of a data frame, expects release of the communication medium. IEEE 802.11 stations, in contrast to wired Ethernet, are not capable of detecting collisions in a communication medium [1, 5]. Due to this fact, detection of collisions and non-conflict transmissions of protocol-based data units is based on the time-outs mechanism and on the algorithm of positive decision feedback.

Let us analyze the cycle of a data frame transmission from the sending station to the recipient station. First and foremost, the sending station senses the medium to determine if another station is transmitting. Thereafter, at the end of the inter-frame interval, the random delaying algorithm is initiated to select a random backoff interval (the number of a slot in which the data transmission may be started). The slot number is selected with equal probability from the interval $[0, S_n - 1]$, where S_n is the size of the contention window measured in slot intervals t_c and determined by the relation $S_n = 2^{N_0+m}$, $m = n$ if $n \leq 10 - N_0$ and $m = 10 - N_0$ if $n \geq 10 - N_0$. Here $N_0 = \overline{1, 10}$ is the initial value predetermining the width of the contention window during the first attempt of a sender to transfer data, and $n \geq 0$ is the number of retransmission. The width of the contention window may not exceed the maximum value established by the standard. For all physical layers and methods of modulation, the IEEE 802.11 standard has established the maximum width of the contention window

equal to $S_{max} = 1024$ [2]. The number of a selected slot shall be assigned to the backoff interval counter t_o, after which the countdown of slot intervals begins. At the end of each slot interval, the backoff interval counter shall decrement as long as medium is idle. If the medium is determined to be busy at any time during a backoff slot, then the backoff procedure is suspended. Decrementing is resumed when the medium is idle again. Transmission shall commence when the backoff interval counter reaches zero ($t_o = 0$). When the transmission is completed, the sender waits for a acknowledgement during the time t_{out}, after which it is considered that a conflict has occurred, and stations having got into such conflict increase the n value by one, and the actions targeted at data transmission are repeated. The width of the contention window is doubled with each attempt of data frame transmission, until the maximum value is achieved; and the width of the contention window remains equal to S_{max} with each subsequent attempt of data frame transmission. After successful transmission, the window width obtains the initial value S_0.

Thus, the wireless access technology, due to lack of possibility to detect collisions in a communication medium, has three significant differences from the random access method implemented in the wired medium. Firstly, the wireless transmission method employs the mechanism of positive feedback (positive acknowledgements). Secondly, in contrast to the random access method, in wired networks the WiFi technology employs the random delay mechanism as early as during the first transmission. And at last, the wireless access protocol employs the mechanism of "suspension" of the delaying timer from the time of detection of the medium occupation until expiration of the random delay timer.

2. Mathematic modelling of 802.11 wireless LAN

Let us analyze the operation of a wireless local area network until the first error-free data frame transmission with obtained acknowledgement on successful delivery of data. Let us suppose that the wireless LAN contains K stations which are data sources. Consider that all the sources are independent and equal, and always have data frames for sending, and all interval spaces are expressed in slot intervals t_c. Let all the stations exchange frames of equal sizes. Then, according to the sequence of protocol actions, the elementary cycle of data frame transfer to the recipient will be determined by the size of the interframe space t_m, random delay period t_o, duration of "suspension" of the random delay timer t_z, time of data frame transmission t_k, and the value of timeout for expecting a positive acknowledgement t_{out}, which consists of a short interframe space plus the time of transmission of a positive acknowledgement [2, 4]. The average time of data frame transmission $T(K, N_0)$ consists of the weighted sum of average periods of waiting for failed transmissions and the time of successful transmission [6]:

$$T(K, N_0) = d + \sum_{N=0}^{\infty} \left[Nd + \sum_{n=0}^{N-1} t(n, K, N_0) + \tau(N, K, N_0) \right] f(N, K, N_0). \qquad (1)$$

Here $d = t_m + t_k + t_{out}$, $t(n, K, N_0)$ and $\tau(N, K, N_0)$ are the average conditional times until failed and successful N-th repeated attempts to send a data frame by a subscriber, and

$f(N, K, N_0)$ is the function of probability [7] of the duration of competition between subscribers for the medium, which is determined by the probability of successful data frame transmission on the N-th repeated step after $N-1$ failures [6]:

$$f(N, K, N_0) = P(N, K, N_0) \prod_{n=0}^{N-1} \pi(n, K, N_0).$$

Along with the average time of data frame transmission, one of the main indicators showing the efficiency of functioning the data transfer network is the throughput performance. In the case under analysis, we will look for an individual throughput performance, the standardized value of which shall be determined as a ratio between the time necessary for data frame transmission t_k and the average time of data frame transmission $T(K, N_0)$:

$$C(K, N_0) = \frac{t_k}{T(K, N_0)}. \tag{2}$$

3. The competition of two wireless stations

Let us analyze the competition of two wireless stations ($K = 2$) of a local area network. We denote the competing (conflicting) stations through A and B. Let us find the probability timing characteristics of the data transmission process executed by the A station. Let us denote via $p_n(i)$ the probability of selection of random backoff interval with a duration equal to i slot intervals on the n-th repeated transmission by the A station, and via $f_n(j)$ the probability of selection of random backoff interval with a duration equal to j slot intervals on the n-th repeated transmission by the B station. Then the conditional probability of a conflict on the n-th repeated transmission for the A station is determined by the relation

$$\pi(n, 2, N_0) = \begin{cases} \sum\limits_{i=0}^{S_0-1} p_0(i) \sum\limits_{j=0}^{i} f_0(j) L_{i-j}, & n = 0; \\ \sum\limits_{k=1}^{n} E_k(n) \left[\sum\limits_{i=0}^{S_k-1} p_n(i) \sum\limits_{j=0}^{i} f_k(j) L_{i-j} + \sum\limits_{i=S_k}^{S_n-1} p_n(i) \sum\limits_{j=0}^{S_k-1} f_k(j) L_{i-j} \right], & n \geq 1. \end{cases} \tag{3}$$

Here L_k represents recurrent probabilities of movement of the B station "bottom-up" from originally selected slot interval j to a conflict slot interval i selected by the A station (k is a difference between j-th and i-th slots), for many steps with successful transmissions:

$$L_k = \begin{cases} \sum\limits_{i=0}^{\infty} f_0^i(0) \sum\limits_{i=1}^{k} f_0(i) L_{k-i}, & k = \overline{1, S_0 - 1}, \; L_0 = 1; \\ \sum\limits_{i=0}^{\infty} f_0^i(0) \sum\limits_{i=1}^{S_0-1} f_0(i) L_{k-i}, & k = \overline{S_0, S_n - 1}. \end{cases}$$

In other words elements L_k include probabilities of all possible actions of the B station before collision with the A station, if the B station originally selected slot interval j and the A station selected slot interval i. From this point, it is not difficult to see

that, before the conflict with the A rival, the competing B station may carry out an unlimited number of successful transmissions in case of "fallout" of random delay having zero duration. Using the relations for the arithmetic-geometrical progression [8] for L_k with $k = \overline{1, S_0 - 1}$, we obtain the final relation:

$$L_k = \frac{S_0^{k-1}}{(S_0 - 1)^k}, \quad k = \overline{1, S_0 - 1}. \tag{4}$$

Inserting (4) into (3), we find the probability of a conflict on the first attempt of data frame transmission:

$$\pi(0, 2, N_0) = \frac{S_0 - 1}{S_0^2} \left[\left(\frac{S_0}{S_0 - 1} \right)^{S_0} - 1 \right].$$

The coefficients $E_k(n)$ in the relation (3) are the probabilities that on the n-th repeated transmission by the A station, the B station will be in the condition of the k-th repeated transmission:

$$E_1(1) = 1; \quad E_1(n) = \sum_{k=1}^{n-1} \frac{E_k(n-1)}{\pi(n-1, 2, N_0)} \left[\sum_{i=1}^{S_k - 1} p_{n-1}(i) \sum_{j=0}^{i-1} f_k(j) L_{i-j} + \right.$$
$$\left. + \sum_{i=S_k}^{S_{n-1} - 1} p_{n-1}(i) \sum_{j=0}^{S_k - 1} f_k(j) L_{i-j} \right], \quad n \geq 2;$$
$$E_k(n) = \frac{E_{k-1}(n-1) \sum\limits_{i=0}^{S_{k-1} - 1} p_{n-1}(i) f_{k-1}(i)}{\pi(n-1, 2, N_0)}, \quad n \geq 2, \ k = \overline{2, n}.$$

The average conditional times until failed and successful n-th attempt of data transmission $t(N, K, N_0)$ and $\tau(N, K, N_0)$ consist of the average duration of random delay $N_s(n)$ (average number of slots until the start of transmission) and the average number of suspensions caused by medium capture by the B station, $Z_t(n, N_0)$ in case of failure and $Z_\tau(n, N_0)$ in case of success, respectively:

$$t(n, 2, N_0) = N_s(n) + Z_t(n, N_0) d, \quad \tau(n, 2, N_0) = N_s(n) + Z_\tau(n, N_0) d.$$

Here $N_s(n) = \sum_{i=0}^{S_n - 1} i p_n(i) = (S_n - 1)/2$, and the average numbers of suspensions $Z_t(n, N_0)$ and $Z_\tau(n, N_0)$ look similar:

$$Z_t(n, N_0) = \begin{cases} \sum\limits_{i=1}^{S_0 - 1} p_0(i) \sum\limits_{j=0}^{i-1} f_0(j) M_{i-j}, & n = 0; \\ \sum\limits_{k=1}^{n} E_k(n) \left[\sum\limits_{i=1}^{S_k - 1} p_n(i) \sum\limits_{j=0}^{i-1} f_k(j) M_{i-j} + \sum\limits_{i=S_k}^{S_n - 1} p_n(i) \sum\limits_{j=0}^{S_k - 1} f_k(j) M_{i-j} \right], & n \geq 1; \end{cases} \tag{5}$$

$$Z_\tau(n, N_0) = \begin{cases} \sum\limits_{i=1}^{S_0 - 1} p_0(i) \sum\limits_{j=0}^{i-1} f_0(j) V_{i-j}, & n = 0; \\ \sum\limits_{k=1}^{n} E_k(n) \left[\sum\limits_{i=1}^{S_k - 1} p_n(i) \sum\limits_{j=0}^{i-1} f_k(j) V_{i-j} + \sum\limits_{i=S_k}^{S_n - 1} p_n(i) \sum\limits_{j=0}^{S_k - 1} f_k(j) V_{i-j} \right], & n \geq 1. \end{cases} \tag{6}$$

The elements M_k and V_k are indicators of the average number of suspensions of the delaying timer for the A station after selection of random delay with the duration i on the n-th repeated transmission upon selection of the j-th slot preceding to the i-th one by the competing B station (k is a difference between j-th and i-th slots):

$$M_k = \begin{cases} \sum\limits_{m=1}^{k} f_0(m) \sum\limits_{i=0}^{\infty} (i+1+M_{k-m}) f_0^i(0), & k = \overline{1, S_0 - 1},\ M_0 = 0; \\ \sum\limits_{m=1}^{S_0-1} f_0(m) \sum\limits_{i=0}^{\infty} (i+1+M_{k-m}) f_0^i(0), & k = \overline{S_0, S_n - 1}; \end{cases}$$

$$V_k = \begin{cases} \sum\limits_{i=0}^{\infty} (i+1) f_0^i(0) \sum\limits_{m=k+1}^{S_0-1} f_0(m) + \sum\limits_{m=1}^{k-1} f_0(m) \sum\limits_{i=0}^{\infty} (i+1+V_{k-m}) f_0^i(0), & k = \overline{1, S_0 - 1}; \\ \sum\limits_{m=1}^{S_0-1} f_0(m) \sum\limits_{i=0}^{\infty} (i+1+V_{k-m}) f_0^i(0), & k = \overline{S_0, S_n - 1}. \end{cases}$$

After inserting here the probabilities of fallout of delay duration $f_0(m)$, we obtain the following relations:

$$M_k = \begin{cases} \dfrac{S_0}{S_0 - 1} \left[\left(\dfrac{S_0}{S_0 - 1} \right)^k - 1 \right], & k = \overline{1, S_0 - 1}; \\ \dfrac{S_0}{S_0 - 1} + \dfrac{\sum_{m=1}^{S_0-1} M_{k-m}}{S_0 - 1}, & k = \overline{S_0, S_n - 1}. \end{cases}$$

$$V_k = \begin{cases} \dfrac{S_0 - 2}{S_0 - 1} \left(\dfrac{S_0}{S_0 - 1} \right)^k, & k = \overline{1, S_0 - 1}; \\ \dfrac{S_0}{S_0 - 1} + \dfrac{\sum_{m=1}^{S_0-1} V_{k-m}}{S_0 - 1}, & k = \overline{S_0, S_n - 1}. \end{cases}$$

The indicator of the general throughput performance can be found by analogy with individual operational speed (2), therewith the numerator of such relation should be adjusted not only for the package successfully transferred by the A station, but also for the average number of packages transferred by the B station for the concerned period:

$$C_{\mathrm{g}}(2, N_0) = \frac{(G(N_0) + 1) t_k}{T(2, N_0)},$$

where $G(N_0)$ will be determined by the weighted amount of the average number of suspensions of the delaying timer of the A station in expectation of failed and successful transmissions, which are determined by the relations (5) and (6):

$$G(N_0) = \sum_{N=0}^{\infty} \left[\sum_{n=0}^{N-1} Z_t(n, N_0) + Z_\tau(N, N_0) \right] f(N, 2, N_0).$$

4. Numerical results

The numeric research into the average time of data frame transmission by the A station shows that the function (1) has a strongly manifested minimum at the coordinate N_0 (see Fig. 1) determining the initial size of the competition window and, subsequently, the degree of scattering of stations by durations of delays before the start

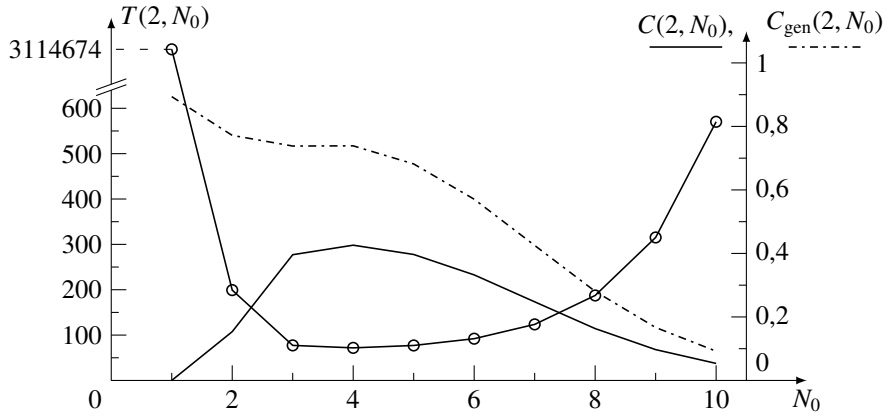


Figure 1: Average time of data frame transmission, and individual and general throughput performances

of the competition procedure. For two competing stations, the minimum is reached at $N_0 = 4$. It is obvious that the value N_0 minimizing the average time of data frame transmission maximizes the individual throughput (see Fig. 1). Moreover, as early as at the stage of formalization of the task, the probability of capture of the communication medium by one of the subscribers mentioned in [9, 10] has become obvious. This effect manifests itself especially strongly with small values N_0. The effect of capturing the communication medium causes discrimination-related individual indicators against a good level of the general throughput performance of the network (see Fig. 1).

As early as at the first attempt of competition between two stations, capture of the communication medium becomes possible (e.g. by the B station), and its probability will be determined by the probabilities that for one of the stations (B) the delay duration will turn out to be shorter than the duration of delay of the other station (A); then the "succeeded" station (B) will have fallout of zero duration, which will alternate with shorter delays than the residual value of the station's A delaying timer:

$$P_z(0, 2, N_0) = \sum_{i=1}^{S_0-1} p_0(i) L_{i-1} \sum_{k=1}^{\infty} f_0^k(0) = \frac{1}{S_0^2} \left(\frac{S_0}{S_0 - 1} \right)^{S_0-1}.$$

From this point, it is not difficult to see that the probability of capture is considerably determined by the initial width of the contention window S_0 (see Fig. 2). After several conflicts, the possibility of capture for the "succeeded" station becomes yet more probable.

The main reason for the effect of capturing the communication medium is the protocol action — "suspension of delay", because this results in a fact that after a non-

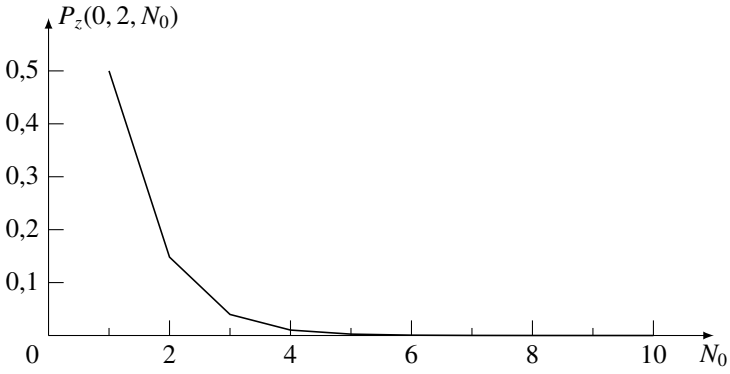


Figure 2: Probability of the medium capture by one of the stations

conflict transmission the station may capture the communication medium for an infinitely long time, getting into the delay interval from 0 to the residual value of the delay of other stations.

Another reason for an increase in the probability of capturing the communication medium by one of the subscribers after several conflicts, consists in various sizes of the contention window for stations withdrawn from the conflict and stations continuing resolution of the conflict in the condition of waiting for expiration of delay time and suspension periods. After a positive resolution of the conflict by one of the stations (or by several stations), the size of its contention window is reduced in multiples down to the initial value $S_0 < S_n$, which gives this station a priority right in subsequent competition for the medium with "conflicting" stations, because the shorter duration of an occasional delay for such station has a significantly higher probability as compared with the similar operational indicator of the "conflicting" station.

It is obvious that to reduce the probability of the effect of medium capturing for an infinitely long time, it is possible to offer, on one hand, to fix the size of the contention window for the first and all subsequent transmissions, and on the other hand – the duration of random delay t_o should be selected within the interval from 1 to $2^{N_0} - 1$ of slot periods t_c, thus excluding the delay of the zero size. Therewith, medium capturing by one station will never exceed $2^{N_0} - 2$ successful transmissions until the subsequent conflict or its resolution.

5. Conclusion

The performed analysis is targeted at studying the method of carrier sense multiple access with collision avoidance. Analytic correlations have been obtained for probability timing characteristics of the competition process between two stations. The "medium capture effect" and the extreme dependence of operational parameters on the initial contention window size have been revealed.

It has been suggested to change the parameters of the protocol procedure of competition, ensuring prevention of the capture effect by saving high values of individual and integral indices of operational speed.

It has been shown that the optimal initial width of the contention window (S_0) is determined by the active size of the network (the number of competing stations), and it ensures almost uniform distribution of a jointly used time resource of the medium between competing subscribers.

REFERENCES

1. *Tanenbaum A.S., Wetherall D.J.* Computer Networks, Fifth Edition. Boston: Prentice Hall, 2010. 960 p.
2. IEEE Std 802.11 — 2007, Revision of IEEE Std 802.11 — 1999. IEEE Std 802.11 — 2007, 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. IEEE Computer Society, 2007. 1184 p.
3. *Geier J.* Wireless Networks first-step. Indianapolis: Cisco Press, 2005. 192 p.
4. *Roshan P., Leary J.* 802.11 Wireless LAN Fundamentals. Cisco Press, 2004. 312 p.
5. *Bianchi G.* Performance Analysis of the IEEE 802.11 Distributed Coordination Function // IEEE Journal on Selected Areas in Communications. 2000. Issue 18(3). P. 535–547.
6. *Kustov N.T., Suschenko S.P.* Capacity of the Random Multiple Access Method // Autom. Remote Control. 2001. V. 62. 1. P. 76–85.
7. *Hastings N.A.J., Peacock J.B.* Statistical Distributions in Scientific Work Series: A Handbook for Students and Practitioners. London: Butterworth, 1975. 130 p.
8. *Prudnikov A.P., Brychkov Yu.A., Marichev O.I.* Integrals and Series: Special functions. Amsterdam: Overseas Publishers Association, 1986. 751 p.
9. *Bononi L., Conti M., Donatiello L.* Design and Performance Evaluation of Distributed Contention Control (DCC) Mechanism for IEEE 802.11 Wireless Local Area Network // J. Parallel Distrib. Comput. 2000. V 60. 4.
10. *Vishnevsky V.M., Lyakhov A.I.,* IEEE 802.11 Wireless LAN: Saturation Throughput Analysis with Seizing Effect Consideration // Cluster Computing. 2002. 5. P. 133–144.

PROBABILITY CHARACTERISTIC COMPUTATION ALGORITHM OF UPSTREAM TRAFFIC IN PASSIVE OPTICAL NETWORK

G. Basharin, N. Rusina

Peoples' Friendship University of Russia, Moscow, Russia,
gbasharin@sci.pfu.edu.ru, rusina_nadezda@inbox.ru

Abstract

Nowadays, access network evaluation is being conducted in both directions: high bit rate access development for providing high quality of service and decrease length of cooper wiring in local line networks. It's paid special attention to the networks, which are based on optical and optoelectronic components. Passive optical network (PON) is an all optical network, which is based on passive optical components only and excludes the conversion of electrical signal into optical form and *vice versa*. This paper is concerned with the algorithm for calculation of call blocking probability in upstream traffic multiservice model for PON considering the functioning process of optical network units and the principle of spectral channel dynamic distribution. These results are used in the blocking probability analysis of the model.

Keywords: Passive Optical Network (PON), Optical Line Terminal (OLT), Optical Network Unit (ONU), upstream, Wavelength Division Multiplexing (WDM), Time Division Multiple Access (TDMA), blocking probability.

1. Introduction

PON is an optical access network, which supports transmitting various classes of the network traffic between optical line terminal (OLT) and optical network units (ONU) through a passive optical multiplexer/demultiplexor, which combines/ splits signals in various spectral channels through a fiber [1, § 1.5], [2]-[6].

According to the time division multiple access (TDMA) technology [4, 5], ONU may be in the ON-state, in other words, transmits data through its earlier assigned frame, or the ONU may be in the OFF-state, in other words, it is in period, when no data transmission occurs.

According to the wavelength division multiplexing (WDM) technology [1, § 7.5, 7.6], [2]-[6], PON employs a set of spectral channels W to transmit upstream traffic from ONU to OLT.

Each ONU uses a reconfigurable laser and can transmit upstream traffic in the selected range of wavelengths. Thus, the dynamically allocating wavelengths mechanism allows increasing the system capacity, providing highly scalable. As well as WDM-TDMA PON provides new users and required quality of service level through existing frequency plan.

Let the WDM-TDMA PON with dynamic distribution of W spectral channels, which we propose is called the Network I.

The Network I solves the problem of the limited number $W \leq L$ of spectral channels distribution between finite number L ONU. If there is no free spectral channel in OLT at the ONU activation moment data transmitting is blocked in the ONU time domain.

2. Optical network units confunction in the Network I fragment

Examine ONU confunction model in the Network I fragment, the parameters of which are presented in Table 1.

Parameter	Description
L	number of optical network units.
W, $W \leq L$	number of upstream traffic spectral channels.
λ_l, $l = \overline{1, L}$	ONU_l transition rate from the passive (OFF) state to the active (ON) state.
μ_l, $l = \overline{1, L}$	ONU_l transition rate from ON-state to OFF-state.

Table 1: The model Parameters.

Let us denote

$g(L, \mathrm{w})$ - nonnormalized probability that the first L ONU use w spectral channels;

$g_{-l}(\mathrm{w})$ - nonnormalized probability that optical network units use w spectral channels, and ONU_l is in the OFF-state.

Here and elsewhere, α_l is a probability of ONU_l, $l = \overline{1, L}$, being in an ON-state or having the opportunity to go into it at its activation moment and is calculated by the formulas [7, 8]

$$\alpha_l = 1 - G^{-1} g_{-l}(W), G = \sum_{\mathrm{w}=0}^{W} g(L, \mathrm{w}). \tag{1}$$

$$g_{-l}(\mathrm{w}) = \begin{cases} 1, & \mathrm{w} = 0, \\ g(L, \mathrm{w}) - \rho_l g_{-l}(\mathrm{w} - 1), & \mathrm{w} = \overline{1, W}, \end{cases} \tag{2}$$

$$g(l, \mathrm{w}) = \begin{cases} 0, & l = 0, \quad \mathrm{w} = \overline{1, W}, \\ 1, & l = \overline{0, L}, \quad \mathrm{w} = 0, \\ g(l-1, \mathrm{w}) + \rho_l g(l-1, \mathrm{w}-1), & l = \overline{1, L}, \quad \mathrm{w} = \overline{1, W}, \end{cases} \tag{3}$$

where

$\rho_l := \frac{\lambda_l}{\mu_l}$, $l = \overline{1, L}$,

α_l, $l = \overline{1, L}$, is one of the key performance factors of the Network I.

3. Upstream traffic transmission model of the Network I

Let us consider the upstream traffic transmitting process in the Network I fragment involving L ONU [9]. Every ONU_l has a finite memory buffer of size R_l time-slots, $0 < R_l < \infty$, $l = \overline{1, L}$. The analyzed queue system supports K types of service classes. k-calls arrive at an ONU_l according to a Poisson process with mean arrival rate $\lambda_{l,k}$, $0 < \lambda_{l,k} < \infty$, $l = \overline{1, L}$, $k = \overline{1, K}$. Arrival streams are independent in total for each ONU_l. A k-call is realized by allocating a specific number of time slots, b_k, $0 < b_k \leq \min\limits_{l=\overline{1,L}} R_l$, $k = \overline{1, K}$. The k-call holds b_k time slots in the ONU_l buffer until it is serviced, then it unbuffers immediately after the completion of service, thus idling the spectral channel. A service policy for each ONU_l is *first come first serve* (FCFS).

A new k-call, $k = \overline{1, K}$, will be rejected by the ONU_l, $l = \overline{1, L}$, if there are no more than $R_l - b_k$ free time slots in the buffer. The k-call leaves the system and has no effect on the stream arrival rate.

μ_k, $0 < \mu_k < \infty$, $k = \overline{1, K}$, is the mean service time of a k-call in ONU_l, which is exponentially distributed. This model of the upstream traffic transmitting will encode $\begin{array}{c|c|c|c} \mathbf{M} & \boldsymbol{M} & \boldsymbol{1} & d_1 \\ \boldsymbol{\Lambda}, \boldsymbol{b} & \boldsymbol{\mu} & & \boldsymbol{R} \end{array}$ [1, 10].

However, the model described above does not account for the fact that there may not be free spectral channels in OLT at the ONU_l activation moment. This fact leads to blocking the data transmitting in the ONU_l time domain. Then the k-call service rate, taking into consideration the fact and the results of the previous section, will be

$$\alpha_l \mu_k, l = \overline{1, L}, k = \overline{1, K}. \tag{4}$$

The queue system scheme of the upstream traffic transmitting from ONU to OLT is shown in Figure 1.

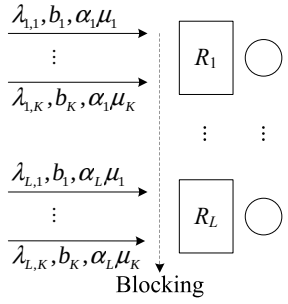


Figure 1: The queue system scheme.

The queue system in Figure 1 is described using the following parameters. The state matrix of the system

$$\mathbf{M} := (m_{l,k})_{l=\overline{1,L}, k=\overline{1,K}}, m_{l,k} \in \left\{0, 1, \ldots, \left\lfloor \frac{R_l}{b_k} \right\rfloor\right\}. \tag{5}$$

The state space of the system

$$\mathrm{S} := \{\mathbf{M} | 0 \leq \sum_{k=1}^{K} b_k m_{l,k} \leq R_l \ , l = \overline{1, L}\}, k = \overline{1, K}, l = \overline{1, L}. \tag{6}$$

The state subspace of k-calls received by ONU_l

$$\mathrm{S}_{l,k} := \{\mathbf{M} \in \mathrm{S} | \sum_{k=1}^{K} b_k m_{l,k} \leq R_l - b_k\}, k = \overline{1, K}, l = \overline{1, L}. \tag{7}$$

The state subspace of k-calls blocked by ONU_l

$$\bar{\mathrm{S}}_{l,k} = \mathrm{S} \backslash \mathrm{S}_{l,k} = \{\mathbf{M} \in \mathrm{S} | \sum_{k=1}^{K} b_k m_{l,k} > R_l - b_k\}, k = \overline{1, K}, l = \overline{1, L}. \tag{8}$$

4. Formula derivation for the blocking probability calculation

The queue system functioning is described by a two-dimensional stationary Markov process $\mathbf{Y}(t) := (Y_{l,k}(t))_{l=\overline{1,L}, k=\overline{1,K}}$, where $Y_{l,k}(t)$ is the number of k-calls in an ONU_l, at some instant $t > 0$.

$\mathbf{E}_{l,k}$ is a matrix of $L \times K$, where at the intersection of the l^{th} row and k^{th} column is "1" and the rest matrix elements are "0". The transition rate diagram for any ONU_l is shown in the Figure 2.

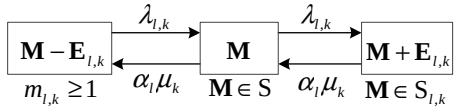


Figure 2: The state transition scheme for ONU_l, $l = \overline{1, L}$, and k-call, $k = \overline{1, K}$.

Theorem 1. *The stationary Markov process* $\mathbf{Y}(t)$, $t > 0$, *has got a stationary probability distribution,which does not depend on the starting probability distribution and is multiplicative*

$$p(\mathbf{M}) = G^{-1} \prod_{l=1}^{L} \frac{1}{\alpha_l^{m_{l,\bullet}}} \prod_{k=1}^{K} \rho_{l,k}^{m_{l,k}}, \tag{9}$$

$$G = \frac{1}{p(\mathbf{0})} = \sum_{\mathbf{M} \in \mathrm{S}} \prod_{l=1}^{L} \frac{1}{\alpha_l^{m_{l,\bullet}}} \prod_{k=1}^{K} \rho_{l,k}^{m_{l,k}}, \tag{10}$$

where
$\mathbf{M} \in \mathrm{S}$, $\rho_{l,k} := \frac{\lambda_{l,k}}{\mu_k}$, $k = \overline{1, K}$, $m_{l,\bullet} := \sum_{k=1}^{K} m_{l,k}$, $l = \overline{1, L}$.

The Theorem 1 proof is similar to that presented in [1, § 2.4].

Then, the blocking probability of k-calls for ONU$_l$

$$\pi_{l,k} = \sum_{\mathbf{M} \in \bar{\mathbf{S}}_{l,k}} p(\mathbf{M}), l = \overline{1,L}, k = \overline{1,K}. \tag{11}$$

5. Computation algorithm for the blocking probability

Based on the model described in the section 4, we need to solve an allocation problem of buffer limited capacity R_l between finite number of K call types for each ONU$_l$, $l = \overline{1,L}$. We apply a convolution Buzen type algorithm [11], which is widely used at the teletraffic theory nowadays [1, §4.5], [12].

It is necessary to calculate the norming quantity to calculate the blocking probability of k-calls for ONU$_l$.

Let us denote

$g_l(k,r)$ - nonnormalized probability that the first $(1,\ldots,k)$ call types use all of the r time slots of the ONU$_l$ buffer.

Theorem 2. *Norming quantity* G (10) *calculated by the formulas*

$$G = \prod_{l=1}^{L} \sum_{r=0}^{R_l} g_l(K,r), \tag{12}$$

$$g_l(k,r) = \begin{cases} 0, & k = \overline{1,K}, \quad r < 0, \\ 0, & k = 0, \quad r = \overline{1,R_l}, \\ 1, & k = \overline{1,K}, \quad r = 0, \\ g_l(k-1,r) + \frac{\rho_{l,k}}{\alpha_l} g_l(k, r-b_k), & k = \overline{1,K}, \quad r = \overline{1,R_l}, \end{cases} \tag{13}$$

where
$l = \overline{1,L}$, $\rho_{l,k} := \frac{\lambda_{l,k}}{\mu_k}$, $k = \overline{1,K}$.

Then, the blocking probability of k-calls for ONU$_l$ in the Network I will be

$$\pi_k = \frac{1}{G} \sum_{r=R-b_k+1}^{R} g(K,r), k = \overline{1,K}. \tag{14}$$

6. Numerical analysis example

Let us consider the dependence of the k-calls blocking probability for ONU$_l$, $l = \overline{1,L}$, and the number of spectral channel $W = \overline{0,16}$ allocated for the upstream traffic transmitting process from L ONU to OLT in the Network I. The buffer capacity is fixed, $R_l = 28$. The model Parameters values are represented in the Table 2.

The chart of π_{11} and π_{12} from the number of spectral channels W is shown in the Figure 3.

Parameter values α^T are calculated according to formulas (1)-(3) based on the ONU co-operation model in the Network I. This model Parameters values are represented in the Table 3.

Param.	Value	Param.	Value	Param.	Value
K	2	$R_l, l = \overline{1, L}$	28	ρ_{lk}, $l = \overline{2, L}$, $k = \overline{1, K}$	1
L	16	$\boldsymbol{b}^T$	$\begin{pmatrix} 1 & 2 \end{pmatrix}$		
W	$\overline{0, 16}$	ρ_{11}, ρ_{12}	0.5, 0.25		

Table 2: The model Parameters.

Param.	Value	Param.	Value	Param.	Value	Param.	Value
L	16	ρ_4, ρ_6	1.7	ρ_9	3	ρ_{13}	2.7
ρ_1	1.2	ρ_5	1.6	ρ_{10}	2.1	ρ_{14}	1.9
ρ_2	1.3	ρ_7	2	ρ_{11}	2.4	ρ_{15}	1.8
ρ_3	1.1	ρ_8	2.5	ρ_{12}	2.8	ρ_{16}	2.2

Table 3: The model Parameters.

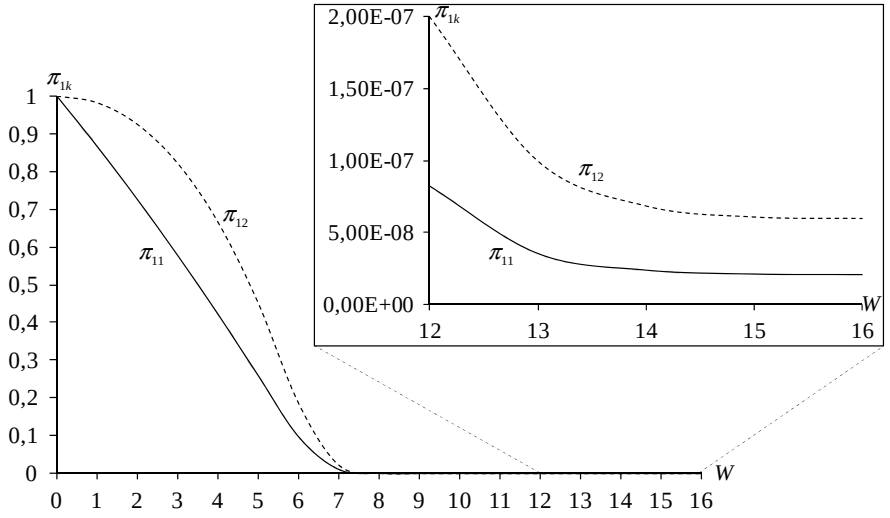


Figure 3: The chart of blocking probabilities π_{11} and π_{12} from W.

The k-calls blocking probability for ONU_l, $l = \overline{1, L}$, is reduced with increasing the spectral channels number allocated in the upstream traffic transmitting model from L ONU to OLT in the Network I.

7. Conclusion

In this paper, the mathematical model of the upstream multiservice traffic transmitting from L ONU to OLT in the Network I is presented. The algorithm for calculating of the blocking probability is proposed. In addition, there is the numerical analysis example in which the role of W is shown to select the optimal functioning behavior of the ONU.

The authors recommend the use of the concepts described in this paper for the algorithm development to calculate of the blocking probability for the upstream priority traffic transmitting model in the Network I.

REFERENCES

1. Basharin, G.P. Lectures on mathematical teletraffic theory. The 3^{rd} publication. PFUR, Moscow, 2009. 342 p. (The 1^{st} pub. – 2004. – 186 p.; The 2^{nd} pub. – 2007. – 268 p.)
2. Listvin V.N., Treschikov V.N. DWDM-systems. The 2^{nd} publication. TECHNOSPHERA, Moscow, 2015. 296 p.
3. Efimushkin V.A., Savangukov I.M. Resource allocation in optical transport networks: Textbook. CRIC, Moscow, 2010. 50 p.
4. Ramaswami R., Sivarajan K.N., Sasaki G.H. Optical Networks A practical Perspective, Third Edition, 2010. 893 p.
5. Mukherjee, B. Optical WDM Networks. Springer, 2006. 973 p.
6. Siva Ram Murthy C., Gurusamy M. WDM Optical Networks: Concepts, Design and Algorithms. Prentice Hall PTR, 2002. 430 p.
7. G.P. Basharin, Yu.V. Gaidamaka, N.V. Rusina. Probability characteristic computation algorithm of ONUs functioning in PON // Vestnik of PFUR, "Mathematics. Informatics. Physics". 2015. No 2. In the press.
8. Naymov, V.A., Samyilov, K.Y., Yarkina N.V. Teletraffic theory of multiservice networks. Monography. PFUR, Moscow, 2008. 191 p.
9. Basharin G., Rusina N. Multirate Loss Model for Optical Network Unit in Passive Optical Networks // Distributed Computer and Communication Networks: Communications in Computer and Information Networks. Springer, Cham. 2014. Pp. 219-228.
10. Basharin G. P. Maintenance of two streams with relative priority for a fully accessible system with a limited waiting space // Proceedings of the AS of the USSR, Technical Cybernetics. 1967. No 2. Pp. 72-86.
11. Busen, J. P. Computational algorithms for closed queueing networks with exponential servers // Communications of the ACM. 1973. V. 16. Pp 527-531.
12. G. P. Basharin, Yu. V. Gaidamaka, and K. E. Samouylov Mathematical Theory of Teletraffic and Its Application to the Analysis of Multiservice Communication of Next Generation Networks // Automatic Control and Computer Sciences. 2013. Vol. 47, No. 2. Pp. 62-69.

SYNTHESIS OF NOISE-LIKE SIGNAL BASED ON ATEB-FUNCTIONS

I. Dronyuk, M. Nazarkevych, O. Fedevych
Lviv Polytechnic National University, Lviv, Ukraine

Abstract

The mathematical model of the noise signal based on Ateb-functions was proposed. Connectivity of Ateb-functions with the curve of superellipse was shown. Algorithm for synthesis Gaussian white noise signal based on the periodical Ateb-functions is created. Investigation for statistical characteristics of generated signals using MATLAB 7.0 environment is presented. Application the proposed method of noise-like signals synthesis to protection information during data transfer in computer networks, in particular networks with CDMA technology is discussed.

СИНТЕЗ ШУМОПОДОБНЫХ СИГНАЛОВ С ПОМОЩЬЮ АТЕВ-ФУНКЦИЙ

И. Дронюк, М. Назаркевич, О. Федевич
Национальный университет "Львовская политехника", г. Львов, Украина
ivanna.droniuk, maria.nazarkevych, olha.fedevych@gmail.com

Аннотация

Предложена математическая модель шумового сигнала на основе Ateb-функций. Показана связь Ateb-функций с кривой суперэллипса. Создан алгоритм синтеза Гауссовского белого шума на основе периодических Ateb-функций. Исследованы статистические характеристики генерируемых сигналов с помощью среды MATLAB 7.0. Обсуждается применение предложенного метода синтеза шумоподобных сигналов для защиты информации при передаче данных в компьютерных сетях, в частности сети с CDMA технологией.

Ключевые слова: *Шумоподобный сигнал, Ateb-функции, Гауссовский белый шум, суперэллипс*

1. Введение

Для защиты данных в компьютерных сетях часто используются некоторые шумовые сигналы [1]. До недавнего времени эти исследования в большинстве принадлежали к закрытой тематике. С развитием информационных технологий шумовые сигналы начали использоваться во многих

системах конфиденциальной связи. Это особенно верно для систем беспроводной связи, в том числе сетей, которые используют технологию кодового разделения каналов (CDMA) [2].Технология CDMA основана на технологии передачи SST (DH-SS прямой последовательности распространения спектра), когда информация, как бы размазывается по широком диапазоне частот. Гармонические колебания часто играют роль информационного носителя. В книге [2] автор описывает классический подход для описания примеси, белого и Гауссовой моделей шума. Шум присутствует во всех коммуникационных системах, и это было препятствием для обнаружения сигналов и должно быть устранено фильтрами. С дальнейшим развитием телекоммуникационных технологий, возникли методы передачи информации, основанные на шумовых сигналах [3]. При строительстве современных конфиденциальных систем связи, шумовые сигналы преобразуются из препятствий на пути основных носителей сигнала в непосредственно сами носители сигналов.

В течение последнего десятилетия началось успешное использование шумовых сигналов в качестве носителя основной информации, передаваемой по каналам связи. Шумовые сигналы имеют ряд преимуществ с точки зрения информационной безопасности как в коммуникационных кабельных каналах, так и в радиоканалах. Передача данных в шумовом сигнале может быть хорошей альтернативой для методов криптографической защиты.

Известно [1], что колебание несущей, формируется на основе шумового сигнала, позволяет эффективно распознавать сигналы по форме. У шумовых сигналов есть и другое преимущество - оно дает возможность обеспечить передающую защиту на уровне физического канала, что особенно важно при создании защищенных систем с множественным доступом связи. Шумоподобные сигналы - это тип шумовые сигналы, где роль носителя может быть сыграна сигнальными конструкциями, которые осуществляются на основе гармонических колебаний.

Колебания, которые возникают в нелинейных системах с несколькими степенями свободы обобщают гармонические колебания и описаны в соответствующими дифференциальными уравнениями. Математически эти колебания могут быть описаны с помощью Ateb-функций. Это идея этой статьи: применить метод построения шумоподобных сигналов на основе гармонических колебаний для того, чтобы обобщить колебания, которые описываются Ateb-функциями, а также для создания шума, как колебания на основе Ateb -функций. Так как шумоподобные сигналы используются в технологии CDMA, в частности, предлагаемый подход является актуальным и может иметь широкое практическое применение для защиты передачи информации в компьютерных сетях. Предложенный метод может быть использован для синтеза шумовых сигналов в аппаратных и программных областях. Целью данного исследования является разработка математической модели, основанной на Ateb-функциях для задачи синтеза дискретно-

го шумового сигнала с заданными спектральными и автокорреляционными свойствами.

2. Построение математической модели

Введем необходимые обозначения и формулы, связанные с *Ateb*-функциями, в дальнейшем необходимые для создания шумового сигнала. Подробности относительно свойств *Ateb*-функций можно найти в [4]. В этой статье мы введем обозначение $sa(n, m, \omega)$, для Ateb-синуса и $ca(m, n, \omega)$, для *Ateb*-косинуса. Как известно, [5] что Ateb-функции удовлетворяют тождеству

$$ca^{m+1}(m, n, \omega) + sa^{n+1}(n, m, \omega) = 1 \tag{1}$$

Пусть $\Gamma(x)$ обозначает Гамма-функцию. Период $\Pi(m, n)$ *Ateb*-функций рассчитывается по следующей формуле

$$\Pi(m, n) = \frac{\Gamma\left(\frac{1}{n+1}\right)\,\Gamma\left(\frac{1}{m+1}\right)}{\Gamma\left(\frac{1}{n+1} + \frac{1}{m+1}\right)} \tag{2}$$

в случае, когда $n = m = 1$ мы получаем $ca(1, 1, \omega) = \cos(\omega), sa(1, 1, \omega) = sin(\omega)$, что удовлетворяет уравнению (1), а именно $\cos(\omega)^2 + \sin(\omega)^2 = 1$. Таким образом, *Ateb*-функции есть обобщением обычных тригонометрических функций. Теперь покажем связь между рассматриваемыми *Ateb*-функциями и плоской алгебраической кривой Ламэ, которая также носит название обобщенного суперэллипса. Тождество (1), которому удовлетворяют периодические *Ateb*-функции, может быть представлено графически в виде суперэллипса. Рассмотрим формулу обобщенного суперэллипса заданную в [6]

$$\left|\frac{x}{a}\right|^p + \left|\frac{y}{b}\right|^q = 1, \qquad \text{где } p, q > 0. \tag{3}$$

Для простоты, будем считать, что $a = b = 1$ и p, q связаны с параметрами *Ateb*-функций, которые следуют из уравнений $p = m+1; q = n+1$. Замена $x = ca(n, m, \omega), y = sa(m, n, \omega)$ в уравнении (3) обобщенного суперэллипса, преобразовывает (3) в тождество (1). Таким образом, это показывает, что основное тождество *Ateb*-функций (1) есть представлением уравнения суперэллипса (3). Теперь, обратимся к описанию способа формирования шумового сигнала на основе *Ateb*-функций. Принимая во внимание характерные черты реального шумового сигнала, реальный шумовой сигнал $s(t)$ - это набор одновременно существующих электрических колебаний частоты, фазы и амплитуды, которые являются случайными. Спектр шумовых сигналов включает в себя широкий диапазон частот. Если этот спектр равномерный на всех частотах от 0 до ∞, тогда такой шум называют белым. На практике, этот шум не может быть получен, но для каких-либо аппаратных средств, чьи полосы пропускания во много раз меньше, чем спектр

шумового сигнала, шум можно считать белым. Мощность шумового сигнала, который используется, определяется шириной полосы устройства на вход которого он приходит. Если излучается управляемый шумовой сигнал с длительностью T, а прием осуществляется с помощью согласованного фильтра или корреляционной схемы, тогда на выходе коррелятора при совпадении по времени принятого и исходного сигналов ($\tau = 0$) выделится энергия

$$E = \int_0^T s(t)^2 dt \tag{4}$$

При условиях, что эта энергия также случайная величина со средним значением E_0 и дисперсией , флуктуации энергии колебаний носителя можно рассматривать как интерференцию, которая искажает сигнал. Если шум ограничен полосой Δf и имеет большую базу , то относительное стандартное отклонение значения равно [7]:

$$\frac{\sigma^2}{E_0} \approx \frac{1}{\Delta f T} \tag{5}$$

Для того чтобы минимизировать искажение информации, передаваемой шумом, целесообразно использовать сигнал с большой базой. Затем, на основе (4), результаты измерения энергии шума станут средними за время, которое в несколько раз превышает время корреляции τ исследуемого сигнала. Белый шум является идеальным шумовым сигналом, который имеет бесконечный спектр и корреляционную функцию в форме дельта-функции. На практике, белый шум не может быть получен, потому что необходимо оборудование с неограниченной полосой пропускания для его генерирования и обработки. По этой причине в реальных системах используется шум с ограниченной полосой Δf. Работа [6] рассматривает метод построения шумоподобных сигналов на основе гармоничных колебаний. Как *Ateb*-функции могут генерировать широкий спектр флуктуаций, что являются обобщениями гармоник и являются периодическими, настоящая работа обобщает предложенный в [4] метод, основанный на *Ateb*-функциях. Основное преимущество предлагаемого подхода заключается в переменном периоде Ateb-функций, определенном по формуле (2). В зависимости от параметров n и m Ateb-функций, которые позволяют выбрать такой период времени, который наилучшим образом соответствует задаче сокрытия информации на основе сгенерированного шумоподобного сигнала. Рис. 1 показывает зависимость периода $\Pi(m, n)$ от параметров n и m, и изменением параметров последнего в диапазоне от -5 до 5. Расчеты проводились с использованием уравнения (2) в вычислительной среде Wolfram Mathematika 7.0. Рис. 1 показывает, что размер периода в основном увеличивается с увеличением аргументов, однако он также имеет ряд локальных экстремумов и некоторые асимптоты для отрицательных значений аргументов. Это позволяет выбрать требуемое значение периода в

зависимости от ширины канала. При построении шумоподобного сигнала с целью сделать его ближе всего к белому шуму, ряд следующих требований должен быть принят во внимание: сигнал должен быть широкополосным, т.е. базовый сигнал $B = FT >> 1$, где T - длительность сигнала; F - ширина полосы пропускания частот сигнала.

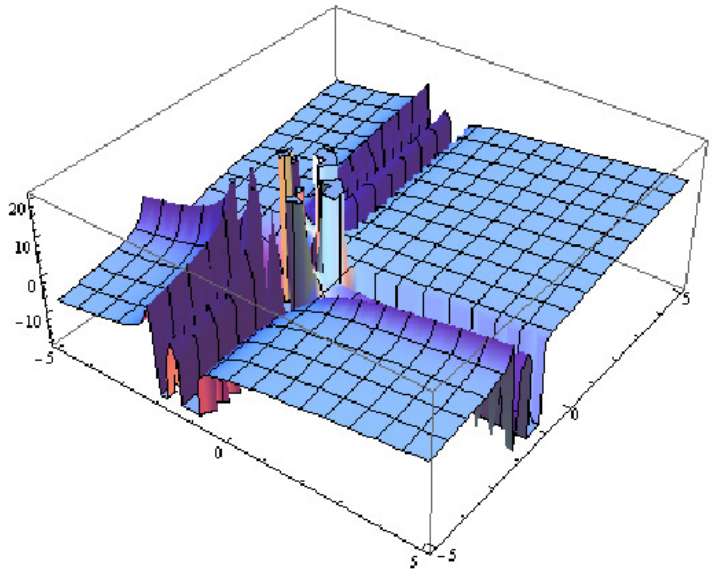


Рис. 1: Зависимость периода $\Pi(m,n)$ от параметров n и m, которые принадлежат интервалу $[-5;5]$

3. Алгоритм для синтеза сигнала белого шума на основе периодических *Ateb*-функций

Затем шумоподобный сигнал может быть записан в виде:

$$u(m,n,t) = U(t)v(t)sa(m,n,2\Pi(n,m)s_0t + \phi(t)) \tag{6}$$

где $sa(m;n;t)$ - *Ateb*-синус с параметрами m,n, $U(t)$ - амплитуда и $\phi(t)$ - фаза, взаимно независимых случайных функций, которые изменяются медленно в сравнении с *Ateb*-синусом;s_0 - центральная частота спектра шума; $v(t)$ - нормализированная псевдослучайная последовательность чисел с заданным законом распределения. Рассмотрим в деталях преимущества уравнения (6), используя свойства *Ateb*-функций. Можно увидеть, (Рис.1), что период функции $ca(3,3,t)$ на порядок больше, чем период функции $ca(0,05;0,05;t)$. Мы должны оценить скорость роста и убывания функции *Ateb*-синуса в зависимости от параметров. Как показано в работе [8], производная *Ateb*-синуса задается формулой

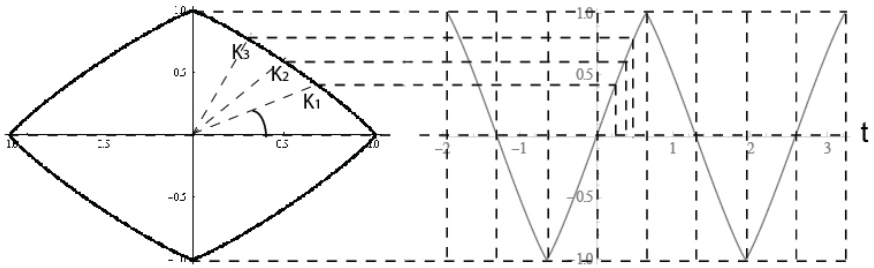


Рис. 2: Алгоритм формирования дискретного шумового сигнала на основании $sa\left(\frac{1}{7}, \frac{1}{7}, t\right)$(слева представлен суперэллипс, справа - график функции *Ateb*-синуса), пунктирные линии соответствуют дискретным сигнальным точкам.

$$\frac{dsa(n,m,t)}{dt} = \frac{2}{n+1} ca^m(m,n,t) \tag{7}$$

График функции *Ateb*-синуса аналогичен косинусу и представлен на Рис. 2. Когда $|ca(n,m,t) << 1|$, функция $sa(m,n;t)$ растет более медленно с увеличением параметров n и m. Это свойство также продемонстрировано на Рис. 2. С помощью параметров n и m, вышеупомянутые свойства позволяют влияние спектра шумоподобного сигнала $u(m,n,t)$ сгенерированного с помощью формулы (6). Способ формирования шумового сигнала на основе *Ateb*-синуса показано на Рис. 2. Метод будет применяться для формирования непрерывных или дискретных сигналов. Этот рисунок показывает систему прямоугольных координат с отправной точкой K с координатами (1,0) когда начальная фаза в (6) $\phi(t) = 0$, которая движется против часовой стрелки с постоянной скоростью по суперэллипсу (точка $K_1; K_2; K_3$ в дискретном случае), и шумовой сигнал $u(m,n,t)$ формируется по формуле (6). Графики синтезированных сигналов з дифференциальной базой B показаны на рисунках 3, 4. Эти и последующие фигуры созданы при помощи вычислительной системы MatLab 7.0. Для того чтобы улучшить мощность шума и исследовать статистические характеристики сгенерированных сигналов, были обсуждены нормированные функции автокорреляции и спектры. Белый шум - это случайный сигнал с постоянной плотностью спектральной мощности. Особым типом является Гауссовский белый шум (ГБШ), который имеет Гауссовскую амплитуду распределения. Это математический подход для описания реальных шумовых сигналов. В соответствии с этими математическими описаниями, сгенерированными с помощью уравнения (6) широкополосные шумоподобные сигналы относятся к Гауссовскому белому шуму, который иногда называют как шумопо-

добные сигналы. Статистические характеристики ГБШ: среднее значение равно нулю, и дисперсия ограниченная.

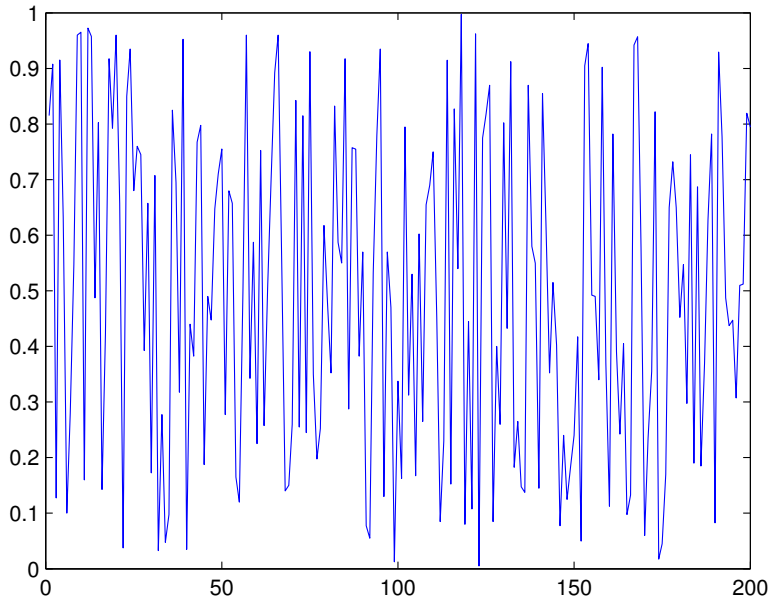


Рис. 3: Шум, производимый на основании уравнения (6), базовый сигнал $B = 200$ и *Ateb* параметры $n = \frac{1}{7}$; $m = \frac{1}{7}$

4. Исследование статистических характеристик сгенерированных шумовых сигналов

Составляет большой интерес исследовать характеристики шумовых сигналов, которые были сгенерированы на основе Ateb-функций. Используя выражение (6) и имитацию метода моделирования [4], была сформирована последовательность шумов $u(m, n, t)$. Представим результаты исследования АКФ этих сигналов и их спектров с использованием расчетной системы MatLab 7.0. Анализ спектров всех исследованных реализаций сигналов показывает равномерность спектров сигналов в полосе Δf, то есть их близость к белому шуму в этой полосе. Функция автокорреляции (АКФ) шумового сигнала показана на Рис. 5 с параметрами $B = 200$ $n = \frac{1}{7}, m = \frac{1}{7}$, а на Рис.6 с базой сигнала $B = 1000$ и параметрами *Ateb*-функции $n = \frac{1}{20}, m = \frac{1}{20}$. АКФ во всех исследуемых случаях имеет один большой пик около нуля и после этого быстро стремится к нулю. Этот пик также представлен на Рис.6, но это визуально сливается с вертикальной осью. Анализ АКФ и спектра сигнала $u(m, n, t)$ показывает его принадлежность к Гауссовскому

типу шумового сигнала. Полученный спектр синтезированного шумового сигнала $u(m, n, t)$ (Рис. 5) является постоянным, непрерывным, и его спектральная плотность мощности распределяется по всей полосе частот, что указывает на его лучшие маскировочные свойства в канале по сравнению с гармоническим сигналом.

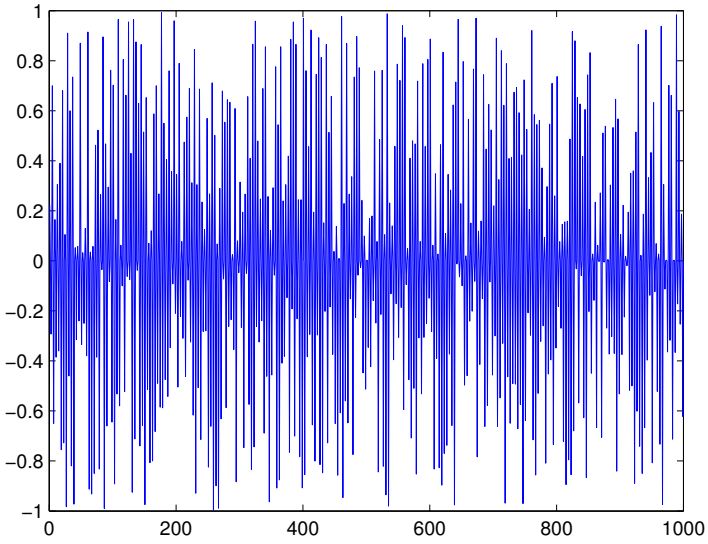


Рис. 4: Шум, производимый на основании уравнения (6), базовый сигнал $B = 200$ и *Ateb* параметры $n = \frac{1}{20}$; $m = \frac{1}{20}$

В основе технологии CDMA находится технология передачи SST (DH-SS Direct Sequence Spread Spectrum), где информация "размазывается"по широкому кругу частот. Последовательность информационных битов умножается на псевдослучайную последовательность коротких импульсов. Затем получается сигнал в широком диапазоне частот с гораздо меньшей интенсивностью. Для декодирования этой последовательности, вы должны знать, псевдослучайную последовательность, которая была использована в процессе передачи. Этот механизм кодирования обеспечивает защиту сигнала от подслушивания. Необходимо знать псевдослучайную последовательность - ключ. Ширина полосы сигнала позволяет возобновлять сигнал, особенно если помехи являются узкополосными. Аналогично сигнал защищен и от временного исчезновения на отдельных частотах (затухание). Для технологии CDMA могут быть использованы широкополосные сигналы [1], в частности шумоподобные сигналы. Итак предложенный метод синтеза шумоподобных сигналов может быть использован для защиты

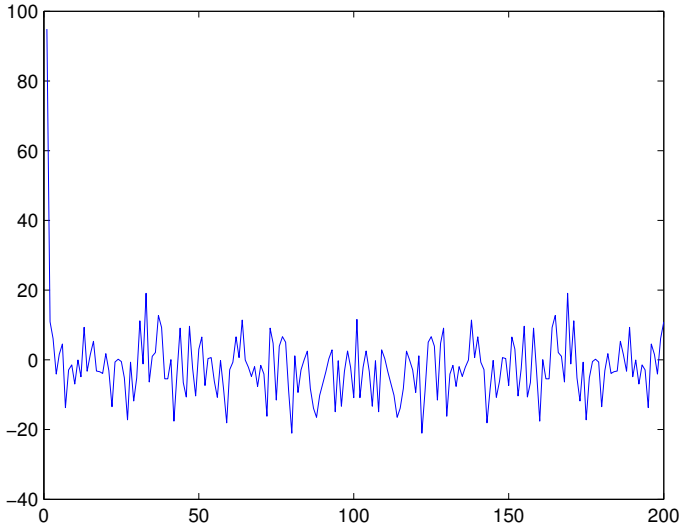


Рис. 5: АКФ сгенерированного шумового сигнала $B = 200$, $n = \frac{1}{7}$; $m = \frac{1}{7}$

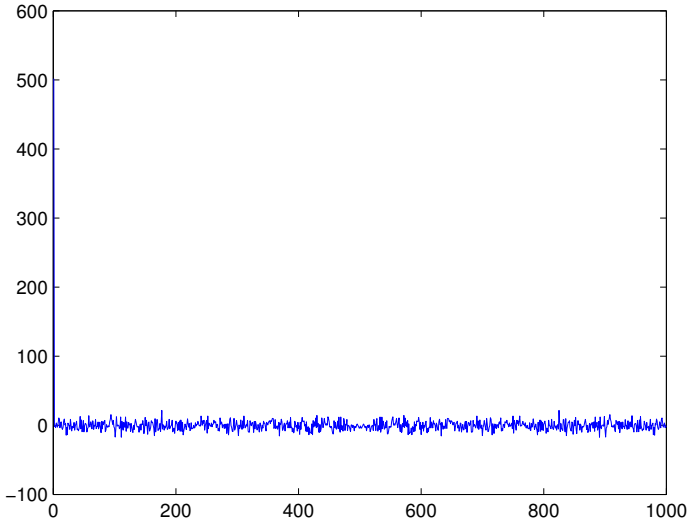


Рис. 6: АКФ сгенерированного шумового сигнала $B = 1000$, $n = \frac{1}{20}$; $m = \frac{1}{20}$

информации в процессе передачи данных, а также и в других случаях, когда потребуется Гауссовские шумовые сигналы с большой информационной емкостью.

5. Заключение

Предложена математическая модель шумового сигнала на основе Ateb-функций. Показана связь между Ateb-функциями и кривой суперэллипса. Представлено осуществление способа синтеза непрерывных или дискретных сигнальных структур с заданными спектральными и автокорреляционными характеристиками на основе Ateb-функций. Необходимые свойства реализации сформированного шума предоставляются путем изменения параметров Ateb-функций. Было показано применение шумовых сигналов для целей информационной безопасности в современных системах связи, в том числе конфиденциальных. В дальнейших исследованиях, будет интересным изучить изменение амплитуды, частоты и фазы сигнала, которые изменяются в соответствии с законом распределения случайных или псевдослучайных чисел. Также будет интересно исследовать сравнительный анализ реализаций шумовых сигналов, основанных на разных законах распределения.

ЛИТЕРАТУРА

1. Захарченко Н.В. Повышение скрытности передачи конфиденциальной информации на базе хаотических сигналов и таймерных сигнальных конструкций / Н.В. Захарченко, В.В. Корчинский, Б.К. Радзимовский, В.И. Кильдишев // Восточно-Европейский журнал передовых технологий. - 2012. - No. 3/9 (57). - С. 45-49.
2. Spread Spectrum and CDMA: Principles and Applications. Valery P. Ipatov.- 2005, John Wiley and Sons, Ltd. ISBN: 0-470-09178-9
3. Bernard Sklar Digital Communications: Fundamentals and Applications (2nd Edition), 2001.
4. Carmine Rizzo, Charles Brookson ETSI White Paper No. 1 Security for ICT - the Work of ETSI Fourth edition. January 2012.
5. Информационные технологии защиты документов средствами Ateb-функций. Ч. 1. Построение базы данных Ateb-функций для защиты документов / В. В. Грыцык, И. М. Дронюк, М. А. Назаркевич // Пробл. упр. и информатики. - 2009. -No. 2. - С. 139-152. - Библиогр.: 13 назв. - рус.
6. Sokolov, D.D. (2001), “Lame curve”, in Hazewinkel, Michiel, Encyclopedia of Mathematics, Springer, ISBN 978-1-55608-010-4.
7. Корчинский В.В. Метод моделирования шумовых сигналов для систем передачи конфиденциальной информации / В.В. Корчинский // Вестник НТУ Харьков, 2013. - No. 38(1011). - С. 99-104.
8. Сенык П. М. Про Ateb-функции / П. М. Сенык //Докл. АН УРСР, сер. А. : 1968. : No. 1. - С. 23 - 27.

LIMITING PROPERTIES OF ADAPTIVE CONTROL SYSTEM CONFLICT FLOWS NONHOMOGENEOUS ARRIVALS

M. Fedotkin, E. Kudryavtsev
Lobachevsky State University of Nizhni Novgorod,
Nizhni Novgorod, Russian Federation

Abstract

In this paper, we consider the process of managing conflict flows heterogeneous requirements of the class of adaptive algorithms. This algorithm takes into account: 1) the state of the operating device; 2) information on the quantities queues flows; 3) the information on the arrival time requirements for each of the streams. Recurrence relations for the generating functions of one-dimensional distributions of the vector sequence, which determines the state of the operating device and the amount of flow queues. We study the limit properties of these recurrence relations. This allows you to develop a method to obtain easily verifiable necessary and sufficient conditions of existence of steady state operation of the system maintenance requirements and flow control. The method is based on an analysis of the recurrence relations for multidimensional generating functions.

ПРЕДЕЛЬНЫЕ СВОЙСТВА СИСТЕМЫ АДАПТИВНОГО УПРАВЛЕНИЯ КОНФЛИКТНЫМИ ПОТОКАМИ НЕОДНОРОДНЫХ ТРЕБОВАНИЙ

М.А. Федоткин, Е.В. Кудрявцев
Нижегородский государственный университет им. Н.И. Лобачевского,
г. Нижний Новгород, Российская Федерация
fma5@rambler.ru, evgkudryavcev@gmail.com

Аннотация

В данный работе рассматривается процесс управления конфликтными потоками неоднородных требований в классе адаптивных алгоритмов. При этом алгоритм учитывает: 1) состояние обслуживающего устройства; 2) информацию о величинах очередей по потокам; 3) информацию о времени поступления требований по каждому из потоков. Получены рекуррентные соотношения для производящих функций одномерных распределений векторной последовательности, которая определяет состояние обслуживающего устройства и величины очередей по потокам. Изучены предельные свойства указан-

ных рекуррентных соотношений. Это позволяет разработать метод получения легко проверяемых необходимых и достаточных условий существования стационарного режима функционирования системы обслуживания требований и управления потоками. Метод основан на анализе рекуррентных соотношений для многомерных производящих функций.

Ключевые слова: конфликтные потоки, предельные свойства, адаптивный алгоритм управления

1. Введение

В работе рассматривается транспортный перекресток как система массового обслуживания. Обслуживаются 2 конфликтных неординарных пуассоновских потока Π_1, Π_2. Вероятность получить количество требований i в каждый вызывающий момент по потоку j равна $P_j(i), i \geq 1, j = 1, 2,$

$$P_j(1) = (1 + \alpha_j + \alpha_j\beta_j/(1 - \gamma_j))^{-1},$$
$$P_j(2) = \alpha_j(1 + \alpha_j + \alpha_j\beta_j/(1 - \gamma_j))^{-1},$$
$$P_j(m) = \alpha_j\beta_j\gamma_j^{m-3}(1 + \alpha_j + \alpha_j\beta_j/(1 - \gamma_j))^{-1}, m \geq 3,$$

где α_j, β_j и γ_j — некоторые параметры распределения. Интенсивность поступления вызывающих моментов по потоку Π_j равна λ_j. Свойства таких потоков изучены в [1, 2, 3, 4].

Обслуживание производится с помощью адаптивного нециклического алгоритма, подробное описание которого приведено в работе [5].

2. Описание системы

В системе обслуживающим устройством является светофор, а требованиями — автомобили, подъезжающие к светофору. Множество состояний светофора $\Gamma = \{\Gamma^{(1)}, \Gamma^{(2)}, \Gamma^{(3)}, \Gamma^{(4)}, \Gamma^{(5)}, \Gamma^{(6)}, \Gamma^{(7)}, \Gamma^{(8)}\}$. Вкратце опишем назначение каждого из состояний.

Состояние $\Gamma^{(3j-2)}$ (зеленый сигнал светофора для j-го потока) соответствует первому этапу периода обслуживания j-го потока. Длительность обслуживания одной заявки, поступившей из накопителя, равна постоянной величине $\mu_{j,1}^{-1}$. Длительность пребывания в $\Gamma^{(3j-2)}$ равна T_{3j-2}.

Состояние $\Gamma^{(3j-1)}$ (зеленый сигнал светофора для j-го потока) соответствует второму этапу периода обслуживания j-го потока. Длительность обслуживания одной заявки равна величине $\mu_{j,2}^{-1} < \mu_{j,1}^{-1}$. Длительность пребывания в этом состоянии — случайная величина, принимающая значения kT_{3j-1} , $k = \overline{1, n_j}$, где n_j — максимальное число продлений и T_{3j-1} — длительность одного продления. Продление происходит в 2 случаях: 1) длина очереди по j-му потоку не меньше параметра K_j, 2) на предыдущем этапе продлений поступили требования, которые необходимо обслужить.

Состояние $\Gamma^{(3j)}$ (желтый сигнал светофора для j-го потока) соответствует режиму переналадки для j-го потока. Длительность пребывания в этом состоянии равна T_{3j}.

Состояние $\Gamma^{(6+j)}$ (зеленый сигнал светофора для j-го потока) соответствует первому этапу периода обслуживания j-го потока, в случае, когда возможен мгновенный переход в состояние $\Gamma^{(3j)}$. Длительность пребывания в $\Gamma^{(6+j)}$ является случайной величиной. Максимальное время пребывания в этом состоянии равно T_{3j-2} .

Далее систему будем рассматривать в моменты $\tau_i, i \geq 0$, или на промежутках $[\tau_i, \tau_{i+1})$. Здесь τ_0 — начальный момент времени, а $\tau_i, i \geq 0$ — моменты смены состояний обслуживающего устройства. Пусть $y_0 = (0,0)$, $y_1 = (1,0), y_2 = (0,1)$ и X — целочисленная одномерная неотрицательная решетка. Для нелокального описания системы при $i = 0, 1, \ldots$ введем следующие случайные величины и элементы:

1) $\Gamma_i \in \Gamma$ — состояние обслуживающего устройства на интервале $[\tau_i, \tau_{i+1})$;
2) $\eta_{j,i} \in X$ — число заявок j-го потока, поступивших в систему за промежуток $[\tau_i, \tau_{i+1})$, $\eta_i = (\eta_{1,i}, \eta_{2,i})$;
3) $\eta'_{j,i}$ — случайный вектор, принимающий значения y_0, если на i-ом такте $[\tau_i, \tau_{i+1})$ в систему не поступило ни одной заявки, y_j, если на i-ом такте первой пришла заявка (или заявки) j-го потока;
4) $\kappa_{j,i} \in X$ — число заявок j-го потока, которые находятся в системе в момент τ_i, $\kappa_i = (\kappa_{1,i}, \kappa_{2,i})$;
5) $\xi_{j,i}$ — максимально возможное число заявок j-го потока, которые система может обслужить на интервале $[\tau_i, \tau_{i+1})$, $\xi_i = (\xi_{1,j}, \xi_{2,j})$.

Примем следующие соотношения:

$$T_{3j-2} = \mu_{j,1}^{-1} + l_{3j-2}\alpha_j\mu_{j,1}^{-1},$$
$$T_{3j-1} = l_{3j-1}\alpha_j\mu_{j,2}^{-1},$$
$$T_{3j} = l_{3j}\alpha_j\mu_{j,2}^{-1},$$

где $l_{3j-1}, l_{3j-2} \in X, l_{3j-1} > 0$. Параметр l_{3j} выбирается так, чтобы выполнялось неравенство $T_{3j} \geq \mu_{j,1}^{-1}$. Величина $0 < \alpha_j \leq 1$ обозначает часть обслуживания, которую необходимо пройти требованию, чтобы можно было начать обслуживать следующую заявку. В случае $\alpha_j < 1$ одновременно может обслуживаться несколько требований.

3. Формализация работы обслуживающего устройства и изменения длины очереди по потокам Π_1, Π_2

Адаптивный алгоритм смены состояний обслуживающего устройства из множества Γ задается с помощью рекуррентного соотношения:

$$\Gamma_{i+1} = \begin{cases} \Gamma^{(3j-2)}, & \left\{\left[\Gamma_i = \Gamma^{(3s)}\right] \& \left[(\kappa_{j,i} > 0) \vee (\kappa_{s,i} \geq K_s) \vee (\eta'_i = y_j)\right]\right\} \vee \\ & \vee \left\{\left[\Gamma_i = \Gamma^{(3j)}\right] \&[\kappa_{s,i} = 0] \&[\kappa_{j,i} \leq K_s] \&[\eta'_i = y_j]\right\}, \\ \Gamma^{(3j-1)}, & \left\{\Gamma_i = \Gamma^{(3j-2)}\right\} \vee \left\{\left[\Gamma_i = \Gamma^{(6+j)}\right] \&[\eta'_i = y_j]\right\}, \\ \Gamma^{(3j)}, & \left\{\Gamma_i = \Gamma^{(3j-1)}\right\} \vee \left\{\left[\Gamma_i = \Gamma^{(6+j)}\right] \&[\eta'_i \neq y_j]\right\}, \\ \Gamma^{(6+j)}, & \left[\Gamma_i = \Gamma^{(3s)}\right] \&[\kappa_{j,i} = 0] \&[\kappa_{s,i} < K_s] \&[\eta'_i = y_0]. \end{cases}$$

Как видно из приведенного соотношения состояние обслуживающего устройства на следующем шаге зависит от состояния на предыдущем шаге, длины очередей и очередности прихода заявок. При этом динамика длины очереди задается следующими рекуррентными соотношениями

$$\kappa_{j,i+1} = \begin{cases} \max\{0; \kappa_{j,i} + \eta_{j,i} - \xi_{j,i}\}, \text{ если } \Gamma_i \in \Gamma \backslash \{\Gamma^{(3)}, \Gamma^{(6)}\}; \\ \eta_{j,i} + \max\{0; \kappa_{j,i} - \xi_{j,i}\}, \text{ если } \Gamma_i \in \{\Gamma^{(3)}, \Gamma^{(6)}\}. \end{cases}$$

4. Свойства последовательности $\{(\Gamma_i, \kappa_i); i = 0, 1, \ldots\}$

Состояние системы на i-м такте времени $[\tau_i, \tau_{i+1})$ описывается случайным элементом $(\Gamma_i(\omega), \kappa_i(\omega)), i = 0, 1, \ldots$ Для векторной последовательности $\{(\Gamma_i, \kappa_i); i = 0, 1, \ldots\}$ доказана марковость и проведена классификация ее состояний.

Теорема 1. *Случайная последовательность $\{(\Gamma_i, \kappa_i); i = 0, 1, \ldots\}$ с заданным начальным распределением вектора (Γ_0, κ_0) является марковской.*

Теорема 2. *Пусть $j, s = 1, 2, j \neq s, x = (x_1, x_2) \in X^2$ и*

$$G = \{(\Gamma^h, x) : \Gamma^h \in \Gamma, x \in X^2\},$$
$$G^{(3j-2)} = \{(\Gamma^{(3j-2)}, x_s y_s) : x_s < K_s - l_{3s}\},$$
$$G^{(3j-1)} = \{(\Gamma^{(3j-1)}, x_s y_s) : x_s < K_s - l_{3s}\},$$
$$G^{(6+j)} = \{(\Gamma^{(6+j)}, x) : x_j > 0\} \cup \{(\Gamma^{(6+j)}, x) : x_s \geq K_s - l_{3s}\},$$
$$G_j = \begin{cases} G^{(6+j)} \cup G^{(3j-2)}, l_{3j-2} > 0, \\ G^{(6+j)} \cup G^{(3j-2)} \cup G^{(3j-1)}, l_{3j-2} = 0. \end{cases}$$

Тогда состояния из G_j являются несущественными и множество вида $G_0 = G \backslash (G_1 \cup G_2)$ является неразложимым апериодическим классом существенных состояний.

5. Рекуррентные соотношения для распределений последовательности $\{(\Gamma_i, \kappa_i); i = 0, 1, \ldots\}$

Для любого $i \geq 0, r = \overline{1,8}; x \in X^2$ введем обозначение:

$$Q_i^{(r)}(x) = \mathbf{P}(\Gamma_i = \Gamma^{(r)}, \kappa_i = x).$$

В работе [5] были найдены рекуррентные соотношения для одномерных распределений $\{Q_i^{(r)}(x) : r = \overline{1,8}, x \in X^2\}, i \geq 0$, марковской последовательности $\{(\Gamma_i, \kappa_i); i = 0, 1, \ldots\}$. Приведем одно из таких соотношений

$$\begin{aligned} &Q_{i+1}^{(3j-2)}(w_j y_j) = \\ &= \sum_{x_j=1}^{w_j} \sum_{x_s=0}^{l_{3s}} Q_i^{(3s)}(x)\varphi_{3s}((w_j - x_j)y_j) + \sum_{x_s=0}^{l_{3s}} Q_i^{(3s)}(x)\varphi_{3s,j}(w_j y_j) + \\ &+ \sum_{x_j=0}^{l_{3j}} Q_i^{(3j)}(x)\varphi_{3j,j}(w_j y_j) + \sum_{x_j=l_{3j}+1}^{\min\{K_j-1, w_j+l_{3j}\}} Q_i^{(3j)}(x)\varphi_{3j,j}((w_j - l_{3j})y_j), \end{aligned} \tag{1}$$

где $w_j \in X$ и $\varphi_{3s}(x), \varphi_{3s,j}(x), \varphi_{3j,j}(x)$ — вспомогательные функции, вид которых приведен в [5] и определяется распределением входных потоков.

Пусть $z = (z_1, z_2)$, где z_1, z_2 действительные или комплексные переменные и $|z_1| \leq 1$, $|z_2| \leq 1$. Положим $z^x = z_1^{x_1} z_2^{x_2}$, где $x = (x_1, x_2) \in X^2$. Рассмотрим теперь производящие функции

$$W_i^{(r)}(z) = \sum_{x \in X^2} Q_i^{(r)}(x) z^x, r = \overline{1,8};$$

$$W_i(z) = \sum_{r=1}^{8} W_i^{(r)}(z).$$

Используя рекуррентные соотношения типа (1) для одномерных распределений $\{Q_i^{(r)}(x) : r = \overline{1,8}, x \in X^2\}, i \geq 0$, векторной последовательности $\{(\Gamma_i, \kappa_i); i = 0, 1, \ldots\}$, стандартным образом были получены рекуррентные соотношения для производящих функций $W_i^{(r)}(z)$, $r = \overline{1,8}, i > 0$. В качестве примера приведем здесь только одно из полученных рекуррентных соотношений

$$\begin{aligned} W_{i+1}^{(3j)}(z) &= \exp(-(\lambda_1 + \lambda_2)T_{3j-2}) W_i^{(6+j)}(z) + q_{6+j,s}(z) W_i^{(6+j)}(z) + \\ &+ q_{3j-1}^{n_j}(z) W_i^{(3j-1)}(z) + \sum_{k=1}^{n_j-1} (1 - q_{3j-1}^{n_j-k}(z)) \Phi_{i,k}^{(3j-1)}(z) + R_i^{(3j-1)}(z), \end{aligned} \tag{2}$$

где $j, s = 1, 2$ при $j \neq s$ и

$$q_{6+j,s}(z) = \sum_{a_j=0}^{l_{3j-2}} \sum_{k=1}^{\infty} \varphi_{6+j,s,a_j}(ky_s) z_s^k,$$

$$q_r(z) = z_j^{-l_r} \exp((\lambda_1(z_1 - 1) + \lambda_2(z_2 - 1))T_r),\ r = \overline{3j-2, 3j},$$

$$\Phi_{i,k}^{(3j-1)}(z) = \sum_{x_j=0}^{K_j + kl_{3j-1}-1} \sum_{x_s=0}^{\infty} Q_i^{(3j-1)}(x) z^x z_j^{-kl_{3j-1}} \sum_{b \in X^2} \varphi_{3j-1,k}(x_j, b) z^b,$$

$$k = \overline{1, n_j - 1},$$

$$R_i^{(3j-1)}(z) = \sum_{k=1}^{n_j} \sum_{x_s=0}^{\infty} \sum_{x_j=0}^{kl_{3j-1}-1} Q_i^{(3j-1)}(x) z_s^{x_s} \times$$

$$\times \sum_{b_j=0}^{kl_{3j-1}-x_j} \sum_{b_s=0}^{\infty} \varphi_{3j-1,k}(x_j, b) z_s^{b_s} (1 - z_j^{x_j + b_j - kl_{3j-1}}).$$

Заметим, что функции $\varphi_{6+j,s,a_j}(x)$ определены в [5] и определяются распределением каждого из входных потоков.

Если очереди неограниченно возрастают, то это является признаком отсутствия стационарного режима на перекрестке. В случае больших значений длин очередей адаптивный алгоритм работает как циклический, переключаясь по состояниям $\Gamma^{(1)} \to \Gamma^{(2)} \to \ldots \to \Gamma^{(6)} \to \Gamma^{(1)}$. В силу этого состояния $\Gamma^{(r)}$, $r = \overline{1, 6}$, назовем основными состояниями обслуживающего устройства. Целесообразно проитерировать соотношения типа (2) для производящих функций шесть раз (по числу основных состояний).

Введем теперь следующие операции для произвольных чисел $v, v' \in X$:

$$a \oplus b = a + b \text{ при } a + b \leq 6; a \oplus b = a + b - 6 \text{ при } a + b > 6;$$
$$a \ominus b = a - b \text{ при } a - b \geq 0; a \ominus b = a - b + 6 \text{ при } a - b < 0.$$

Обозначим $q(z) = q_1(z) q_2^{n_1}(z) q_3(z) q_4(z) q_5^{n_2}(z) q_6(z)$. Далее, для $r, r' = \overline{1, 6}$; $i \geq 0$; $j, s = 1, 2$; $j \neq s$ введем обозначения:

$$\bar{r} = r - 1,\ \ r \neq 1;\ \bar{r} = 6,\ \ r = 1;$$
$$\delta(r) = n_1 \delta_{2,r} + n_2 \delta_{5,r}, \quad \delta_{u,v} — \text{символ Кронекера};$$
$$\bar{q}_{r,r'}(z) = q_r^{\delta(r)}(z) \cdot q_{r \oplus 1}^{\delta(r \oplus 1)}(z) \cdot \ldots \cdot q_{r'}^{\delta(r')}(z),\ \ r' \neq \bar{r};$$
$$\bar{q}_{r,r'}(z) = 1,\ \ r' = \bar{r};$$

$$\mathcal{R}^{(r)}_{j,6i}(z) = \sum_{v=3j-2}^{3j-1} \bar{q}_{v\oplus 1,\bar{r}}(z) R^{(v)}_{6i+(v\ominus r)}(z) + \bar{q}_{3s-2,\bar{r}}(z) R^{(3j)}_{6i+(3j\ominus r),s}(z) +$$

$$+\bar{q}_{3j-2,\bar{r}}(z) R^{(3j)}_{6i+(3s\ominus r),j}(z) + \bar{q}_{3j-1,\bar{r}}(z) R^{(6+j)}_{6i+((3j-2)\ominus r),j};$$

$$\Psi^{(r)}_{j,6i}(z) = \bar{q}_{3j,\bar{r}}(z) \sum_{k=1}^{n_j-1} (1 - q^{n_j-k}_{3j-1}(z)) \Phi^{(3j-1)}_{6i+((3j-1)\ominus r),k}(z).$$

Наконец, обозначим

$$\Upsilon^{(r)}_{6i}(z) = \sum_{j=1}^{2} \Big\{ \bar{q}_{3j,\bar{r}}(z)(\exp(-\lambda T_{3j-2}) + q_{6+j,s}(z)) W^{(6+j)}_{6i+((3j-1)\ominus r)}(z) +$$

$$+q_{6+j,j}(z)\bar{q}_{3j-1,\bar{r}}(z) W^{(6+j)}_{6i+((3j-2)\ominus r)}(z) -$$

$$q_{3j,0}(z)\bar{q}_{3s-2,\bar{r}}(z) \bar{W}^{(3j)}_{6i+(3j\ominus r)}(z) -$$

$$- q_{3j,j}(z)\bar{q}_{3s-2,\bar{r}}(z) \bar{W}^{(3j)}_{6i+(3j\ominus r)}(z) + q_{3j,j}(z)\bar{q}_{3j-2,\bar{r}}(z) \bar{W}^{(3j)}_{6i+(3s\ominus r)}(z).$$

Тогда для $i \geq 0$, $r = \overline{1,6}$ имеем место соотношение:

$$W^{(r)}_{6(i+1)}(z) = q(z) W^{(r)}_{6i}(z) + \Upsilon^{(r)}_{6i}(z) + \sum_{j=1}^{2} \left\{ \mathcal{R}^{(r)}_{j,6i}(z) + \Psi^{(r)}_{j,6i}(z) \right\}. \qquad (3)$$

Для $i \geq 0$, $j = 1, 2$ введем следующие обозначения:

$$\mathcal{R}_{j,6i}(z) = \sum_{r=1}^{6} \mathcal{R}^{(r)}_{j,6i}(z);$$

$$\Psi_{j,6i}(z) = \sum_{r=1}^{6} \Psi^{(r)}_{j,6i}(z);$$

$$\Upsilon_{6i}(z) = \sum_{r=1}^{6} \Upsilon^{(r)}_{6i}(z) - \sum_{j=1}^{2} q(z) W^{(6+j)}_{6i}(z) + \sum_{j=1}^{2} W^{(6+j)}_{6(i+1)}(z).$$

Используя введенные обозначения, найдем

$$W_{6(i+1)}(z) = q(z) W_{6i}(z) + \Upsilon_{6i}(z) + \sum_{j=1}^{2} \left\{ \mathcal{R}_{j,6i}(z) + \Psi_{j,6i}(z) \right\}, \; i \geq 0. \qquad (4)$$

Соотношения (3) и (4) позволят в дальнейшем найти необходимые и достаточные условия существования предельного распределения изучаемой марковской последовательности. В частности было доказано следующее утверждение.

Введем обозначения

$$T_0 = T_1 + T_3 + T_4 + T_6; \quad T = T_0 + n_1 T_2 + n_2 T_5;$$
$$L_j = l_{3j-2} + n_j l_{3j-1} + l_{3j};$$
$$\Lambda_j = \lambda_j (1 + \alpha_j + \alpha_j \beta_j / (1 - \gamma_j))^{-1} (1 + 2\alpha + \alpha\beta(\frac{2}{1-\gamma} + \frac{1}{(1-\gamma)^2})).$$

Теорема 3. *Для существования предельного распределения марковской последовательности* $\{(\Gamma_i, \kappa_i); i \geq 0\}$ *необходимо и достаточно, чтобы выполнялось следующее условие: существует число* $\varepsilon > 0$*, что при любом распределении случайного элемента* (Γ_0, κ_0) *найдется номер* I *такой, что для любых* $i > I$ *выполняется:*

$$\Lambda_j T - L_j + \frac{d}{dz_j} \Upsilon_{6i}(\overline{z}_j)|_{z_j=1} + (\Lambda_j T_{3j-1} - l_{3j-1}) \times$$
$$\times \sum_{v=0}^{5} \sum_{k=1}^{n_j - 1} (k - n_j) \Phi_{6i+v,k}^{(3j-1)}(\overline{1}) + \Lambda_j T_{3s-1} \sum_{v=0}^{5} \sum_{k=1}^{n_s - 1} (k - n_s) \Phi_{6i+v,k}^{(3s-1)}(\overline{1}) < -\varepsilon. \tag{5}$$

ЛИТЕРАТУРА

1. Fedotkin M.A., Kudryavtsev E.V., Rachinskaya M.A. About correctness of probabilistic models of traffic flows dynamics on a motorway // Proceedings of the International Workshop «Distributed com-puter and communication networks» (DCCN-2010), ISBN 978-9901871-2-2. Moscow, 2010, pp. 86-93
2. Федоткин М.А., Кудрявцев Е.В. Оценка параметров вероятностной модели интенсивного транспортного потока // Proceedings of International Workshop «Distributed com-puter and communication networks» (DCCN-2013). ISBN 978-5-94836-366 -0. Moscow, 2013, pp. 365-372
3. Fedotkin M.A., Fedotkin A.M., Kudryavtsev E.V. Construction and Analysis of a Mathematical Model of Spatial and Temporal Characteristics of Traffic Flows // Automatic Control and Computer Sciences, 2014, Vol. 48, No 6, pp. 358-367, Allerton Press, Inc.
4. Fedotkin M.A., Fedotkin A.M., Kudryavtsev E.V. Nonlocal description of the time characteristic for input flows by means of observations // Automatic Control and Computer Sciences, 2015, Vol. 49, No 1, pp. 29-36, Allerton Press, Inc.
5. Федоткин М.А., Кудрявцев Е.В. Построение математической модели адаптивного управления неординарными потоками // Материалы Международной научной конференции «Теория вероятностей, случайные процессы, математическая статистика и приложения», Минск, БГУ, 2015, с. 106-111.

PERFORMANCE EVALUATION OF BROADBAND WIRELESS NETWORKS ALONG THE LONG TRANSPORT ROUTES

V.M. Vishnevsky[1], *A.N. Dudin*[2], *D.V. Kozyrev*[1,3], *A.A. Larionov*[1]
[1] V.A. Trapeznikov Institute of Control Sciences of Russian Academy of Sciences, Moscow, Russia
[2] Belarusian State University, Minsk, Belarus
[3] Peoples' Friendship University of Russia, Moscow, Russia

Abstract

This paper presents the description of the method of performance evaluation and assessment of main characteristics of broadband wireless networks with linear topology, which is based on a model of stochastic multiphase queueing systems with correlated MAP input flows and a cross-traffic. The results of analytical calculations for the networks of small dimension are given. A simulation model for assessment of performance characteristics of large-scale wireless networks with linear topology is developed.

Keywords: wireless networks, stochastic multi-stage systems, Markov arrival flow, performance evaluation, analytical modeling and simulation

ОЦЕНКА ПРОИЗВОДИТЕЛЬНОСТИ ШИРОКОПОЛОСНЫХ БЕСПРОВОДНЫХ СЕТЕЙ ВДОЛЬ ПРОТЯЖЕННЫХ ТРАНСПОРТНЫХ МАГИСТРАЛЕЙ[1]

В.М. Вишневский[1], *А.Н. Дудин*[2], *Д.В. Козырев*[1,3], *А.А. Ларионов*[1]
[1]Институт проблем управления им. В.А. Трапезникова РАН, Москва, Россия
[2] Белорусский государственный университет, Минск, Беларусь
[3] Российский университет дружбы народов, Москва, Россия
vishn@inbox.ru, dudin@bsu.by, larioandr@gmail.com, kozyrevdv@gmail.com

Аннотация

В статье дано описание метода оценки производительности и основных характеристик широкополосных беспроводных сетей с линейной топологией на базе модели стохастической многофазной системы

[1]Работа выполнена при финансовой поддержке Министерства образования и науки Российской федерации, проект №14.613.21.0020 от 22.10.2014 (RFMEFI61314X0020).

с коррелированными входными MAP-потоками и кросс-трафиком. Приводятся результаты аналитических расчетов для сетей небольшой размерности. Разработана имитационная модель для исследования характеристик беспроводных сетей с линейной топологией большой размерности.

Ключевые слова: беспроводные сети, стохастические многофазные системы, MAP-потоки, оценка производительности, аналитическое и имитационное моделирование

1. Введение

Одной из важнейших проблем при разработке новых и функционировании существующих транспортных магистралей (железнодорожных и автомобильных дорог, нефтяных и газовых трубопроводов) является создание современной инфраструктуры беспроводной связи, базирующейся на международном стандарте IEEE 802.11–2012 [1]. Указанный стандарт регламентирует создание высокоскоростных каналов связи и беспроводных сетей, функционирующих под управлением протоколов IEEE 802.11n и IEEE 802.11s, на основе которых могут эффективно реализовываться беспроводные сети вдоль протяженных транспортных магистралей. Указанные сети обеспечивают не только высокоскоростную передачу мультимедийной информации между базовыми станциями, расположенными на высотных зданиях и вышках вдоль транспортных магистралей, но и оперативную связь со стационарными и мобильными абонентами (автомобили, поезда, дорожные знаки, пункты весового контроля и пункты контроля ПДД, управления светофорами и т.д.) [2, 3]. Развертывание и развитие сетей беспроводной связи вдоль протяженных магистралей требует решения ряда сложных организационно-технических задач в условиях жестких ограничений на использование частотных, экономических и аппаратных ресурсов. В связи с этим возрастает актуальность решения проблемы оценки характеристик производительности, которая является одной из важнейших при проектировании широкополосных беспроводных сетей этого класса. Ее решение направлено как на реализацию высокоскоростной магистральной сети, так и максимальное телекоммуникационное покрытие трассы с целью обеспечения подключения мобильных пользователей, а также минимизации интерференции и временных задержек при передаче мультимедийной информации по сети. Исследованию этой проблемы посвящены многочисленные публикации [4]–[6]. В настоящей работе предложен новый подход к оценке производительности беспроводных сетей с линейной топологией на базе модели многофазной стохастической системы с входящим MAP-потоком и кросс-трафиком. Работа является развитием исследований, начатых в [9, 11], в части разработки эффективных вычислительных алгоритмов и проведения имитационных экспериментов,

позволяющих осуществлять оценку характеристик производительности сетей большой размерности.

2. Описание модели беспроводной сети с линейной топологией

Широкополосная беспроводная сеть вдоль протяженных транспортных магистралей представляет собой совокупность базовых станций, объединенных высокоскоростными беспроводными каналами связи. Адекватной математической моделью такой сети является многофазная система массового обслуживания с входящим MAP-потоком, PH-распределением времени обслуживания на фазах системы и кросс-трафиком (рис.1). Под «обслуживанием» каждого сообщения понимается множество технических процессов в реальной сети, длительность которых является случайной величиной. Так, сообщения обрабатываются одним или несколькими программными компонентами, время которых зависит от текущей загрузки процессора и памяти базовой станции, числа одновременно обрабатываемых сообщений, количества ядер процессора и прочих параметров. Затем сообщения поступают в устройство вывода в сеть, и передаются по одному или нескольким последовательным каналам связи до следующей станции. Время передачи по каналам связи также является случайной величиной, поскольку на него влияет фоновый трафик, используемые сетевые технологии, настройки параметров профилирования трафика и множество других факторов.

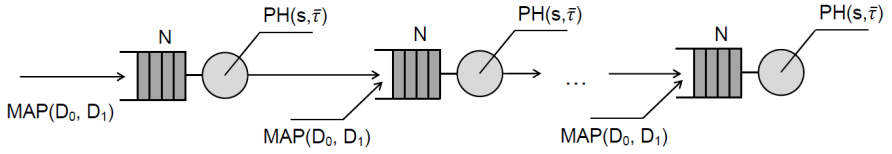


Рис. 1: схема системы массового обслуживания

В общем случае моделировать такое время обслуживания целесообразно при помощи PH-распределений: каждая фаза соответствует определенному техническому процессу (обработка программным компонентом, передача по линии связи). PH-распределение описывается цепью Маркова с непрерывным временем с $M+1$ состоянием, в котором нулевое состояние является "поглощающим" - после попадания в него сообщение считается обслуженным. $PH(\overline{\tau},\ S)$ (распределение фазового типа) - распределение времени до попадания в поглощающее состояние 0 в цепи Маркова с непрерывным временем, дискретным пространством состояний $\{0,1,\ldots,M\}$, стационарным распределением $(\tau_0,\overline{\tau})$, где $\tau_0 = 1-\overline{\tau}\overline{\mathbf{1}}_M$, и инфинитезимальным генератором T:

$$T = \begin{pmatrix} 0 & 0 \\ -S\overline{\mathbf{1}}_M & S \end{pmatrix},\ S \in \mathbb{R}^{M\times M}$$

$$\overline{\mathbf{1}}_M = [1\ \ldots\ 1]^T \in \mathbb{R}^{M\times 1}$$

Каждая базовая станция беспроводной сети передает не только сообщения, полученные от соседней станции, но и сообщения, полученные от мобильных абонентов (кросс-трафик). Будем полагать, что входные потоки и потоки сообщений кросс-трафика являются MAP-потоками, что позволяет учитывать сложный, коррелированный характер потоков в беспроводных сетях. В цепи Маркова, управляющей MAP-потоком, все переходы делятся на наблюдаемые и невидимые. Когда происходит наблюдаемый переход, поток генерирует сообщение и, если начальное и конечное состояния различаются, производится смена состояний. При невидимом переходе происходит только смена состояний. Интенсивности наблюдаемых и невидимых переходов записываются в матрицы D_1 и D_0 соответственно:

$$D_0 = \begin{cases} \lambda_{ij}^{(0)} & ,\ i \neq j \\ -\lambda_i & ,\ i = j \end{cases},\quad i,\ j = \overline{1, W}$$

$$D_1 = \left\{\lambda_{ij}^{(1)}\right\},\ i,\ j = \overline{1, W}$$

$$\lambda_i = \lambda_{ii}^{(1)} + \sum_{\substack{j=1 \\ j\neq i}}^{W} \left(\lambda_{ij}^{(0)} + \lambda_{ij}^{(1)}\right)$$

здесь $\lambda_{ij}^{(0)}$ - интенсивности невидимых переходов, $\lambda_{ij}^{(1)}$ - интенсивности наблюдаемых переходов, а λ_i - суммарная интенсивность выходов из состояния или генерации сообщения без смены состояния. Сложение матриц D_0 и D_1 дает инфинитезимальный генератор марковской цепи:

$$D = D_0 + D_1$$

Элементарной проверкой, с учетом определения λ_i, можно убедиться в том, что сумма всех элементов каждой строки генератора D равна нулю.

Стационарные вероятности $\overline{\theta} \in \mathbb{R}^W$ марковского процесса вычисляются из уравнений баланса и условия нормировки:

$$\begin{cases} \overline{\theta} D & = \overline{\mathbf{0}}_W \\ \overline{\theta}\overline{\mathbf{1}}_W & = 1 \end{cases} \tag{1}$$

где $\overline{\mathbf{0}}_W = \begin{Vmatrix} 0 & 0 & \ldots & 0 \end{Vmatrix} \in \mathbb{R}^W$ - вектор-строка, состоящий из всех нулей, а $\overline{\mathbf{1}}_W = \begin{Vmatrix} 1 & 1 & \ldots & 1 \end{Vmatrix}^T \in \mathbb{R}^W$ - вектор-столбец, состоящий из всех единиц. Используя найденное стационарное распределение вероятностей, можно вычислить среднюю интенсивность сообщений, генерируемых

MAP-потоком, как математическое ожидание случайной величины, равной суммарной интенсивности наблюдаемых переходов из заданного состояния:

$$\lambda = \sum_{i=0}^{W} \left[\theta_i \sum_{j=0}^{W} \lambda_{ij}^{(1)} \right] = \overline{\theta} D_1 \overline{\mathbf{1}}_W \tag{2}$$

Так как базовые станции беспроводной сети имеют ограниченный объем буферов, то в математической модели необходимо учитывать ограничения на размеры длин очередей обслуживающих приборов каждой фазы.

Таким образом, описана многофазная система массового обслуживания $MAP/PH/1/N \rightarrow \bullet/PH/1/N \rightarrow \cdots \rightarrow \bullet/PH/1/N$, адекватно описывающая функционирование широкополосной беспроводной сети с линейной топологией.

3. Свойства и характеристики системы MAP/PH/1/N

Прежде, чем переходить к отысканию стационарных характеристик системы MAP/PH/1/N, сформулируем следующие две теоремы:

Теорема 1. *Пусть на вход системы $MAP/PH/1/N$ поступает поток $X = MAP(D_0^{(X)}, D_1^{(X)})$, $D_0^{(X)},\ D_1^{(X)} \in \mathbb{R}^{W \times W}$, время обслуживания распределено согласно PH-распределению $Y = PH(S, \tau)$, $S \in \mathbb{R}^{M \times M}, \tau \in \mathbb{R}^{1 \times M}$. Тогда на выходе системы будет MAP-поток $Z = MAP(D_0^{(Z)}, D_1^{(Z)})$, матрицы переходов которого $D_{0,1}^{(Z)} \in \mathbb{R}^{(WM(N+2)) \times (WM(N+2))}$ определены как:*

$$D_0^{(Z)} = \begin{pmatrix} \tilde{D_0} & B_1 & 0 & 0 & \dots & 0 & 0 & 0 \\ 0 & R_0 & \tilde{D_1} & 0 & \dots & 0 & 0 & 0 \\ 0 & 0 & R_0 & \tilde{D_1} & \dots & 0 & 0 & 0 \\ 0 & 0 & 0 & R_0 & \dots & 0 & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & 0 & \dots & R_0 & \tilde{D_1} & 0 \\ 0 & 0 & 0 & 0 & \dots & 0 & R_0 & \tilde{D_1} \\ 0 & 0 & 0 & 0 & \dots & 0 & 0 & R_A \end{pmatrix} \tag{3}$$

$$D_1^{(Z)} = \begin{pmatrix} 0 & \dots & 0 & 0 & 0 \\ I_W \otimes C_t & \dots & 0 & 0 & 0 \\ 0 & \dots & 0 & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & I_W \otimes C_t & 0 & 0 \\ 0 & \dots & 0 & I_W \otimes C_t & 0 \end{pmatrix}$$

где матрицы $\tilde{D}_0$, $\tilde{D}_1$ B_1, R_0, R_A, C_t определены следующим образом:

$$
\begin{aligned}
&\tilde{D_0} = D_0 \otimes I_M \\
&\tilde{D_1} = D_1 \otimes I_M \\
&B_1 = D_1 \otimes \left(\overline{\tau} \otimes \overline{\mathbf{1}}_M\right) \\
&R_0 = D_0 \otimes I_M + I_W \otimes S - I_W \otimes C_t \\
&R_A = (D_0 + D_1) \otimes I_M + I_W \otimes S - I_W \otimes C_t \\
&C_t = \begin{bmatrix} \mu_{10} \\ \mu_{20} \\ \cdots \\ \mu_{M0} \end{bmatrix} \otimes \overline{\tau} = (-S\overline{\mathbf{1}}_M) \otimes \overline{\tau}
\end{aligned}
\tag{4}
$$

Здесь и далее $A \otimes B$ — кронекерово произведение матриц A и B, а $I_K \in \mathbb{R}^{K \times K}$ - единичная матрица порядка K.

Таким образом, приведенная теорема утверждает, что выходной поток из системы MAP/PH/1/N является также MAP-потоком.

Учитывая то, что на каждую фазу многофазной системы поступает MAP-поток кросс-трафика и MAP-поток с предыдущей фазы, можно доказать следующую теорему:

Теорема 2. *Композиция MAP-потоков X и Y с матрицами переходов* $D_0^{(X)},\ D_1^{(X)} \in \mathbb{R}^{M \times M}$ *и* $D_0^{(Y)},\ D_1^{(Y)} \in \mathbb{R}^{N \times N}$ *- MAP-поток* $Z = X \otimes Y$, *матрицы переходов* $D_0^{(Z)},\ D_1^{(Z)} \in \mathbb{R}^{(MN) \times (MN)}$ *которого определены как:*

$$
\begin{aligned}
D_0^{(Z)} &= I_N \otimes D_0^{(X)} + D_0^{(Y)} \otimes I_M \\
D_1^{(Z)} &= I_N \otimes D_1^{(X)} + D_1^{(Y)} \otimes I_M
\end{aligned}
\tag{5}
$$

Таким образом, приведенная теорема утверждает, что композиция (наложение) двух MAP-потоков образует MAP-поток.

Сформулированные выше теоремы будут использованы для расчета характеристик производительности многофазной системы массового обслуживания.

Далее рассмотрим метод расчета стационарных вероятностей состояний вероятности потерь, интенсивности поступления сообщений, маргинальные распределения длин очередей и других параметров системы MAP/PH/1/N.

Стационарное распределение $\overline{\theta} \in \mathbb{R}^L$ вероятностей состояний выходящего (обслуженного) MAP-потока является решением системы линейных алгебраических уравнений:

$$
\begin{cases}
\overline{\theta} D^{(o)} & = \overline{\mathbf{0}}_L \\
\overline{\theta} \overline{\mathbf{1}}_L & = 1
\end{cases}
\tag{6}
$$

Обозначим через W — число состояний входящего MAP-потока , M — число непоглощающих состояний управляющей цепи PH-распределения, N — число мест для ожидания в очереди прибора.

Для удобства примем, что $\overline{\theta} = \left(\overline{\theta}_0 \dots \overline{\theta}_{N+1}\right)$, где каждый вектор $\overline{\theta}_i \in \mathbb{R}^{WM}$ — компоненты распределения, отвечающие состоянию i, т.е. когда в системе i заявок.

Пусть $\eta \in \mathbb{R}^{N+2}$ - распределение вероятностей количества сообщений в системе. Определить значения $\eta_n, n = \overline{0, N+1}$ можно из вектора $\overline{\theta}$:

$$\eta_n = \sum_i \overline{\theta}_{n,i} \tag{7}$$

Из матриц D_0, D_1, D входящего MAP-потока можно получить стационарное распределение $\overline{\phi} \in \mathbb{R}^W$ вероятностей состояний входящего MAP-потока, решив систему линейных алгебраических уравнений:

$$\begin{cases} \overline{\phi} D & = \overline{\mathbf{0}}_W \\ \overline{\phi} \mathbf{1}_W & = 1 \end{cases} \tag{8}$$

По полученному стационарному распределению рассчитывается средняя интенсивность поступления сообщений в систему λ_{avg}:

$$\lambda_{avg} = \overline{\phi} D_1 \overline{\mathbf{1}}_{W+1} \tag{9}$$

Среднее число сообщений в системе определяется как математическое ожидание случайной величины n:

$$N_{avg} = \sum_{n=0}^{N+1} \eta_n n \tag{10}$$

Наконец, определим вероятность потери поступившего на вход прибора сообщения:

$$P_{loss} = \overline{\theta}_{N+1} \frac{D_1}{\lambda_{avg}} \overline{\mathbf{1}}_{W+1} \tag{11}$$

4. Расчет параметров тандемной системы $MAP/PH/1/N \to \bullet/PH/1/N \to \dots \to \bullet/PH/1/N$

В предыдущем разделе было описано, как строится входящий MAP-поток на i-ую фазу , а также приведены формулы для расчета различных параметров в системе MAP/PH/1/N. Ниже приводится алгоритм, позволяющий использовать эти сведения для расчета стационарных вероятностей потери сообщений, интенсивности входящих потоков и средней длины очередей на каждой фазе.

Пусть K - число фаз, $\Psi_0 = MAP\,(A_0, A_1)$ - поток, описывающий кросс-трафик, входящий в каждую фазу, $\Phi_i = MAP\left(B_0^{(i)}, B_1^{(i)}\right)$ - исходящий поток из $i-$й фазы, а $\Psi_i = MAP\left(A_0^{(i)}, A_1^{(i)}\right)$ - поток, входящий в $i-$ю фазу, $1 \le i \le K$. Пусть обслуживание в каждом приборе осуществляется распределением $\Omega = PH\,(\overline{\tau},\ S)$.

На основании теоремы 2 можно утверждать, что:

$$\begin{aligned} &\Psi_1 = \Psi_0 \\ &\Psi_i = \Phi_{i-1} \otimes \Psi_0,\ 2 \le i \le K \end{aligned} \tag{12}$$

Для вычисления вероятности потери сообщения, средней интенсивности поступления сообщений и среднего количества сообщений в системе на каждой фазе предлагается алгоритм 1, представленный в виде псевдо-кода.

В соответствии с предложенным алгоритмом был разработан программный комплекс расчета основных характеристик беспроводной сети. Основной сложностью, возникающей при расчете системы $MAP/PH/1/N \rightarrow \bullet/PH/1/N \rightarrow \cdots \rightarrow \bullet/PH/1/N$, является огромный размер матриц, описывающих MAP-потоки, поступающие на i-ую фазу, $i = \overline{1,k}$. Их сложность растет экспоненциально, поэтому для решения задач большой размерности была создана имитационная модель. Разработанное программное обеспечение позволяет получить точное аналитическое решение для тандемной системы с малым количеством фаз и MAP-потоком с матрицами небольшой размерности. Точное решение используется для калибровки имитационной модели.

5. Сравнение результатов имитационного и аналитического моделирования

Ввиду огромного размера матриц даже для MAP-потоков с тремя состояниями и PH-распределений времени обслуживания на фазах с двумя состояниями, точное решение было получено для двух случаев: простейшей тандемной системы $M/M/1/N \rightarrow \bullet/M/1/N \rightarrow \cdots \rightarrow \bullet/M/1/N$ с пуассоновским кросс-трафиком, а также для системы с входящим MAP-потоком с двумя состояниями. Параметры MAP-потока были получены из реальной статистики, собранной на автомобильных дорогах города Москвы в разное время суток.

На рис. 2 приведена плотность распределения вероятностей интервалов между автомобилями, полученная из собранной статистики, и плотность распределения вероятностей MAP-потока с матрицами

$$D_0 = \begin{pmatrix} -0.85 & 0.85 & 0.0 \\ 0.0 & -1.1 & 0.2 \\ 0.0 & 0.5 & -4.0 \end{pmatrix},\ D_1 = \begin{pmatrix} 0.0 & 0.0 & 0.0 \\ 0.9 & 0.0 & 0.0 \\ 3.5 & 0.0 & 0.0 \end{pmatrix}$$

MAP-поток аппроксимирует статистику по первым двум моментам.

Data: K — число фаз, N — число мест, $\Psi_0 = MAP\,(A_0, A_1)$ — кросс-трафик, $\Omega = PH\,(\overline{\tau},\ S)$ — распределение времени обслуживания

Result: $P_{loss}^{(i)}$ — вероятность потери сообщения на i-й фазе, $\lambda^{(i)}$ — средняя интенсивность поступления сообщений на i-ю фазу, $N_{avg}^{(i)}$ — среднее число заявок на i-й фазе

$i := 1$;
while $i \leq K$ **do**
  if $i = 1$ **then**
    $A_0^{(i)} := A_0$, $A_1^{(i)} := A_1$;
  else
    /* $\Psi_i = \Phi_{i-1} \otimes \Psi_0$ */
    вычислить $A_0^{(i)}, A_1^{(i)}$ с помощью (5) из $\Psi_0 = MAP\,(A_0, A_1)$ и $\Phi_i = MAP\left(B_0^{(i)}, B_1^{(i)}\right)$;
  end
  вычислить $\Phi_i = MAP(B_0^{(i)}, B_1^{(i)})$ с помощью (3);
  вычислить с помощью (1) $\overline{\pi}^{(i)} = \left(\overline{\pi}_0^{(i)}, \ldots, \overline{\pi}_{N+1}^{(i)}\right)$ – стационарные вероятности потока Φ_i;
  вычислить с помощью (1) $\overline{\alpha}^{(i)}$ – стационарные вероятности потока Ψ_i;
  вычислить $\lambda^{(i)}$, используя вектор $\overline{\alpha}$, с помощью (9);
  вычислить $N_{avg}^{(i)}$, используя вектор $\overline{\pi}$, с помощью (10);
  вычислить $P_{loss}^{(i)}$, используя вектор $\overline{\pi}$, с помощью (11);
  $i := i + 1$;
end

Algorithm 1: Алгоритм расчета параметров тандемной системы

В обоих случаях с помощью имитационной модели и точного аналитического расчета были вычислены вероятности потери пакетов P_{loss} на каждом приборе, а также вероятности занятости прибора P_{busy}.

Для системы $MAP/M/1/N \to \bullet/M/1/N \to \cdots \to \bullet/M/1/N$ с кросс-трафиком интенсивности входящих потоков давались матрицами $D_0,\ D_1$:

$$D_0 = \begin{pmatrix} -1 & 1 \\ 0 & -1 \end{pmatrix},\ D_1 = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix},$$

что соответствует эрланговскому распределению с двумя фазами и интенсивностями, равными единице, интенсивность обслуживания равнялась $\mu = 2.0$, длина очереди $N = 2$, а число станций $Q = 4$. Результаты сравнения приведены в таблице 1.

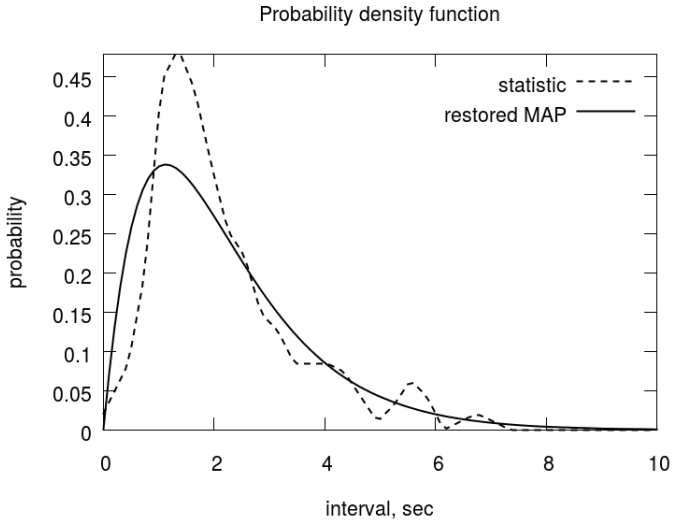


Рис. 2: MAP-поток и статистическое распределение интервалов

Номер прибора	P_{loss}^{sim}	$P_{loss}^{analytic}$	P_{busy}^{sim}	$P_{busy}^{analytic}$
1	0.0018	0.0019	0.2503	0.2495
2	0.0363	0.0365	0.4813	0.4813
3	0.1138	0.1139	0.6469	0.6480
4	0.1797	0.1815	0.7336	0.7350

Таблица 1: Сравнение характеристик системы: P_{loss}^{sim} и $P_{loss}^{analytic}$ – стационарные вероятности потери пакета, полученные из имитационной и аналитической моделей соответственно; P_{busy}^{sim} и $P_{busy}^{analytic}$ – стационарные вероятности занятости прибора.

Из таблицы видно, что результаты аналитического и имитационного моделирования совпадают с высокой точностью.

6. Расчет характеристик широкополосной сети большой размерности

В отличие от аналитического подхода, имитационное моделирование позволяет получить результаты для тандемной системы $MAP/PH/1/N \rightarrow \bullet/PH/1/N \rightarrow \cdots \rightarrow \bullet/PH/1/N$ с кросс-трафиком при большом количестве фаз. Все поступающие на вход потоки описывались MAP-потоком, матрицы которого были приведены выше. Число станций в тандеме $Q = 20$,

а максимальная длина очереди $N = 10$. Такие ограничения на практике встречаются, если распределенная система работает на очень слабых встроенных платформах – программные компоненты имеют буферы ограниченного размера, и при их переполнении новые сообщения отбрасываются. В случае малого объема доступной памяти буферы некоторых приложений ограничиваются местами под десять сообщений.

Обслуживание моделировалось с помощью экспоненциальных распределений с различными интенсивностями. Рассматривались случаи

$$\mu \in \{2, 5, 10, 20, 100, 1000\},$$

начиная с быстрой сети и аппаратуры ($\mu = 1000$), кончая совсем медленной ($\mu = 2$). Последний случай оказывается релевантным, поскольку при плохих телекоммуникационных каналах между станциями и частых простоях средне-дневная скорость обработки заявок значительно деградирует.

На рис. 3 приведена средняя межконцевая задержка для каждой станции: на оси абсцисс указан номер станции, которая производит первую передачу сообщения, на оси ординат - стационарное время, через которое сообщение успешно покинет сеть (будет обслужен последней станцией тандема). Из рисунка видно, что начиная со значения $\mu = 100$ задержка укладывается в 200 милисекунд, однако уже при $\mu = 20$ удается передать данные от самой удаленной станции за время порядка 1.4 секунды.

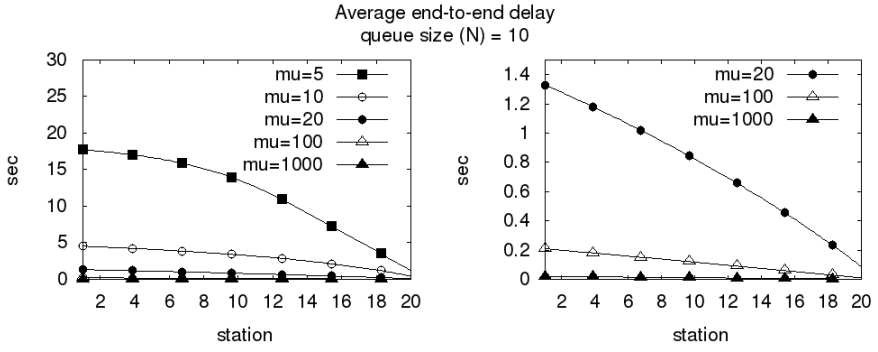


Рис. 3: стационарная межконцевая задержка

На рис. 4 приведена средняя вероятность потери сообщения на каждой станции. Из графика видно, что начиная с $\mu = 20$ потерь практически не происходит. Однако, ввиду случайного характера трафика, при большем числе соединенных станций желательно иметь более высокую производительность.

На рис. 5 приведена вероятность потери сообщения, передаваемого заданной станцией, на пути до последней станции. Очевидно, что эта вели-

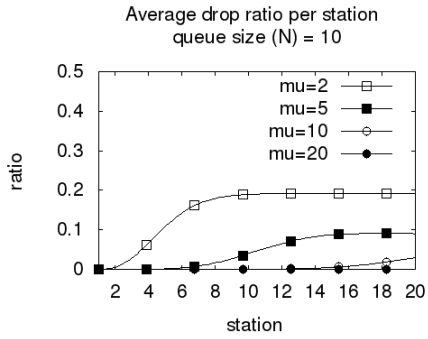


Рис. 4: стационарная вероятность потери сообщения на каждой станции

чина является невозрастающей относительно номера станции, а начиная с $\mu = 20$ равна нулю практически всюду.

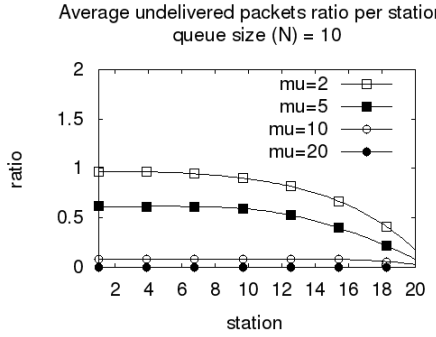


Рис. 5: стационарная вероятность неуспешной доставки сообщения

На рис. 6 показаны стационарные средние длины очередей для каждой станции. Поскольку на каждой станции добавляется кросс-трафик, интенсивность входящих потоков не убывают относительно номера станции, а поскольку во всех приборах время обслуживания распределено одинаково, стационарные средние длины очередей оказываются неубывающими функциями номеров станций. Из рисунка видно, что чем выше интенсивность обслуживания, тем позже "насыщается" очередь. В частности, для случая $\mu = 20$ очередь каждой станции оказывается большую часть времени свободной.

Наконец, на рис. 7 и рис. 8 показаны распределения задержек в обслуживании сообщений на каждой станции и распределения интервалов между исходящими сообщениями на каждой станции. Графики приведе-

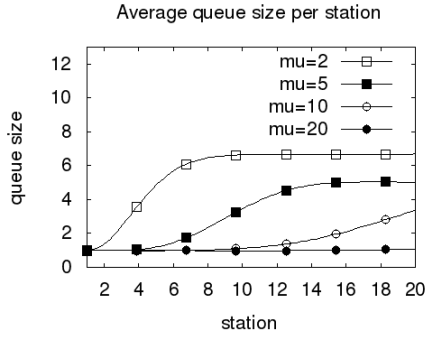


Рис. 6: стационарные средние длины очередей

ны для различных интенсивностей обслуживания и нескольких станций. В частности, из графика 8 видно, что, начиная с некоторой фазы, зависящей от интенсивности обслуживания, исходящие потоки из приборов становятся практически неразличимы.

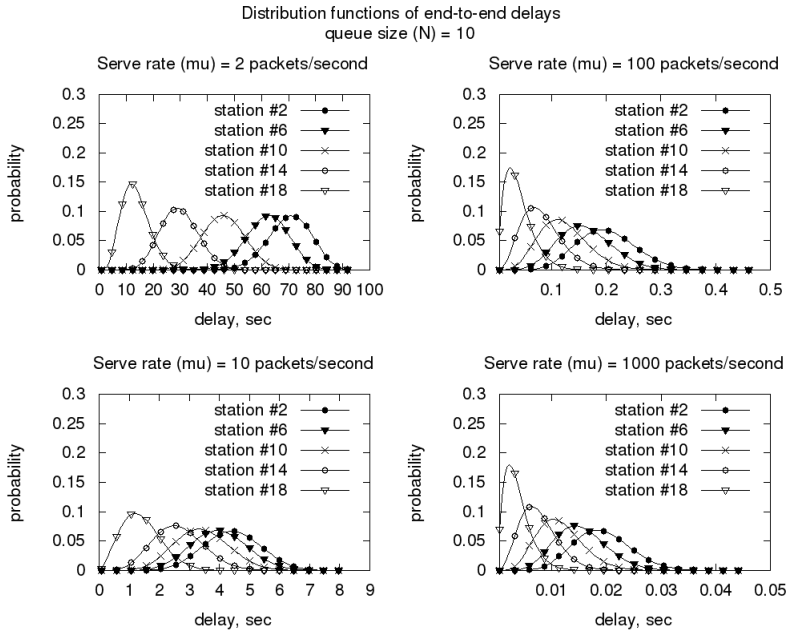


Рис. 7: распределение длительностей пребывания в системе

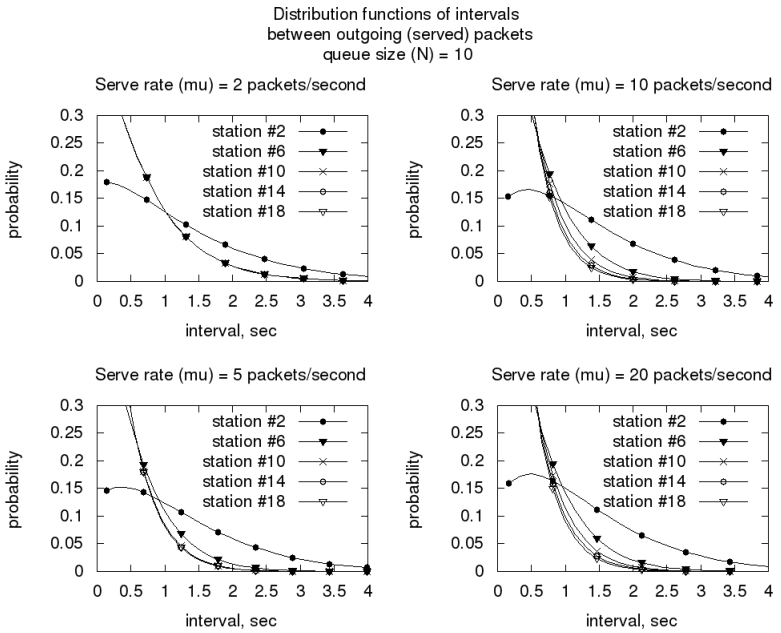


Рис. 8: распеределения интервалов между исходящими (обслуженными) сообщениями

Как видно, при низкой интенсивности обслуживания сообщений система быстро деградирует. Это связано, в первую очередь, с наличием кросс-трафика: каждая последующая базовая станция не только вынуждена обрабатывать все, что передано от предыдущей станции, но и сообщения кросс-трафика. В частности, из приведенных результатов следует, что для эффективной работы системы на протяженных магистралях длиной в несколько сот километров и несколькими десятками ретрансляторов, следует уделять внимание качеству линий связи, соединяющих станции. Для стабильной работы требуется, в среднем, обслуживать от 20 сообщений в секунду. Желательно иметь производительность порядка ста сообщений в секунду.

7. Заключение

Разработан новый подход к оценке основных характеристик многофазных стохастических систем $MAP/PH/1/N \to \bullet/PH/1/N \to \cdots \to \bullet/PH/1/N$ с входным MAP-потоком, кросс-трафиком и PH-распределением времени обслуживания, адекватно описывающих функционирование широкополосных беспроводных сетей вдоль протяженных транспортных магистралей. Предложен аналитический алгоритм для расчета характеристик произво-

дительности сетей небольшой размерности и имитационная модель, обеспечивающая исследование сетей с большим количеством базовых станций.

Разработанные методы расчета характеристик многофазных стохастических систем эффективно использовались при проектировании и реализации широкополосной беспроводной сети вдоль окружной автомобильной дороги города Казань (М7 Волга).

ЛИТЕРАТУРА

1. 802.11-2012 IEEE Standard for Information technology. Telecommunications and information exchange between Local and metropolitan area networks. Specific requirements Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications. — IEEE Std., March 2012.
2. Vishnevsky V.M., Portnoi S.L., Shakhnovich I.V. WiMAX Encyclopaedia. Way to 4G. Tekhnosfera, Moscow, 2010. 470 p. (in Russian)
3. Vishnevsky V.M., Semenova O.V. Polling Systems: Theory and Applications for Broadband Wireless Networks. London: Academic Publishing, 2012. 317p.
4. Wu Q., Zheng J. Performance modeling and analysis of IEEE 802.11 DCF based fair channel access for vehicle-to-roadside communication in a non-saturated state // Springer Wireless Networks, Volume 21, Issue 1, 2014. Pp. 1–11.
5. Chakraborty S., Nandi S. IEEE 802.11s Mesh Backbone for Vehicular Communication: Fairness and Throughput // IEEE Transactions on Vehicular Technology, vol. 62, issue 5, 2013. Pp. 2193 – 2203.
6. Campolo C., Cozzetti H.A., Molinaro A., Scopigno R. Augmenting Vehicle-to-Roadside connectivity in multi-channel vehicular Ad Hoc Networks // Journal of Network and Computer Applications, vol. 36, issue 5, 2013. Pp. 1275–1286.
7. В.М. Вишневский, А.А. Ларионов, О.В. Семенова. Оценка производительности высокоскоростной беспроводной тандемной сети с использованием каналов сантиметрового и миллиметрового диапазона радиоволн в системах управления безопасностью дорожного движения // Проблемы управления. 2013. No.4. С.50–56.
8. V. Vishnevsky, O. Semenova, A. Dudin, V. Klimenok. Queueing model with gated service and adaptive vacations // Communications Workshops, 2009. ICC Workshops 2009. IEEE International Conference on. — June 2009. — P.5.
9. V. Klimenok, A. Dudin, V. Vishnevsky. Tandem queueing system with correlated input and cross-traffic // Computer Networks, ser. Communications in Computer and Information Science. Springer Berlin Heidelberg, 2013. — Vol.370. — Pp.416–425.

10. V. M. Vishnevsky, A. N. Dudin, O. V. Semenova, V. I. Klimenok. Performance analysis of the BMAP/G/1 queue with gated servicing and adaptive vacations. — Performance Evaluation. — Vol. 68, No. 5. — May 2011. — Pp.446–462.
11. V. Klimenok, A. Dudin, V. Vishnevsky. On the stationary distribution of tandem queue consisting of a finite number of stations. Ц Computer Networks, ser. Communications in Computer and Information Science. — Vol. 291. — Springer Berlin Heidelberg, 2012. — Pp.383–392.

STABILITY AND ADMISSIBLE DENSITIES IN TRANSPORTATION FLOW MODELS

A. A. Lykov[1], *V. A. Malyshev*[2], *M. V. Melikian*[3]
Moscow State University, Faculty of Mechanics and Mathematics
[1]alekslyk@yandex.ru, [2]2malyshev@mail.ru, [3]magaarm@list.ru

Abstract

One-way road traffic model with a local control is considered. For given density of transportation units we discover phase transitions in the control parameter space when there exists or not safe transportation. Also we discuss possible densities and instability in several road networks with crosses, if no control is imposed.

1. Introduction

Theoretical modelling and computer simulation of transportation systems is a very popular field, see very impressive review [3]. There are two main directions in this research - macro and micro models. Macro approach does not distinguish individual transportation units and uses analogy with the notion of flow in hydrodynamics, see [2]. Stochastic micro models are most popular and use almost all types of stochastic processes: mean field, queueing type and local interaction models. We consider here deterministic transportation flows. Although not as popular as stochastic traffic, there is also a big activity in this field, see [1, 4, 5, 6, 7, 8].

In section 2 we do not pursue maximal generality but rather consider the simplest flow, that is the following one-way road traffic model. Namely, at any time $t \geq 0$ there are many (even infinite) number of point particles (may be called also cars, units etc.) with coordinates $z_k(t)$ on the real line, enumerated as follows

$$... < z_N(t) < ... < z_1(t) < z_0(t) \tag{1}$$

For this infinite chain of cars we try to find control mechanism which garanties that the distance between any pair of neighbouring cars is greater or equal (on all time interval $(0, \infty)$) to some fixed number d, called safe distance, but at the same time is not too far from it (then we say that the density of cars is admisible).

This control mechanism is assumed to be local - any car has information only about the previous car. Moreover, this mechanism is of physical nature, like forces between molecules in crystals but our "forces" are not symmetric. The safety (stability) conditions appear to be similar to the dynamical phase transition in the model of the molecular chain rapture under the action of external force. However here we do not need the double scaling limit, used in [9].

In section 3 we consider idealistic models of transportation network as the collection of roads with crosses (intersections). Along each road the units (cars) move deterministically. We get admissible density of units on the given transportation network and show the dependence of this density on the number and multiplicity of crosses and number of cycles in the network.

2. Local flow control

In this section we consider the simplest flow (1), where it is assumed that the rightmost unit moves "as it wants". More exactly, the trajectory $z_0(t)$ is only assumed to be sufficiently smooth with positive velocity $v_0(t) = \dot{z}_0(t)$ and natural upper bounds on the velocity and acceleration

$$\sup_{t>0} v_0(t) \leqslant v_{max}, \quad \sup_{t>0} |\ddot{z_0}(t)| \leqslant a_{max} \tag{2}$$

We would like to organize such traffic so that for any t and k the distances $r_k(t) = z_{k-1}(t) - z_k(t)$ were greater or equal to some number $d > 0$, which is chosen to avoid collisions and keep maximal possible density of traffic. Moreover, the organization should use only (maximally) local control. More exactly, the k-th driver at any time t knows only its own coordinate and velocity and the coordinate $z_{k-1}(t)$ of the previous car. Thus, for any $k \geq 1$ the trajectory $z_k(t)$ is uniquely defined by the trajectory $z_{k-1}(t)$ of the previous particle.

Using physical terminology one could say that if, for example, $r_k(t)$ becomes larger than d, then some positive force F_k increases acceleration of the particle k. We will see that, besides F_k, for such stability, also friction force $-\alpha v_k(t)$, which, on the contrary, restrains the growth of the velocity $v_k(t)$, is necessary. The constant $\alpha > 0$ should be chosen appropriately.

Thus the trajectories are uniquely defined by the system of equations for $k \geq 1$

$$z_k^{''}(t) = F_k(t) - \alpha \frac{dz_k}{dt} = \omega^2(z_{k-1}(t) - z_k(t) - d) - \alpha \frac{dz_k}{dt} \tag{3}$$

where F_k is taken to be simplest possible

$$F_k(t) = \omega^2(z_{k-1}(t) - z_k(t) - d)$$

Stability conditions For given $\alpha, \omega, d, z_0(t)$ and initial conditions the trajectories are uniquely defined and we can denote

$$I = \inf_{k \geqslant 1} \inf_{t \geqslant 0} r_k(t), \quad S = \sup_{k \geqslant 1} \sup_{t \geqslant 0} r_k(t)$$

Put also

$$d^* = d^*(v_{max}, a_{max}) = \frac{1}{\omega^2}(a_{max} + \alpha v_{max})$$

Consider firstly the simplest initial conditions

$$z_k(0) = -kd, \frac{dz_k}{dt}(0) = v, k \geq 0 \tag{4}$$

Theorem 1. *Assume (4) and $\alpha > 2\omega > 0$. Then for any chosen "safe distance" parameter $d > d^*$ in the equations (3) the following bounds hold*

$$I > (d - d^*) > 0, \quad S < 2d.$$

Now consider more general initial conditions satisfying

$$z_0(0) = 0, z_0'(0) = v, 0 < a \leqslant r_k(0) \leqslant b, \quad |\frac{dr_k}{dt}(0)| \leqslant c, \tag{5}$$

for any $k > 0$ and some non-negative a, b. Denote also

$$D^* = \max\{A, d^*\},$$

where

$$A = \frac{\alpha a^{'} + 2c}{2\gamma}, \gamma = \sqrt{\frac{\alpha^2}{4} - \omega^2}, a^{'} = \max\{|a - d|, |b - d|\}.$$

Theorem 2. *If $\alpha > 2\omega$, $\frac{\alpha a - 2c}{\alpha - 2\gamma} > \frac{a+b}{2}$, $\frac{a+b}{2} < d^* < \frac{\alpha a - 2c}{\alpha - 2\gamma}$ then for any initial conditions (5) and any smooth function $z_0(t)$, satisfying (2), there exists open subset $\mathcal{D} \subset \mathbb{R}$ such that for any choosen "safe distance" parameter $d \in \mathcal{D}$ in the equaitons (3), the following bounds hold*

$$I \geq (d - D^*) > 0, \quad S \leq d + D^*.$$

The proof of both theorems is based on the analysis of the chain of equations for $x_k = r_k - d, k > 0$

$$x_k^{''}(t) + \alpha x_k^{'}(t) + \omega^2 x_k(t) = \omega^2 x_{k-1}(t), \; k = 1, 2, ...$$

where

$$x_k(0) = 0, \; x_k^{'}(0) = 0, \; x_0(t) = \frac{1}{\omega^2}(z_0^{''}(t) + \alpha z_0^{'}(t))$$

Density and currents Let $n(t, I)$ be the number of units on the interval $I \subset R$ at time t and $n(T, x)$ be the number of units passing the point x in the time interval $(0, T)$. Then the density and the current through some point x are defined as follows ($|I|$ is the length of I)

$$\mu(t) = \lim_{|I| \to \infty} \frac{n(t, I)}{|I|}$$

$$J(x) = \lim_{T \to \infty} \frac{n(T, x)}{T}$$

if these limits exist. For fixed N define also mean length of the chain of cars $0, 1, ..., N$

$$L_N(t) = \frac{z_0(t) - z_N(t)}{N}.$$

Theorem 3. *Under the conditions of theorem 2 assume also that the initial conditions are such that the following finite limits exist*

$$\lim_{N\to\infty} L_N(0) = L(0), \quad \lim_{N\to\infty} \dot{L}_N(0) = \dot{L}(0),$$

Then for any t there exist

$$\lim_{N\to\infty} L_N(t) = L(t),$$

and moreover

$$L(t) = L(0) + \frac{1}{\alpha}(1 - e^{-\alpha t})\frac{dL}{dt}(0)$$

Note that if moreover $\frac{dz_k(0)}{dt}$ are uniformly bounded then $\frac{dL}{dt}(0) = 0$, that is the mean length does not change with time.

Instability First reason for the instability is the absence of dissipation.

Theorem 4. *Let $\alpha = 0$, then for the initial conditions (4) for $k > 0$ and for $z_0(t) = tv + \sin\omega' t,\ v > 0,\ \omega' \neq 0$ and $k \geqslant 2$ we have*

$$\inf_{t\geqslant 0} r_k(t) = -\infty$$

It is sufficiently easy to explain by the resonance effect: $x_1(t)$ is the harmonic movement with frequencies ω, ω', and the proper frequency of x_2 is ω.

Assume again (4). We will prove instability (even for rather simple behaviour of the leading unit $z_0(t)$) when k and t tend simultaneously to ∞ so that $t = \mu k,\ k \to \infty$ for some constant $\mu > 0$. The following theorem consists of two parts: the first one exhibits the possible zero density, the second one exhibits the possibility of collisions.

Theorem 5. *Let $\frac{dz_0}{dt}(t) = v$ for all $t \geqslant 0$. Then*

1. for any $\alpha > 0$ there exists $\omega > 0$ and constants $q_+ > 1,\ \mu_+ > 0, c_+ > 0$, so that for any $d > 0$

$$r_k(t) \sim \frac{c_+}{\sqrt{k}} q_+^k$$

as $t = \mu_+ k,\ k \to \infty$;

2. for any $\alpha > 0$ there exists $\omega > 0$ and constants $q_- > 1,\ \mu_- > 0, c_- < 0$, so that for any $d > 0$

$$r_k(t) \sim \frac{c_-}{\sqrt{k}} q_-^k$$

as $t = \mu_- k,\ k \to \infty$.

3. Transport in networks without control

Transportation without control means that any unit moves with its own velocity and does not know anything about the movement of other units. In other words the functions $z_k(t)$ are fixed. We will consider only the case when

$$z_k(t) = c_k + v_k t$$

for some constants $v_k > 0$ and

$$... < c_{k-1} < c_k = z_k(0) < ...$$

The transportation network is defined to be safe if at any time moment t the distance between any two units is greater or equal to some fixed number $d > 0$ called safety distance. It is clear that for the safe network all v_k should be equal to some $v > 0$, their joint constant velocity. It is also clear that $\mu(t) \geq d^{-1}$ and this density is attained in the case when

$$z_i(t) = z_i(0) + vt, \;\; z_{i-1}(0) - z_i(0) = d \implies z_{i-1}(t) - z_i(t) = d$$

It is easy to see that in our case at time t any density $\mu(t)$ is admissible iff

$$\mu(t) \leq d^{-1}$$

The network (of roads) is defined here as the one-dimensional topological space, which is the disjoint union of some number of real lines (roads) $k = 1, ..., M \leq \infty$ with coordinates z_k. It is assumed that any pair of real lines has finite number of identified points (crosses). Consider the graph G, which vertices are these identified points (intersections of the real lines) and the edges are all segments of the roads in-between the vertices. That is we delete all infinite intervals of the roads. The metrics is defined as usual: the distance $\rho(x_k, x_l)$ between two points x_k and x_l on the network is the minimal length of paths between these two points.

On any road k the transportation units are labeled as (k, i), where k is the road along which it moves and i is its order on this road. Here the sequences are assumed to be infinite to both sides

$$... < z_{k,i}(t) < z_{k,i-1}(t) < ...$$

and we again consider networks where the velocities and safe distance are the same for all roads

Proposition 1. *1. Assume there are $M < \infty$ roads with only one common cross. In this case the graph G consists of one vertex only. Then at time t any density $\mu(t)$ is admissible iff*

$$\mu(t) \leq (Md)^{-1} \tag{6}$$

2. Assume there are M roads and let L be maximal multiplicity of their intersections. Assume also that the graph G has no cycles. Then at time t any density $\mu(t)$ is admissible iff

$$\mu(t) \leq (Ld)^{-1}$$

3. Consider 4 roads A, B, C, D with 4 different crosses $1 = A \cap B, 2 = B \cap C, 3 = C \cap D, 4 = D \cap A$, Then the graph G is a quadruple and has 4 edges $12, 23, 34, 41$. Then the admissible densities are

$$\mu(t) \leq \frac{3}{8}(d)^{-1}$$

Note that if the number of cycles grows then the coefficient in front of d^{-1} decreases very quickly. It would be interesting to calculate it explicitely.

There is another reason showing that such transportation network is not only of small density but also is strongly unstable. Namely, consider the case when each road $k = 1, ..., M$ has its own safe distance d_k and its own velocity v_k, Thus

$$z_{k,i}(t) = z_{k,i}(0) + v_k t, \;\; z_{k,i-1}(0) - z_{k,i}(0) = d_k \Longrightarrow z_{k,i-1}(t) - z_{k,i}(t) = d_k$$

Then the corresponding currents are $J_k = v_k d_k^{-1}$. The following proposition shows extreme instability of the simplest network.

Proposition 2. *Consider two roads with one cross. Call the transportation stable if distances between any two units are greater or equal to some $D_0 > 0$ uniformly in t. Then:*

If $\frac{J_1}{J_2}$ is not rational, then such transportation cannot be stable.

If $\frac{J_1}{J_2} = \frac{n_1}{n_2}$ for some integers n_1, n_2 such that $(n_1, n_2) = 1$. Then stable transportation exists iff

$$\frac{n_1}{d_1} + \frac{n_2}{d_2} < \frac{1}{D_0}$$

Remark 1. *(Control types)*

To avoid instability of Proposition 2 some control is necessary. There are two possibilities for the control. The first one we considered in section 2 - local internal control, that is depending only on distances between cars. Second type is the control which forces velocities of cars to change in certain points of the network. The mostly used such control is the organization of traffic lights where the cars should stand still for some time. Other control types are also known, see [1, 4, 5, 6, 7, 8], and it could be interesting to find general classification of control types.

REFERENCES

1. M. Blank. Ergodic properties of a simple deterministic traffic flow model. J. Stat. Phys., 2003, v. 111, 903-930.

2. Prigogine I., Herman R. Kinetic theory of vehicular traffic. N.Y.: Elsevier, 1971.
3. Helbing D. Traffic and related self-driven many particle systems. Rev. Mod. Phys. 73, 1067-1141 (2001).
4. Feintuch A., Francis B. Infinite chains of kinematic points. Automatica 48 (2012) 901-908.
5. Qing Hui, Jordan M. Berg. Semistability theory for spatially distributed systems. Proceedings of the IEEE Conference on decision and control, January 2009.
6. M.R. Jovanovic, B. Bamieh. On the Ill-Posedness of certain vehicular platoon control problems. IEEE transactions on automatic control, Vol.50, NO.9, September 2005.
7. S.M. Melzer, B.C. Kuo. Optimal regulation of systems described by a countably infinite number of objects. Automatica, Vol. 7, pp. 359-366. Pergamon Press, 1971.
8. D. Swaroop, J.K. Hedrick. String stability of interconnected systems. IEEE transactions on automatic control, Vol.41, NO.3, March 1996.
9. Malyshev V.A., Musychka S.A. Dynamical phase transition in the simplest molecular chain model. Theoretical and mathematical physics, 2014, v. 179, No. 1, 123-133.

A NUMBER OF CUSTOMERS IN THE CYCLIC QUEUEING SYSTEM

S. Paul, A. Nazarov
Tomsk State University, Tomsk, Russia
paulsv82@mail.ru, nazarov.tsu@gmail.com

Abstract

We review the queueing system with one service device and a queue with an unlimited number of waiting seats with server vacations. By using method of asymptotic analysis under large load, we found asymptotic distribution of probabilities of a number of customers in the system. It is shown that this probability distribution can be approximated by exponential distribution.

Keywords: queueing system with server vacations, asymptotic analysis, exponential distribution

1. Introduction

Single-line queueing systems with server vacations are mathematical models of telecommunication systems, which are pretty common in practice [1]. In real systems "vacations" are considered as a temporal suspension of service either for device other applications or for its breakdown or repair [2].

2. Mathematical Model

Let's review the queueing system with one service device and a queue with an unlimited number of waiting seats [3]. The system receives Poisson process of requests with intensity λ. Device operation mode consists of two consecutive intervals. During first interval's customers are handled at the device for random time, distributed by exponential law with parameter μ. If there are no requests in the queue at the beginning of this interval or if the device has handled all customers that were in the queue at that interval of time, then device is still operating in the same mode, waiting for requests. When this interval ends, the server goes on a vacation. During vacations, all requests that came into system are gathered in the queue and are waiting for device to return to operating mode.

Durations of these intervals are random and determined by distribution functions $T_1(x)$ and $T_2(x)$ respectively. We will review systems with customers priority service [4].

Let's denote:

$i(t)$ is number of customers in the system at the time t.

$k(t)$ is device mode: 1 – device is on the service, 2 – device is on vacations.

$z(t)$ is remaining time of device staying in corresponding mode.

3. Kolmogorov Equations

Let's review three-dimensional Markov process $\{i(t), k(t), z(t)\}$ and for distribution of probabilities

$$P_k(i, z, t) = P\{i(t) = i, k(t) = k, z(t) < z\}$$

Let is set up the direct system of Kolmogorov's differential equations. We will assume that system operates in stationary mode:

$$\begin{cases} \frac{\partial P_1(i,z)}{\partial z} - \frac{\partial P_1(i,0)}{\partial z} + \\ (P_1(i-1,z) - P_1(i,z))\,\lambda + (P_1(i+1,z) - P_1(i,z))\,\mu + \frac{\partial P_2(i,0)}{\partial z} T_1(z) = 0, \\ \frac{\partial P_2(i,z)}{\partial z} - \frac{\partial P_2(i,0)}{\partial z} + \\ \lambda\,(P_2(i-1,z) - P_2(i,z)) + \frac{\partial P_1(i,0)}{\partial z} T_2(z) = 0. \end{cases}$$

Let's introduce partial characteristic functions $H_k(u, z) = \sum\limits_{i=0}^{\infty} e^{jui} P_k(i, z)$, for which we will rewrite the direct system of Kolmogorov's differential equations in this form:

$$\begin{cases} \frac{\partial H_1(u,z)}{\partial z} - \frac{\partial H_1(u,0)}{\partial z} + \left(\lambda\left(e^{ju} - 1\right) + \mu\left(e^{-ju} - 1\right)\right) H_1(u, z) + \\ + \frac{\partial H_2(u,0)}{\partial z} T_1(z) + \left(1 - e^{-ju}\right)\mu P_1(0, z) = 0, \\ \frac{\partial H_2(u,z)}{\partial z} - \frac{\partial H_2(u,0)}{\partial z} + \lambda H_2(u, z)\left(e^{ju} - 1\right) + \frac{\partial H_1(u,0)}{\partial z} T_2(z) = 0. \end{cases} \quad (1)$$

$P_1(0, z)$ is probability of situation where device stays in the service mode, and there are no customers in the system.

We will solve system (1) by using method of asymptotic analysis in conditions of large system load [5]–[7].

4. Method of Asymptotic Analysis

Let's denote:

$$\lambda = (1 - \varepsilon)S\mu,\ u = \varepsilon w,$$

$$H_k(u, z) = F_k(w, z, \varepsilon), P_1(0, z) = \varepsilon\pi_1(z, \varepsilon) \quad (2)$$

S is the system's load. Its value will be found below. Then we will rewrite system (1) in this form:

$$\begin{cases} \frac{\partial F_1(w,z,\varepsilon)}{\partial z} - \frac{\partial F_1(w,0,\varepsilon)}{\partial z} + \\ + \left((1 - \varepsilon)\,S\mu\left(e^{j\varepsilon w} - 1\right) + \mu\left(e^{-j\varepsilon w} - 1\right)\right) F_1(w, z, \varepsilon) + \\ + \left(1 - e^{-j\varepsilon w}\right)\mu\varepsilon\pi_1(z, \varepsilon) + \frac{\partial F_2(w,0,\varepsilon)}{\partial z} T_1(z) = 0, \\ \frac{\partial F_2(w,z,\varepsilon)}{\partial z} - \frac{\partial F_2(w,0,\varepsilon)}{\partial z} + \\ + (1 - \varepsilon)\,S\mu F_2(w, z, \varepsilon)\left(e^{j\varepsilon w} - 1\right) + \frac{\partial F_1(w,0,\varepsilon)}{\partial z} T_2(z) = 0. \end{cases} \quad (3)$$

Theorem 1. *The limit value for* $\varepsilon \to 0 \quad F_k(w) \quad = \quad F_k(w, \infty)$ *solutions* $F_k(w, z, \varepsilon)$ *of the system (3) has the form*

$$F_k(w, z) \quad = \quad R_k(z)\Phi(w),$$

where

$$R_1 = S, \quad R_2 = 1 - S, \quad S = \frac{T_1}{(T_1 + T_2)},$$

and asymptotic characteristic function $\Phi(w)$ *is defined by*

$$\Phi(w) = \frac{S}{S - jw\{S + \Delta\}},$$

the value of Δ *is*

$$\Delta = -R_1 R_2 \mu \frac{T_1 T_2}{T_1 + T_2} + R_1^2 R_2 \mu \frac{T_2^{(2)}}{2T_2} + R_2^2 R_1 \mu \frac{T_1^{(2)}}{2T_1}, \tag{4}$$

where $\int\limits_0^\infty (1 - T_k(x))dx = T_k$ *is the average time of staying in the corresponding mode.* $T_k^{(2)}$ *is the second initial moment of the time of device's staying in mode* k*.*

Prelimit characteristic function $H(u)$ *the number of customers* $i(t)$ *can be written as*

$$H(u) = \frac{d}{d - ju},$$

where

$$d = \frac{S}{S + \Delta}. \tag{5}$$

I.e in the form of the characteristic function of exponential distribution the parameter d*.*

Proof. We perform the proof by three stages.

Stage 1. By tending ε to zero $\varepsilon \to 0$ in (3) we will obtain:

$$\begin{cases} \frac{\partial F_1(w,z)}{\partial z} - \frac{\partial F_1(w,0)}{\partial z} + \frac{\partial F_2(w,0)}{\partial z} T_1(z) = 0, \\ \frac{\partial F_2(w,z)}{\partial z} - \frac{\partial F_2(w,0)}{\partial z} + \frac{\partial F_1(w,0)}{\partial z} T_2(z) = 0. \end{cases}$$

We will seek solution for this system in this form

$$F_k(w, z) \quad = \quad R_k(z)\Phi(w).$$

We will obtain following system, solution for which we will write down in this form:

$$\begin{cases} R_1(z) = \int\limits_0^z (R_1'(0) - R_2'(0)T_1(x))dx, \\ R_2(z) = \int\limits_0^z (R_2'(0) - R_1'(0)T_2(x))dx. \end{cases}$$

$\{R_1(z), R_2(z)\}$ is a two-dimensional distribution of device mode and value of remaining time of device staying in that mode.

By tending z to infinity $z \to \infty$, considering that $T(\infty) = 1$, we will obtain:

$$\begin{cases} R_1(\infty) = \int\limits_0^\infty (R_1'(0) - R_2'(0)T_1(x))dx, \\ R_2(\infty) = \int\limits_0^\infty (R_2'(0) - R_1'(0)T_2(x))dx. \end{cases}$$

For improper integral to be convergent it is necessary that the following condition is met $R_1'(0) - R_2'(0)T_1(\infty) = 0$, then we will obtain

$$R_1'(0) = R_2'(0) = R'(0).$$

We have

$$R_k(z) = R'(0) \int\limits_0^z (1 - T_k(x))dx.$$

Let's denote $\int\limits_0^\infty (1 - T_k(x))dx = T_k$ as the average time of staying in the corresponding mode.

Then we will obtain

$$R'(0) = \frac{1}{(T_1 + T_2)}.$$

Then

$$\begin{cases} R_1(\infty) = \frac{1}{T_1+T_2} \int\limits_0^\infty (1 - T_1(x))dx = \frac{T_1}{T_1+T_2}, \\ R_2(\infty) = \int\limits_0^\infty (R_2'(0) - R_1'(0)T_2(x))dx = \frac{T_2}{T_1+T_2}. \end{cases}$$

Stage 2. We will substitute the following expansion in the system (3).

$$F_k(w, z, \varepsilon) = \Phi(w) \{R_k(z) + j\varepsilon w f_k(z)\} + O(\varepsilon^2). \tag{6}$$

We will expand the exponent in a row and after simple transformations we will obtain

$$\begin{cases} j\varepsilon w \left(f_1'(z) - f_1'(0) + f_2'(0)T_1(z)\right) + jw\varepsilon \left(S - 1\right) \mu R_1(z) + O(\varepsilon^2) = 0, \\ j\varepsilon w \left(f_2'(z) - f_2'(0) + f_1'(0)T_2(z)\right) + j\varepsilon w S \mu R_2(z) + O(\varepsilon^2) = 0. \end{cases}$$

By dividing both equations by ε and be tending ε to zero $\varepsilon \to 0$, we will obtain:

$$\begin{cases} f_1(z) = \int\limits_0^z \{f_1'(0) - f_2'(0)T_1(x) + R_1(x)\mu(1-S)\}\, dx, \\ f_2(z) = \int\limits_0^z \{f_2'(0) - f_1'(0)T_2(x) - S\mu R_2(x)\} dx. \end{cases}$$

Let's find $f_1(\infty)$, $f_2(\infty)$

$$\begin{cases} f_1(\infty) = \int\limits_0^z \{f_1'(0) - f_2'(0)T_1(x) + R_1(x)\mu(1-S)\}\, dx, \\ f_2(\infty) = \int\limits_0^z \{f_2'(0) - f_1'(0)T_2(x) - S\mu R_2(x)\} dx. \end{cases}$$

It is necessary that the following is true

$$\begin{cases} f_1'(0) - f_2'(0)T_1(\infty) + R_1(\infty)\mu(1-S) = 0, \\ f_2'(0) - f_1'(0)T_2(\infty) - S\mu R_2(\infty) = 0. \end{cases}$$

Let's denote

$$R_1 = R_1(\infty), \quad R_2 = R_2(\infty),$$

and we will obtain:

$$R_1 = S, \quad R_2 = 1 - S.$$

From the other side

$$f_1'(0) - f_2'(0) = S\mu(S-1) = \mu R_1(R_1 - 1) \tag{7}$$

Let's write down:

$$\begin{cases} f_1(\infty) = f_2'(0) \int\limits_0^\infty (1 - T_1(x))\, dx - \mu R_2 \int\limits_0^\infty (R_1 - R_1(x))\, dx, \\ f_2(\infty) = f_1'(0) \int\limits_0^\infty (1 - T_2(x))\, dx + \mu R_1 \int\limits_0^\infty (R_2 - R_2(x)) dx. \end{cases}$$

Let's denote

$$\Delta_k = \int\limits_0^\infty (R_k - R_k(x))\, dx.$$

$$\begin{cases} f_1(\infty) = f_2'(0)T_1 - \mu R_2 \Delta_1, \\ f_2(\infty) = f_1'(0)T_2 + \mu R_1 \Delta_2. \end{cases} \tag{8}$$

Stage 3. In the equation (3) let's tend z to infinity $z \to \infty$.

$$\begin{cases} -\frac{\partial F_1(w,0,\varepsilon)}{\partial z} + \left((1-\varepsilon)\, S\mu \left(e^{j\varepsilon w} - 1\right) + \mu \left(e^{-j\varepsilon w} - 1\right)\right) F_1(w, \infty, \varepsilon) + \\ + \left(1 - e^{-j\varepsilon w}\right) \mu\varepsilon\pi_1(\infty, \varepsilon) + \frac{\partial F_2(w,0,\varepsilon)}{\partial z} T_1(\infty) = 0, \\ -\frac{\partial F_2(w,0,\varepsilon)}{\partial z} + \\ + (1-\varepsilon)\, S\mu F_2(w, \infty, \varepsilon) \left(e^{j\varepsilon w} - 1\right) + \frac{\partial F_1(w,0,\varepsilon)}{\partial z} T_2(\infty) = 0. \end{cases}$$

Let's sum the equations of the last system

$$\left((1-\varepsilon)\,S\mu\left(e^{j\varepsilon w}-1\right)+\mu\left(e^{-j\varepsilon w}-1\right)\right)F_1(w,\infty,\varepsilon)+$$

$$+(1-\varepsilon)\,S\mu\left(e^{j\varepsilon w}-1\right)F_2(w,\infty,\varepsilon)+\left(1-e^{-j\varepsilon w}\right)\mu\varepsilon\pi_1(\infty,\varepsilon)=0.$$

We will substitute expansions (6) in the equality which we got from above. By dividing both parts of the equations by ε^2 and by tending ε^2 to zero $\varepsilon\to 0$ we will obtain:

$$\left[jwS\mu-S\mu+jw\,(S-1)\,\mu f_1(\infty)+jwS\mu f_2(\infty)\right]\Phi(w)+$$

$$+\mu\pi_1(\infty)=0.$$

Let's assume that $w=0$ and considering that $\Phi(0)=1$, we will obtain:

$$\pi_1=S,\qquad \pi_1=\lim_{\varepsilon\to 0}\pi_1(\varepsilon^2).$$

Then

$$\Phi(w)=\frac{S}{S-jw\left\{S+(S-1)\,f_1(\infty)+Sf_2(\infty)\right\}}.$$

Let's review the expression in the denominator separately

$$(S-1)\,f_1(\infty)+Sf_2(\infty).$$

Considering (7)–(8) we will obtain and denote the following

$$(S-1)\,f_1(\infty)+Sf_2(\infty)=R_1f_2(\infty)-R_2f_1(\infty)=R_1f_1'(0)T_2+R_1^2\mu\Delta_2-$$

$$-R_2f_2'(0)T_1+R_2^2\mu\Delta_1=$$

$$=-R_1R_2\mu\frac{T_1T_2}{T_1+T_2}+R_1^2R_2\mu\frac{T_2^{(2)}}{2T_2}+R_2^2R_1\mu\frac{T_1^{(2)}}{2T_1}=\Delta.$$

Here

$$\Delta_k=\int_0^\infty\left(R_k-R_k(x)\right)dx=R_k\frac{T_k^{(2)}}{2T_k},$$

where $T_k^{(2)}$ is the second initial moment of the time of device's staying in mode k. $R_k(z)$ is written down in this form:

$$R_k(z)=\frac{T_k}{T_1+T_2}\left\{\frac{1}{T_k}\int_0^z\left(1-T_k(x)\right)dx\right\}.$$

Then

$$\Phi(w)=\frac{S}{S-jw\left\{S+\Delta\right\}},$$

where Δ is determined by equation (4). I.e. the function $\Phi(w)$ is a characteristic function of the exponentially distributed random variable.

Let's denote

$$d = \frac{S}{S + \Delta}.$$

By making backward substitutions, we will tend z to infinity and ε to zero, $z \to \infty$, $\varepsilon \to 0$, and we'll get the characteristic function of a number of customers in the system

$$H(u) = \sum_k F(w, \varepsilon) = \sum_k \Phi(w) R_k(\infty) + o(\varepsilon) = \frac{d}{d - ju}.$$

■

5. Conclusions

In this work we have researched mathematical model of the system with server vacations. By using method of asymptotic analysis under large load, we have found asymptotic distribution of probabilities of values of a number of customers in the system. It is shown that this distribution is exponential with the parameter determined by the equality (5).

Acknowledgments.

The work is performed under the state order of the Ministry of Education and Science of the Russian Federation (No. 1.511.2014/K).

REFERENCES

1. Pechinkin, A.V., Sokolov I.A. Queueing system with an unreliable device in discrete time // J. Inform. and its appl. 2011. T.5. V.4, P. 6–17 (in Russian).
2. Saksonov E.A. Method of the calculation of probabilities of modes for one-line queueing systems with the server vacation // J. Automat. and tele-mech. 1995.P.101–106 (in Russion).
3. Nazarov A.A., Terpugov A.F. Queueing theory: educational material. Tomsk. NTL. 2004. (in Russian).
4. Nazarov A.A., Paul S.V. Research of queueing system with the server vacation that is controlled by T-strategy. // International science conference: "Theory of probabilies, random processes, mathematical statistics and applications". Minsk: RIVSH. 2015. P. 202–207 (in Russian).
5. Nazarov A.A., Moiseeva S.P. Method of asymptotic analysis in queueing theory.Tomsk. NTL, 2006 (in Russian).
6. Moiseeva E.A., Nazarov A.A. Research of RQ-system MMP—GI—1 by using method of asymptotic analysis under large load. // TSU's herald/messenger, Administration, calculating technics and informatics. 2013. 4(25). P. 83–94 (in Russian)
7. Nazarov A.A., Moiseev A.N. Analysis of an open non-Markovian GI–(GI—?)K queueing network with high-rate renewal arrival process // Problems of Information Transmission, V. 49. No. 2. P. 167 — 178. DOI: 10.1134/S0032946013020063

ASYMPTOTIC ANALYSIS OF REPEATED REQUESTS FLOW TO THE QUEUEING SYSTEM WITH REPEATED SERVICE

L. Zadiranova[1], *S. Moiseeva*[2]
[1,2] Tomsk State University, Tomsk, Russia

Abstract

We consider the QS with repeated service requests. Using the method of asymptotic analysis to obtain expressions for the characteristic functions of the number of repeated requests to each system.

АСИМПТОТИЧЕСКИЙ АНАЛИЗ ПОТОКА ПОВТОРНЫХ ОБРАЩЕНИЙ В СМО С ПОВТОРНЫМ ОБСЛУЖИВАНИЕМ ЗАЯВОК

Л. Задиранова[1], *С. Моисеева*[2]
[1,2] Национальный исследовательский Томский государственный университет, Томск, Россия
[1]zhidkovala@mail.ru, [2]smoiseeva@mail.ru

Аннотация

Рассматриваются СМО с повторным обслуживанием заявок. С помощью метода асимптотического анализа получены выражения для характеристических функций числа повторных обращений в каждую систему.

Ключевые слова: система массового обслуживания; поток повторных обращений; метод асимптотического анализа.

1. Введение

В настоящее время внимание к теории массового обслуживания в значительной степени стимулируется необходимостью применения ее результатов для важных практических задач, возникающих в связи с бурным развитием систем коммуникаций, возникновением информационно-вычислительных систем, созданием автоматизированных систем управления, для задач экономико-математического моделирования.

СМО с неограниченным числом обслуживающих приборов являются математическими моделями вычислительных, информационных системы и различных социально-экономических систем [1, 2, 4].

Основными методами исследования СМО с неограниченным числом приборов, как правило, являются метод вложенных цепей Маркова и метод дополнительной переменной. В последнее время также развиваются матрично-аналитические методы. В случаях, когда не удается найти характеристики системы в явном виде, применяют асимптотические методы [3].

В настоящей работе , помощью метода асимптотического анализа, приводится исследование потока повторных обращений за время t в системы $M|M|\infty$, $MMPP|M|\infty$ и $GI|M|\infty$.

2. Постановка задачи

Рассмотрим системы массового обслуживания с неограниченным числом обслуживающих устройств, в качестве моделей входящих потоков взяты пуассоновский поток (M), марковский модулированный поток (MMPP) и рекуррентный поток (GI). Продолжительность обслуживания заявки является случайной величиной и имеет экспоненциальное распределение с параметром μ. Поступившая заявка занимает любой из свободных приборов, завершив обслуживание на котором, с вероятностью 1-r покидает систему или с вероятностью r возвращается для повторного обслуживания. Ставится задача исследования потока заявок, обратившхся в рассматриваемые системы за время t, для повторного обслуживания.

3. Поток повторных обращений в системе $M|M|\infty$ с повторным обслуживанием

Рассматривается система массового обслуживания с неограниченным числом приборов, на вход которой поступает простейший поток заявок с интенсивностью λ. Обозначим $i(t)$ - число занятых приборов в момент времени t, $n(t)$ - число заявок, обратившихся в систему за время t для повторного обслуживания, тогда двумерный поток $\{i(t), n(t)\}$ является марковским.

Для распределений вероятностей $P(i,n,t) = P\{i(t) = i, n(t) = n\}$ запишем систему дифференциальных уравнений Колмогорова для характеристических функций

$$\frac{\partial H(u,w,t)}{\partial t} = \lambda(e^{ju}-1)H(u,w,t)+j\mu(1-re^{jw}-(1-r)e^{-ju})\frac{\partial H(u,w,t)}{\partial u}. \quad (1)$$

Полученное уравнение позволяет определить основные вероятностные характеристики рассматриваемой системы, в том числе и для потока повторных обращений в систему.

Проведем исследование потока повторных обращений в систему за время t, с помощью метода асимптотического анализа в условии растущего времени обслуживания.

Обозначим

$$\mu = \epsilon, u = \epsilon y, H(u,w,t) = F(y,w,t,\epsilon),$$

тогда перепишем (1) в виде

$$\frac{\partial F(y,w,t,\epsilon)}{\partial t} = j(1 - re^{jw} - (1-r)e^{-j\epsilon y})\frac{\partial F(y,w,t,\epsilon)}{\partial y} +$$

$$+\lambda(e^{j\epsilon y} - 1)F(y,w,t,\epsilon). \tag{2}$$

Теорема 1. *Предельное, при $\epsilon \to 0$, значение функции $F(y,w,t)$ решения $F(y,w,t,\epsilon)$ уравнения (2) имеет вид*

$$F(y,w,t) = exp\left\{\frac{\lambda r t(e^{jw}-1)}{1-r} + \frac{j\lambda y}{1-r}\right\}. \tag{3}$$

Доказательство. Выполняя в уравнении (2) предельный переход при $\epsilon \to 0$, получаем дифференциальное уравнение первого порядка

$$\frac{\partial F(y,w,t,\epsilon)}{\partial t} = jr(1 - e^{jw})\frac{\partial F(y,w,t,\epsilon)}{\partial y}. \tag{4}$$

Общее решение полученного уравнения имеет вид

$$F(y,w,t) = \varphi\left(t + \frac{jy}{r(e^{jw}-1)}\right),$$

где $\varphi(y)$ некоторая функция вид которой определим, используя начальное условие.

Рассмотри функцию $F(y,w,t)$ в нулевой момент времени, очевидно, что данная функция не будет зависеть от w, то начальное условие имеет вид

$$F(y,w,0) = \Phi(y), \tag{5}$$

где $\Phi(y)$ асимптотическое приближение характеристической функции распределения числа занятых приборов в системе в условии растущего времени обслуживания заявок

$$\Phi(y) = exp\left\{\frac{j\lambda y}{1-r}\right\}.$$

Таким образом, решение уравнения (2), удовлетворяющее начальному условию (5), имеет вид

$$F(y,w,t) = exp\left\{\frac{\lambda r t(e^{jw}-1)}{1-r} + \frac{j\lambda y}{1-r}\right\}.$$

которое совпадает с равенством (3).

Теорема доказана.

■

Полагая в (3) $y = 0$, имеем асимптотическое приближение характеристической функции числа повторных обращений, в условии растущего времени обслуживания заявок

$$h(w,t) = exp\left\{\frac{\lambda r t(e^{jw}-1)}{1-r}\right\}.$$

4. Поток повторных обращений в системе $MMPP|M|\infty$ с повторным обслуживанием

Рассматривается система массового обслуживания с неограниченным числом приборов, на вход которой поступает марковский модулированный поток (MMPP), управляемый цепью Маркова $k(t)$ с конечным числом состояний, $k(t) = 1, 2, \ldots, K$, заданной матрицей инфинитезимальных характеристик $\mathbf{Q}$, $i, j = 1, 2, \ldots, K$ и матрицей условных интенсивностей Λ.

Обозначим $i(t)$ - число занятых приборов в момент времени t, $n(t)$ - число заявок, обратившихся в систему за время t для повторного обслуживания, $k(t)$ - состояние управляющей цепи Маркова, тогда трехмерный процесс $\{k(t), i(t), n(t)\}$ является марковским.

Для распределений вероятностей $P(k,i,n,t) = P\{k(t) = k, i(t) = i, n(t) = n\}$ запишем систему дифференциальных уравнений Колмогорова для частичных характеристических функций в матричном виде

$$\frac{\partial \mathbf{H}(u,w,t)}{\partial t} + j\mu(re^{jw} - 1 + (1-r)e^{-ju})\frac{\partial \mathbf{H}(u,w,t)}{\partial u} =$$

$$= \mathbf{H}(u,w,t)[(e^{ju}-1)\Lambda + \mathbf{Q}], \qquad (6)$$

где

$$\mathbf{H}(u,w,t) = [H(1,u,w,t), H(2,u,w,t), ..., H(K,u,w,t)],$$

$$\Lambda = \begin{pmatrix} \Lambda_1 & 0 & \cdots & 0 \\ 0 & \Lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \Lambda_K \end{pmatrix},$$

$$\mathbf{Q} = \begin{pmatrix} q_{11} & q_{12} & \cdots & q_{1K} \\ q_{21} & q_{22} & \cdots & q_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ q_{K1} & q_{K2} & \cdots & q_{KK} \end{pmatrix}.$$

Найдем асимптотическую характеристическую функцию числа повторных обращений в системе $MMPP|M|\infty$ за время t в условии растущего времени.

Обозначим

$$\mu = \epsilon, u = \epsilon y, \mathbf{H}(u, w, t) = \mathbf{F}(y, w, t, \epsilon), \tag{7}$$

Перепишем (6) с учетом введенных обозначений

$$\frac{\partial \mathbf{F}(y, w, t, \epsilon)}{\partial t} + j(re^{jw} - 1 + (1 - r)e^{-jy\epsilon})\frac{\partial \mathbf{F}(y, w, t, \epsilon)}{\partial y} =$$

$$= \mathbf{F}(y, w, t, \epsilon)[(e^{jy\epsilon} - 1)\Lambda + \mathbf{Q}]. \tag{8}$$

Теорема 2. *Сумма компонентов предельного, при* $\epsilon \to 0$, *значение вектор-функции* $\boldsymbol{F}(y, w, t)$ *решения* $\boldsymbol{F}(y, w, t, \epsilon)$ *уравнения* (8) *имеет вид*

$$\boldsymbol{F}(y, w, t)\boldsymbol{E} = exp\left\{\frac{\lambda rt(e^{jw} - 1)}{1 - r} + \frac{j\lambda y}{1 - r}\right\}, \tag{9}$$

где величина λ *определяется выражением*

$$\lambda = \boldsymbol{R}\Lambda\boldsymbol{E}.$$

Доказательство. Суммируя все уравнения системы (8) и выполняя предельный переход при $\epsilon \to 0$, получим уравнение в частных производных первого порядка

$$\frac{\partial \mathbf{F}(y, w, t)}{\partial t}\mathbf{E} + jr(e^{jw} - 1)\frac{\partial \mathbf{F}(y, w, t)}{\partial y}\mathbf{E} = 0,$$

решение которого имеет вид

$$\mathbf{F}(y, w, t)\mathbf{E} = \varphi\left(t + \frac{jy}{r(e^{jw} - 1)}\right),$$

где $\varphi(y)$ некоторая функция.

Так как число обслуженных заявок за интервал нулевой длины с вероятностью единица равно нулю, то начальное условие для определения вида функции $\varphi(y)$ имеет вид

$$\mathbf{F}(y, w, 0)\mathbf{E} = \Phi(y),$$

где $\Phi(y)$ - асимптотическое приближение характеристической функции распределения числа занятых приборов в системе в условии растущего времени обслуживания заявок, вид которого получен в работе [5]

$$\mathbf{F}(y, w, 0)\mathbf{E} = exp\left\{\frac{jy\lambda}{1 - r}\right\}. \tag{10}$$

Таким образом, решение уравнения (8), удовлетворяющее начальному условию (10) имеет вид

$$\mathbf{F}(y,w,t)\mathbf{E} = exp\left\{\frac{\lambda rt(e^{jw}-1)}{1-r} + \frac{j\lambda y}{1-r}\right\}.$$

■

Полагая в (9) $y = 0$, имеем асимптотическое приближение характеристической функции числа повторных обращений, в условии растущего времени обслуживания заявок

$$h(w,t) = exp\left\{\frac{\lambda rt(e^{jw}-1)}{1-r}\right\}.$$

5. Поток повторных обращений в системе $GI|M|\infty$ с повторным обслуживанием

Рассмотривается система массового обслуживания с неограниченным числом обслуживающих приборов, на вход которой поступает рекуррентный поток заявок с функцией распределения длин интервалов между моментами поступления заявок $A(x)$. Обозначим $i(t)$ - число занятых приборов в системе в момент времени t, $n(t)$ - число заявок, обратившихся в систему за время t для повторного обслуживания.

Так как полученный случайных процесс $i(t), n(t)$ немарковкий, то марковизируем его, введя дополнительную переменную $z(t)$, равную длине интервала от момента t до момента поступления следующей заявки. Тогда трехмерный процесс $z(t), i(t), m(t)$ будет марковским.

Обозначим распределение вероятностей значений полученного марковского процесса $P(z,i,n,t) = P\{z(t) < z, i(t) = i, n(t) = n\}$, для которого получаем систему дифференциальных уравнений Колмогорова для характеристических функций числа повторных обращений в следующем виде

$$\frac{\partial H(z,u,w,t)}{\partial t} = \frac{\partial H(z,u,w,t)}{\partial z} - j\mu(re^{jw} - 1 + (1-r)e^{-ju})\frac{\partial H(z,u,w,t)}{\partial u} +$$

$$+(e^{ju}A(z) - 1)\frac{\partial H(0,u,w,t)}{\partial z}. \qquad (11)$$

Обозначим

$$\mu = \epsilon, u = \epsilon y, H(z,u,w,t) = F(z,y,w,t,\epsilon),$$

тогда перепишем (11) в виде

$$\frac{\partial F(z,u,w,t,\epsilon)}{\partial t} = \frac{\partial F(z,u,w,t,\epsilon)}{\partial z} - j(re^{jw} - 1 + (1-r)e^{-jy\epsilon})\frac{\partial F(z,u,w,t,\epsilon)}{\partial y} +$$

$$+(e^{jy\epsilon}A(z) - 1)\frac{\partial F(0,u,w,t,\epsilon)}{\partial z}. \qquad (12)$$

Теорема 3. *Предельное, при* $\epsilon \to 0$*, значение функции* $F(y,w,t)$ *решения* $F(y,w,t,\epsilon)$ *уравнения* (12) *имеет вид*

$$F(y,w,t) = exp\left\{\frac{\lambda r t(e^{jw}-1)}{1-r} + \frac{j\lambda y}{1-r}\right\}, \tag{13}$$

где

$$\lambda = \frac{\partial R(0)}{\partial z},$$

$R(z)$ *- стационарное распределение вероятностей значений случайного процесса* $z(t)$*.*

Доказательство. Выполняя в уравнении (12) предельный переход при $\epsilon \to 0$, получаем, что $F(z,y,w,t)$ является решением уравнения

$$\frac{\partial F(z,u,w,t)}{\partial t} = \frac{\partial F(z,u,w,t)}{\partial z} - jr(e^{jw}-1)\frac{\partial F(z,u,w,t)}{\partial y} + (A(z)-1)\frac{\partial F(0,u,w,t)}{\partial z}.$$

В полученном соотношении выполним предельный переход при $z \to \infty$, получаем дифференциальное уравнение вида

$$\frac{\partial F(z,u,w,t)}{\partial t} + jr(e^{jw}-1)\frac{\partial F(z,u,w,t)}{\partial y} = 0,$$

решение которого имет вид

$$F(y,w,t) = \varphi\left(t + \frac{jy}{r(e^{jw}-1)}\right),$$

где $\varphi(y)$ некоторая функция.

Для пределения вида функции $\varphi(y)$ используем начальное условие

$$F(y,w,0) = \Phi(y), \tag{14}$$

где $\Phi(y)$ - асимптотическое приближение характеристической функции числа занятых приборов в системе в момент времени t в условии растущего времени обслуживания, вид которого был определен в работе [6]

$$\Phi(y) = exp\left\{\frac{j\lambda y}{1-r}\right\}.$$

Тогда частное решение уравнения (12), удовлетворяющее начальному условию (14) имет вид

$$F(y,w,t) = exp\left\{\frac{\lambda r t(e^{jw}-1)}{1-r} + \frac{j\lambda y}{1-r}\right\}.$$

Теорема доказана.

■

Полагая в (13) $y = 0$, имеем асимптотическое приближение характеристической функции числа повторных обращений, в условии растущего времени обслуживания заявок

$$h(w,t) = exp\left\{\frac{\lambda rt(e^{jw}-1)}{1-r}\right\}.$$

6. Заключение

В настоящей работе рассмотрены математические модели систем $M|M|\infty$, $MMPP|M|\infty$ и $GI|M|\infty$ с повторными обращениями, получено асимптотическое приближение характеристических функций потока повторных обращений в условии растущего времени обслуживания заявок для каждой системы.

ЛИТЕРАТУРА

1. Моисеева, С. П. Математическая модель параллельного обслуживания кратных заявок с повторными обращениями / С. П. Моисеева, И. А. Захорольная // Автометрия. - 2011. - Т. 47, №6. - С. 51-58.
2. Ананина, И. А. Исследование потоков в системе M/GI/∞ с повторными обращениями методом предельной декомпозиции / И. А. Ананина, С. П. Моисеева, А. А. Назаров // Вестник Томского государственного университета. Управление, вычислительная техника и информатика. - 2009. - №3(8). - С. 56-67.
3. Назаров, А. А. Методы асимптотического анализа в теории массового обслуживания / А. А. Назаров, С. П. Моисеева - Томск : Изд-во НТЛ, 2006. - 112 с.
4. Жидкова, Л. А. Математическая модель потоков покупателей двухпродуктовой торговой компании в виде системы массового обслуживания с повторными обращениями к блокам / Л. А. Жидкова, С. П. Моисеева // Известия Томского политехнического университета. - 2013. - Т. 322, №5. - С. 5-9.
5. Жидкова, Л. А. Исследование числа занятых приборов в системе MMPP|M|∞ с повторными обращениями / Л. А. Жидкова, С. П. Моисеева // Вестник Томского государственного университета. Управление, вычислительная техника и информатика. 2014. №1 (26). С. 53-62
6. Жидкова, Л. А. Исследование системы GI|M|∞ с повторными обращениями / Л. А. Жидкова, С. П. Моисеева // Труды Томского государственного университета. - Серия физико-математическая: Математическое и программное обеспечение информационных, технических и экономических систем: материалы II Всероссийской молодежной научной конференции, Т. 295. - Томск: Издательский Дом Томского государственного университета, 2014. - С. 94-100.

ANALYSIS OF QUEUEING MODELS WITH STATE-DEPENDENT JUMP PRIORITIES

A.Z. Melikov[1], *A. M. Rustamov*[2], *T.I. Jafarzade*[3], *J. Sztrik*[4]

[1]Institute of Control Systems, ANAS, F. Agayev 9,AZ1141, Baku, Azerbaijan,
[2]Qafqaz University, H. Aliyev 120, AZ0101, Khirdalan, Baku, Azerbaijan,
[3]National Aviation Academy, Merdekan pr. 30, AZ1045, Baku, Azerbaijan,
[4]University of Debrecen, 4032, Debrecen, Hungary,
[1]agassi.melikov@gmail.com, [2]anrustemov@qu.edu.az,
[3]turan_jafarzade@hotmail.com, [4]sztrik.janos@inf.unideb.hu

Abstract

In this paper, exact and approximate approaches for studying queuing models with state-dependent jump priorities are developed. Models with finite separate buffers for heterogeneous calls are investigated. It is shown that the investigated models might be described by two-dimensional Markov Chains. One of the main challenge in exact approach for the solution of appropriate system of balance equations for state probabilities becomes big computation for large scale models. To overcome the indicated difficulties an approximate approach based on the state space merging algorithms is developed. This approach allows constructing simple algorithms to calculate the quality of service metrics of the examined models.

Keywords: queuing models, jump priority, Markov chains, space merging, numerical analysis

1. Introduction

Priorities are effective tools to solve the problems of quality of service (QoS) provisioning of heterogeneous calls in queuing systems. By nature the priorities can be broadly divided into two classes: static and dynamic. Static priorities (relative or preemptive) are defined in advance and they do not change during the whole system operation time [1]. In literature relative static priorities in queuing systems with buffers sometimes are called HOL-priorities (Head-Of-Line), .i.e. in static priorities call for service is chosen from the head of line according to the highest priority. Dynamic priorities in turn are divided into two classes: dynamical vs time and dynamical vs state. In dynamical versus time priorities the priority of the calls can be changed according their waiting times (or sojourn time) [2]. In dynamical versus state priorities (they sometimes are called state-dependent priorities) calls can change priority according the state of the system where the state is described by vector whose components indicate the number of heterogeneous calls in the queue (or in the system) [3].

The drawback of static priorities is that when they are used in real systems the delay of low priority calls is too large especially for the system with heavy

loads of high priority calls. Dynamic priorities allow avoiding the starvation of low priority calls. Detailed review of priority schemas might be found in [4].

As a rule, classical priorities (static or dynamic) are used to determine type of call from the buffer which must be send to channel for servicing. However, some scientific and practical interest represents the priorities which are introduced to change (either increase or decrease) the priorities of calls in buffer. These changes are realized instantaneously so such kinds of priorities are called jump priorities (JP).

The pioneer work on the analyzing dynamical vs time HOL-priorities with priority jumps (HOL-JP) is [5]. In this paper dynamical vs time HOL-JP was proposed where calls with low priority can jump to another buffer with high priority after waiting some (deterministic) period of time in native buffer. Formulas for calculation of the mean waiting time of the heterogeneous calls were developed in [5].

Dynamical vs states HOL-JP in discrete-time queuing models were proposed in [6-9]. In [5-9] queuing models with infinite buffers are investigated. So, they have little applicability in the real communication networks. In particular, real communication networks have finite buffer capacity. Secondly, investigated HOL-JP is defined by state-independent probabilities. Therefore they cannot be adapted for real situations according to the changes of loads of heterogeneous calls.

Different approach to study queuing models with dynamical versus state HOL-JP can be found in the papers [10-13] and in chapter 5 of the book [14] where new type of randomized state-dependent JP for continuous-time queuing systems with finite buffers was proposed. They make it possible to pass to from the L-queue (queue for low priority calls, L-calls) into the H-queue (queue for high priority calls, H-calls) only at the instants of arrival of the L-calls, but the probability of such transitions depend only on the number of L-calls in the system. In chapter 5 of the book [14] models with separate buffers which jump priorities depending only on the number of H-calls in the system were examined. In the indicated works [10-14] methods of calculation of main QoS metrics of the investigated models are proposed. To the best of our knowledge, models in which JP depends on the number of both types of calls in the system are not examined. In this paper we investigate such kinds of models. Our contribution consist of two parts; 1) we propose novel kind of state-dependent jump priorities, and 2) both exact and approximate methods to calculate the QoS metrics of queuing models with such kind of priorities are developed.

The rest of the paper is organized as follows. In section 2, model with separate buffers is defined and state-dependent JP is introduced. In section 3, exact method of calculation its QoS metrics is developed. In section 4 an approximate method to solving the same problem is developed. Conclusion remarks are given in section 5.

2. Jump Priorities in Model with Separate Buffer

In the single server queueing system two Poisson traffic of heterogeneous calls have different arrival rate $\lambda_i, i = 1, 2$. We determined first type of calls as high priority calls (H-calls) while second type of calls are treated as low priority calls (L-calls). In general, H-calls have relative priority over L-calls while channel is idle, namely in the case of absence of H-call in the buffer, L-call can be served. If there isn't any call in the buffer, then the channel becomes free. Service intensity of the server is the same for both types of the call where it is determined as μ obeying exponential distribution.

Consider the model with separate buffers, i.e. it is assumed that there are two isolated buffers — H-buffer (for waiting H-calls) and L-buffer (for waiting L-calls) with size of R_1 and R_2 $(0 < R_i < \infty, i = 1, 2)$ respectively.

Decision epochs coincide with the arrival moments of L-calls. In this model state-dependent HOL-JP is defined as follows.

- High priority calls are always accepted to the H-buffer with probability 1 if there is a free place in this buffer. If the H-buffer is full then arriving H-call is dropped with probability 1.
- If upon arrival of L-call the number of calls of this type equals $i, i < R_2$, and the number of H-call equals $j, j < R_1$, then L-call joins the H-buffer with probability $\alpha_i(j)$ and in future it will be served as H-call; and arriving L-call joins the L-buffer with probability $1 - \alpha_i(j)$.
- If upon arrival of L-call the number of H-call equals R_1, then L-call joins the L-buffer if there is free place in this buffer; otherwise, arriving L-call is dropped with probability 1.
- If upon arrival of L-call L-buffer is full and the number of H-call equals $j, j < R_1$, then L-call joins the H-buffer with probability $\alpha_{R_2}(j)$; and arriving L-call is dropped with probability $1 - \alpha_{R_2}(j)$.

The problem is finding the QoS metrics for this model. The main QoS metrics are the following: the stationary probability of losing the calls of the i-th type, the mean number of the i-th type calls in the buffers and the mean call transmission delay of the i-th type calls, $i = 1, 2$.

3. Exact Method

The state of the system is defined by 2-D vectors $\mathbf{n} = (n_1, n_2)$ where the first component indicates the number of H-calls and the second one the number of L-calls respectively. So, operation of this system is described by the 2-D Markov Chain (2-D MC) with the following state space:

$$S = \{\boldsymbol{n} : n_i = 0, 1, ..., R_i, i = 1, 2\}. \tag{1}$$

Transition intensity from state $\boldsymbol{n} \in S$ to state $\boldsymbol{n}' \in S$ are denoted by $q(\boldsymbol{n}, \boldsymbol{n}')$. Then nonnegative elements of the generating matrix (Q-matrix) of the given 2-D MC can be calculated as below:

$$q(\boldsymbol{n},\boldsymbol{n}') = \begin{cases} \lambda_1 + \lambda_2\alpha_{n_2}(n_1), & \text{if } n_1 < R_1, \boldsymbol{n}' = \boldsymbol{n} + \boldsymbol{e_1} \\ \lambda_2(1-\alpha_{n_2}(n_1)), & \text{if } n_1 < R_1, n_2 < R_2, \boldsymbol{n}' = \boldsymbol{n} + \boldsymbol{e_2} \\ \lambda_2, & \text{if } n_1 = R_1, \boldsymbol{n}' = \boldsymbol{n} + \boldsymbol{e_2} \\ \mu, & \text{if } n_1 > 0, \boldsymbol{n}' = \boldsymbol{n} - \boldsymbol{e_1} \text{ or } n_1 = 0, \boldsymbol{n}' = \boldsymbol{n} - \boldsymbol{e_2} \\ 0, & \text{in other cases.} \end{cases} \tag{2}$$

where $\boldsymbol{e}_1 = (1,0), \boldsymbol{e}_2 = (0,1)$.

The stationary probability of state $n \in S$ is denoted by $p(n)$. Construction and solution of the corresponding system of balance equations (SBE) for the given 2-D MC is the standard way for determining the stationary state probabilities. It is constructed with regard to (2) and here is omitted.

After determining the state probabilities from SBE, one can establish its QoS metrics. As indicated above, H-calls are lost if upon their arrivals H-buffer is full. Hence, the loss probability for H-calls (CLP_1) can be determined as follows:

$$CLP_1 = \sum_{i=0}^{R_2} p(R_1, i). \tag{3}$$

Similarly, we conclude that the loss probability of L-calls (CLP_2) is given by

$$CLP_2 = p(R_1, R_2) + \sum_{i=0}^{R_1-1} p(i, R_2)(1 - \alpha_{R_2}(i)). \tag{4}$$

The mean numbers of the H-calls (L_1) and L-calls (L_2) in the queue are determined as the expected values of appropriate discrete random variables:

$$L_1 = \sum_{i=1}^{R_1} i \sum_{j=0}^{R_2} p(i,j); \tag{5}$$

$$L_2 = \sum_{i=1}^{R_2} i \sum_{j=0}^{R_1} p(j,i). \tag{6}$$

Further, formulas (3)-(6) and modified Little's formula can be used to evaluate the mean times of call transmission delay (CTD_i) for the heterogeneous:

$$CTD_1 = \frac{L_1}{\lambda_1^{(c)}}, \tag{7}$$

$$CTD_2 = \frac{L_1 + L_2}{\lambda_1^{(c)} + \lambda_2^{(c)}}, \tag{8}$$

where $\lambda_1^{(c)}$ and $\lambda_2^{(c)}$ are carried loads of H-calls and L-calls, respectively. These parameters are calculated as follows:

$$\lambda_1^{(c)} = \lambda_1 \left(1 - \sum_{j=0}^{R_2} p(R_1, j)\right) + \lambda_2 \sum_{i=0}^{R_1-1} \sum_{j=0}^{R_2} p(i,j)\alpha_j(i),$$

$$\lambda_2^{(c)} = \sum_{i=0}^{R_1} \sum_{j=0}^{R_2-1} p(i,j)\left(1 - \alpha_j(i)\right).$$

By implementation of programming languages it is possible to solve the SBE for the steady-state probabilities $p(\boldsymbol{n}), \boldsymbol{n} \in S$ with a help of numerical methods of the linear algebra. This method of calculation of QoS metrics is called the exact (precise) method. In cases of small dimensions of state space (1) this method is reasonable to calculate QoS metrics of the system. But for large scale system it isn't suitable. Therefore, we need to find out a more efficient method to calculate the QoS metrics of the models with large dimensions of buffers.

4. Approximate Method

Below we consider asymptotic analysis of the QoS metrics for large scale models, i.e. when R_1 and R_2 take large values. The developed approximate method has high accuracy for heavy traffic regime of H-calls. In other words, below we consider asymptotic analysis of the large scale model with heavy loads of H-calls, i.e. it is assumed that $\nu_1 >> \nu_2 >> 1$, where $\nu_i = \lambda_i/\mu, i = 1, 2$. Note that this assumption make sence for the jump priorites for the L-calls in the systems with heavy loads of H-calls.

Consider the following splitting of the state space (1):

$$S = \bigcup_{i=0}^{R_2} S_i, S_i \bigcap S_j = \emptyset, i \neq j, \tag{9}$$

where $S_i = \{\boldsymbol{n} \in S : n_2 = i\}, i = 0, 1, 2, ..., R_2$.

We notice that the assumption made about the relation of the loads of the heterogeneous calls enables one to satisfy the condition for correct use of the algorithms of state space merging of the 2-D MC (see [3, Appendix]): transition intensities within classes $S_i, i = 0, 1, ..., R_2$, are essentially higher than those between states of different classes. The classes of microstates S_i are united into individual merged states $< i >$, and in the original state space S the following merge function is defined:

$$U(\boldsymbol{n}) =< i >, \text{if } \boldsymbol{n} \in S_i. \tag{10}$$

The function (10) defines a merged model with the state space $\Omega = \{< i >: i = 0, 1, ..., R_2\}$. Let us consider the problem of calculation of state probabilities inside the splitting models. The stationary probability of the state (k, i) in the split model with the state space S_i is denoted by $\rho_i(k), i = 0, 1, ..., R_2, k = 0, 1, ..., R_1$.

Each split model with state space S_i. is a 1-D birth and death process with the parameters that are calculated as follows:

$$q_i(k_1, k_2) = \begin{cases} \lambda_1 + \lambda_2\alpha_i(k_1), & \text{if } k_2 = k_1 + 1 \\ \mu, & \text{if } k_2 = k_1 - 1 \\ 0, & \text{otherwise.} \end{cases} \tag{11}$$

Consequently, we have

$$\rho_i(k) = \prod_{j=0}^{k-1} (\nu_1 + \nu_2\alpha_i(j))\rho_i(0), k = 1, ..., R_1, \tag{12}$$

where $\rho_i(0) = \left(1 + \sum_{k=1}^{R_1} \prod_{j=0}^{k-1} (\nu_1 + \nu_2\alpha_i(j))\right)^{-1}$.

The elements of the Q-matrix of the merged model are denoted by $q(< k > , < k' >), < k >, < k' >\in \Omega$. According to the algorithm of state space merging of the 2-D MC (see [3, Appendix]) these elements are given by

$$q(< k >, < k' >) = \sum_{\boldsymbol{n}\in S, \boldsymbol{n}'\in S_{k'}} q(\boldsymbol{n}, \boldsymbol{n}')\rho_{n_1}(n_2). \tag{13}$$

So, by using (2), (12) and (13) after some mathematical transformations the following formulae are obtained

$$q_i(< k >, < k' >) = \begin{cases} \lambda_2 \left(\rho_k(R_1) + \sum_{i=0}^{R_1-1}(1 - \alpha_k(i)\right), & \text{if } k' = k + 1 \\ \mu\rho_k(0), & \text{if } k' = k - 1 \\ 0, & \text{otherwise.} \end{cases} \tag{14}$$

From (14) we can calculate the probabilities of the merged states $\pi(< k >), < k >\in \Omega$ as follows:

$$\pi(< k >) = \nu_2^k \prod_{j=0}^{k-1} \Lambda_j \pi(< 0 >), k = 1, 2, ..., R_2, \tag{15}$$

where $\pi(< 0 >) = \left(1 + \sum_{k=1}^{R_2} \nu_2^k \prod_{j=0}^{k-1} \Lambda_j\right)^{-1}, \Lambda_j = \frac{\rho_j(R_1) + \sum_{i=0}^{R_1-1}(1-\alpha_j(i))\rho_j(i)}{\rho_{j+1}(0)}, j = 0, 1, 2, ..., R_2 - 1$.

The state probabilities of the initial 2-D MC are determined approximately as follows:

$$p(i,j) \approx \rho_j(i)\pi(< j >). \tag{16}$$

By taking into account (16), (12) and (15) we can calculate approximate values of state probabilities of initial 2-D MC, and omitting the intermediate mathematical calculations the following approximate formulae to calculate the QoS metrics (3)-(6) are obtained:

$$CLP_1 \approx \sum_{i=0}^{R_2} \rho_i(R_1)\pi(< i >). \tag{17}$$

$$CLP_2 \approx \pi(< R_2 >) \left(\rho_{R_2}(R_1) + \sum_{i=0}^{R_1-1} \rho_{R_2}(i)(1-\alpha_{R_2}(i)) \right). \tag{18}$$

$$L_1 \approx \sum_{i=1}^{R_1} i \sum_{k=0}^{R_2} \rho_k(i)\pi(< k > . \tag{19}$$

$$L_2 \approx \sum_{k=1}^{R_2} k\pi(< k >). \tag{20}$$

The approximate value of QoS metrics CTD_i (see (7), (8)) are determined from (17)-(20) after the calculation of the parameters CLP_i and $L_i, i = 1, 2$. Here it should be mention that, approximate values of the carried loads of H-calls and L-calls are calculated according to the formula (16).

The developed approximate formulas allow one to carry out an authentic analysis of QoS metrics over any range of change of values of loading parameters of the heterogeneous traffic and also at any buffers sizes. Another goal of performing numerical experiments was the estimation of the proposed approximate formulas accuracy. In order to be short, here the appropriate results are omitted. Let us only note that accuracy of the proposed approximate formulas is acceptable for engineering practice. The bigger the ratio of loads of H-calls to L-calls, the higher the accuracy of approximate value of QoS metrics.

5. Conclusion

This paper proposed a new class of state-dependent JP in queueing systems with finite separate buffers for heterogeneous calls. An exact and approximate approaches for calculating the QoS metrics of heterogeneous calls in such systems are developed. They might be used to investigate the models of queueing systems with finite common buffer for heterogeneous calls as well. The important advantage of approximate approach lies in the use of explicit formulae to calculate the QoS metrics, which enables our approach to be used for models of any dimension. In addition, it is possible to use the proposed formulae to find the optimal (in given sense) values of jump priorities. Latest problems are important especially for the threshold-based non-randomized JP-schemas and they are a subject for further study.

Acknowledgments.

The publication was supported by the T'AMOP-4.2.2.C-11/1/KONV-2012-0001 project. The project has been supported by the European Union, co-financed by the European Social Fund.

REFERENCES

1. Jaiswal HK. Priority queues. New York: Academic Press; 1968.
2. Kleinrock L. A delay dependent queue discipline. Naval Research Logistics Quarterly Journal 1964; 11: 329-341.
3. Melikov A, Ponomarenko L, Kim CS. Performance Analysis and Optimization of Multi-Traffic on Communication Networks. Heidelberg: Springer; 2010.
4. Wittevrongel S, De Vuyst S, Sys C, Bruneel H. A reservation-based scheduling mechanism for fair QoS provisioning in packet-based networks. In: Proceeding of the 26th IEEE International Teletraffic Congress, Karlskrona: 2014, p. 55-62.
5. Lim Y, Kobza JE. Analysis of delay dependent priority discipline in an integrated multiclass traffic fast packet switch. IEEE Transactions on Communication 1990; 38(5): 659-665.
6. Maertens T, Walraevens J, Bruneel H. On priority queues with priority jumps. Performance Evaluation 2006; 63(12): 1235-1252.
7. Maertens T, Walraevens J, Bruneel H. A modified HOL priority scheduling discipline: performance analysis. European Journal of Operation Research 2007; 180(3): 1168-1185.
8. Maertens T, Walraevens J, Bruneel H. Performance comparison of several priority schemes with priority jumps. Annals of Operation Research 2008; 162: 109-125.
9. Walraevens J, Steyaert B, Bruneel H. Performance analysis of single-server ATM queue with priority scheduling. Computers and Operation Research 2003; 30(12): 1807-1829.
10. Melikov AZ, Ponomarenko LA., Kim CS. Algorithmic approach to analysis of queuing system with finite buffers and jump priorities. Journal of Automation and Information Sciences 2012; 44(12): 43-54.
11. Kim CS, Oh Y, Melikov AZ. A space merging approach to the analysis of the performance of queueing models with buffers and priority jumps. Industrial Engineering and Management Systems 2013; 12(3): 274-280.
12. Melikov AZ, Ponomarenko LA, Kim CS. Approximate method for analysis of queuing models with jump priorities. Automation and Remote Control 2013; 74(1): 62-75.
13. Melikov AZ, Ponomarenko LA, Kim CS. Numerical method for analysis of queuing models with priority jumps. Cybernetics and System Analysis 2013; 49(1): 55-61.
14. Melikov A., Ponomarenko L. Multidimensional queueing models in telecommunication networks. Heidelberg: Springer; 2014.

RESEARCH THE POSSIBILITY OF MODIFYING RADIAL BASIS FUNCTION FOR LOCALIZATION SYSTEM IN THE BUILDING

L. Luoh[1], *G. Antiokh*[2], *A. Rozhnov*[3]
[1] Department of Electrical Engineering, Chung Hua University, Hsin-Chu, Taiwan, R.O.C
[2] Moscow Aviation Institute, Moscow, Russia,
[3] V.A. Trapeznikov Institute of Control Sciences of RAS, Moscow, Russia

Abstract

Seeing that GPS fails to function inside a building, indoor positioning system (IPS) has recently gained much attention in the field of rescue, medical application, public facilities, emerging indoor location based services, etc. This study proposes a novel strategy based on wireless sensor network (WSN) to achieve high positioning accuracy for moving entities. Firstly, radial basis function network (RBFN) model for indoor location estimation is constructed in an unknown environment. Secondly, owing to the shortcoming of locating instant received signal strength (RSS), we introduce the benefits of using the average RSS to reduce noise interference and exclude transient APs. In addition, we integrate Zigbee hardware to realize a set of convenient wireless IPS with comparatively low cost. Finally, in order to reach an optimal accuracy, we adopt multiple similar networks within the same environment. Experiments in this study have demonstrated effective enhancement of existing IPS accuracy with the average error as little as 2.8 meters 100% when compared with other approaches. The Russian side has been carried out testing of this approach for the development of bilateral cooperation in the framework of joint research initiative of the University of Chung Hua and creative team based on ICS RAS.

ИССЛЕДОВАНИЕ ВОЗМОЖНОСТЕЙ МОДИФИКАЦИИ РАДИАЛЬНО-БАЗИСНОЙ ФУНКЦИИ ПРИ ПОСТРОЕНИИ СИСТЕМЫ ЛОКАЛИЗАЦИИ В ЗДАНИИ

Л.Луо[1], *Г. Антиох*[2], *А. Рожнов*[3]
[1] Департамент электротехники, Университет Чунг Хуа, Синь-Чу, Тайвань,
[2] Московский Авиационный Институт, Москва, РФ,
[3] Институт проблем управления им. В.А. Трапезникова РАН, Москва, РФ
[1]lluoh@chu.edu.tw, [2]grigory.antiokh@yandex.ru, [3]rozhnov@ipu.ru

Аннотация

В связи с тем, что GPS невозможно применять внутри зданий, в последнее время система навигации в помещениях (IPS) стала по-

лучать больше внимания во многих сферах жизни человека. Данная научная работа предлагает новую стратегию, основанную на сети беспроводных датчиков (WSN) для достижения высокой точности позиционирования для подвижных объектов. В первую очередь, была смоделирована сеть радиально базисных функций для приближенной оценки местоположения в неизвестном помещении. Затем, из-за недостатка нахождения силы мгновенно полученного сигнала (RSS), мы представляем достоинства среднего RSS для снижения шумов и исключения временных APs. В дополнение, мы использовали Zigbee для реализации беспроводной IPS со сравнительно малой ценой. В заключение, с целью достижения оптимальной точности, мы приспособили несколько одинаковых сетей внутри одной окружающей среды. Эксперименты продемонстрировали эффективное улучшение существующей точности IPS со средней ошибкой 2.8 метра по сравнению с другими подходами. В свою очередь, с российской стороны была произведена апробация указанного подхода в интересах развития двухстороннего сотрудничества в рамках совместных инициативных исследований Университета Чунг Хуа и творческого коллектива на базе ИПУ РАН.

Ключевые слова: навигация в помещении, сеть беспроводных датчиков, сеть радиально-базисных функций, ZigBee, IPS, WSN

1. Введение

Сравнительно недавно стали активно изучаться применения сетей беспроводных датчиков (WSN). WSN это тип технологии беспроводной коммуникации для сбора информации об окружающем мире в сети, составленной из большого количества беспроводных датчиков с низким энергопотреблением и стоимостью, использующих самоорганизующуюся связь. Такой подход может быть применен в различных областях, таких как: медицина, военное дело, контроль промышленных объектов, умные дома и так далее. Так или иначе, среди и помимо многочисленных применений сети беспроводных сенсоров, исследование проблемы позиционирования становится актуальным, поскольку сенсоры должны прежде всего определить собственное местоположение. Больше внимания было уделено IPS автономного робота несколько лет назад. Автономного робота можно использовать для исследования, поиска, картирования и спасательных работ (например, на месте катастрофы в Фукусиме в 2011 году) в неизвестном помещении без использования человеческого вмешательства с неопределенной опасностью. Наиболее известной технологией позиционирования вне помещений является GPS; при этом позиционируемый объект должен находится в прямой зоне видимости спутников. При позиционировании в помещениях с помощью WSN технологии несколько сенсоров устанавливаются в известных позициях для нахождения позиций принимающих узлов. Существуют два популярных метода для идентификации по радиочастоте (RFID): позиционирование на интервальное и неинтервальное позиционирование. RFID это

пассивное устройство, применяемое, в основном, для позиционирования на коротких дистанциях, при этом требуется, чтобы робот имел дополнительные встроенные сенсоры и был близко к цели (с меткой). Bluetooth очень похож на RFID в плане обслуживания малых областей. ZigBee обеспечивает позиционирование на средних расстояниях, потребляя мало энергии, имея небольшую стоимость малую сложность и хорошую масштабируемость. В наши дни многие технологии позиционирования включены в беспроводные сети включающие время прихода (TOA), разницу во времени прихода (TDOA), угол прихода (AOA), DOA и «мощность полученного сигнала» (RSS). Так как система позиционирования основана на стандарте IEEE 802.11, то требуется больше электричества, и это не оптимальное решение для долгосрочного отслеживания, например, логистики или технологии умного дома. Кроме того на точность позиционирования влияет количество принимающих узлов, что может очень много стоить. Данная работа направлена на создание системы IPS робота с определенной точностью. Основной упор делается на использование сети ZigBee с низким энергопотреблением и хорошей масштабируемостью, а так же на техники измерения RSS, в которых набор измеренных мощностей беспроводного сигнала дополняется собственной способностью чипа с использованием RBFN сети для RSS позиционирования. Как показывают эксперименты, средняя ошибка позиционирования составляет 1.47 метра, которая достигается, когда приближение оценка местоположения выполняется на основе нескольких RBFN, и общая средняя ошибка оказывается в пределах 2.8 метров.

Структура данной работы выглядит следующим образом. Соответствующие теоретические сведения приведены в разделе 2.3. В разделе 3 дается ссылка на результаты экспериментов и сравнительный анализ, а заключение находится в разделе 4. Доклад основан на работе [1].

2. Подготовительные мероприятия

2.1. Технологии позиционирования. Наиболее часто используемые IPS являются инфра-красными (IR), ультразвуковыми или использующими беспроводные локальные сети (WLAN). Сравнение показано в табл. 1.

Существуют 4 наиболее часто используемых технологии: AOA, TOA, TDOA и RSS (Кеген, 2009). AOA определяет местоположение подвижного объекта на основании угла направления полученного сигнала, как показано на рисунке 1.

Так или иначе, AOA в большей степени зависит от точности измерения направления, что резко увеличивает стоимость. TOA подразумевает подсчет расстояния между обнаруженной целью и маяком на основе времени передачи сигнала, и обнаруживает цель используя алгоритм, показанный на рисунке 2.

В методе TDOA два сигнала передаются из узла передачи с разными скоростями передачи одновременно, а затем производится подсчет рассто-

	Принцип действия	Достоинства / Недостатки
IR	Измеренное расстояние вычисляется по временной задержке полученного излучения	Восприимчивы к солнечной интерференции и их точность зависит от угла излучения
Ультразвуковые	После излучения ультразвуковой волны наблюдаются различные профили отраженной волны	Восприимчивы к интерференции окружающей среды
WLAN	Оценка местоположения основанная на переданном и полученном беспроводном сигнале (RFID и Bluetooth)	Беспроводной сигнал восприимчив к интерференции

Таблица 1: Сравнение техник позиционирования в помещениях.

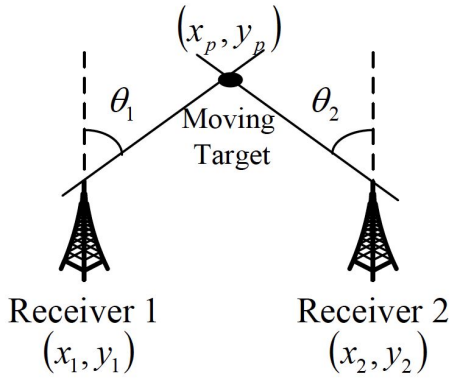


Рис. 1: Конфигурация локализации AOA

яния между этими двумя узлами как показано на рисунке 3. Временная синхронизация необходима и для TDOA и для TOA. Так или иначе, достоинством TDOA является использование относительного времени прибытия вместо абсолютного времени прихода (TOA) для уменьшения ошибки. Кроме того, недостатком является необходимость в дополнительном T/R устройстве.

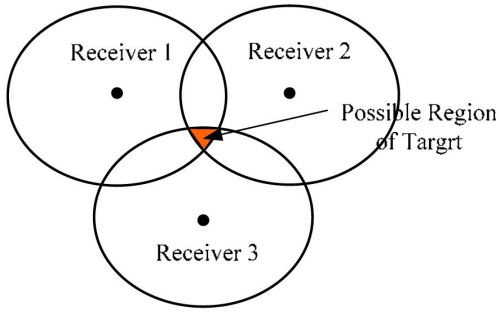


Рис. 2: Конфигурация локализации TOA

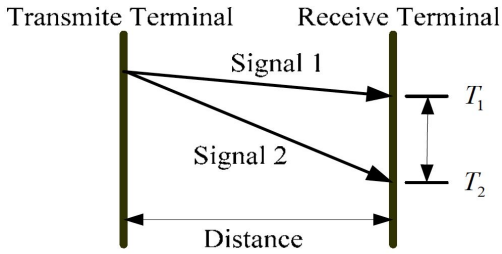


Рис. 3: Конфигурация локализации TDOA

RSS определяет местоположение с помощью нескольких принимающих узлов, определяющих расстояния между принимающими и передающими узлами.

2.2. Структура ZigBee. Питание для сенсора является ключевым элементом беспроводной сети, особенно если это касается установки большого количества сенсоров вне помещения. В связи с этим появился ZigBee в качестве беспроводной сети с низким энергопотреблением и низкой стоимостью. Протокол стандарта IEEE 802.15.4 показан на рисунке 4 (Лу, 2008).

2.2.1. Топология сети

В терминах функционирования программного обеспечения, узлы сети могут быть разделены на координатор PAN, координатор и конечное устройство, как показано в таблице 2.

IEEE 802.15.4 обеспечивает три базовых типа топологий сети, показанных на рисунке 5.

2.2.2. Структура приложения

Существуют три уровня в конфигурации программного обеспечения в узле IEEE 802.15.4, которые показаны на рисунке 6.

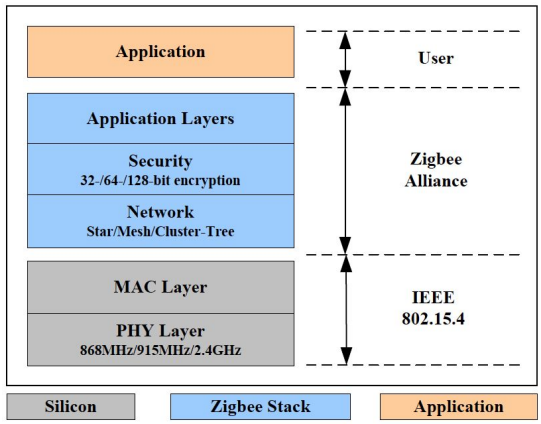


Рис. 4: Структура IEEE 802.15.4/Zigbee

	Функция узла в приложении	Тип узла
Координатор PAN	Присвоение идентификатора PAN сети и управление. Поиск частоты и назначение. Способность связи с другим устройством. Способность маршрутизации. Прием и передача пакетов данных.	FFD
Координатор	Способность связи с другим устройством. Способность маршрутизации. Прием и передача пакетов данных.	FFD
Конечное устройство	Связь с координатором PAN и координатором. Прием и передача пакетов данных.	FFD/RFD

Таблица 2: Функции узлов.

Взаимодействие между приложениями осуществляется с помощью стека API и стека слоев IEEE 802.15.4. Диалоговый доступ приложений к аппаратуре осуществляется встроенным API периферийного оборудования и встроенного в аппаратуру периферийного оборудования. На аппаратном уровне все виды прерываний могут быть сгенерированы и посланы в любой элемент приложения через коды прерываний. Кроме того, следу-

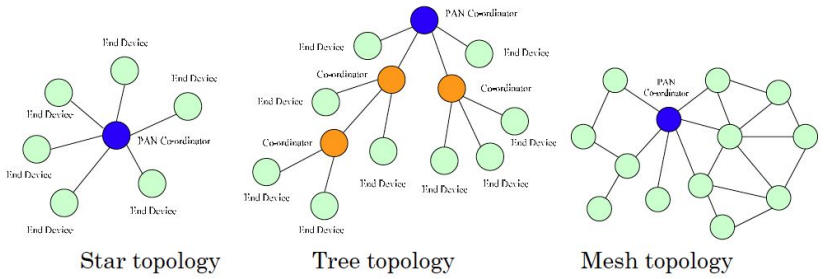


Рис. 5: Топология сети

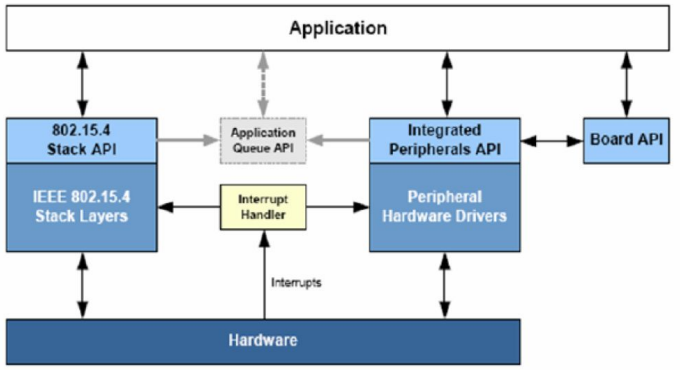


Рис. 6: Базовая конфигурация приложения IEEE 802.15.4

ющие особенности могут быть использованы при разработке приложений: Clear Channel Assessment (CCA), Link Quality Indicator (LQI). Значение LQI больше значения RSS, так, чтобы обеспечивалось большее разрешение.

2.3. Нейросетевой алгоритм.

2.3.1. Алгоритм соревновательного обучения

Искусственная нейронная сеть (нейросеть) имеет возможность обработки информации, схожую с биологическими нейронами. Это позволяет разрешать нелинейные проблемы. Однослойная нейросеть, основанная на алгоритме соревновательного обучения, показана на рисунке 7. Данный обучающий алгоритм обычно используется для кластерного анализа, который заключается в определении структуры и кластерных отношений данных без какой-либо предварительной информации о классах. Если мы предположим, что N-мерный вход может быть представлен как X:

$$X = [x_1, x_2, ..., x_N]^T \tag{1}$$

Вектор весов j-го нейрона выглядит следующим образом:

$$w_j = [w_{j1}, w_{j2}, ..., w_{jN}]^T, j = 1, 2, ..., K \tag{2}$$

где K – это количество выбранных нейронов.

Обучающий алгоритм приведен ниже.

Шаг 1: Соревновательная фаза: Выбирается победитель Искусственный нейрон с наименьшим Евклидовым расстоянием между входным вектором X и весовым вектором w_j является победителем:

$$d(X) = \min_j ||X - w_j||, j = 1, 2, ..., K \tag{3}$$

Шаг 2: Фаза вознаграждения: Изменяется вектор весов победителя

$$w_j(n+1) = w_j(n) + \eta(X - w_j(n)) \tag{4}$$

где η и n это количество итераций скорости обучения и обучающего процесса.

Шаг 3: Итерационная фаза: Если количество итераций достигло верхнего предела, обучение необходимо остановить. Иначе, обучение продолжается с соревновательной фазы.

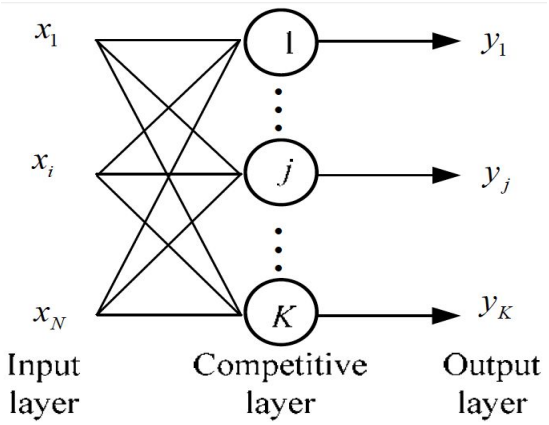


Рис. 7: Сеть соревновательного обучения

2.3.2. Сеть радиально базисных функций (RBFN)

RFBN это многослойная нейронная сеть прямого распространения с возможностью быстрого обучения. На рисунке 8 изображена трехслойная RFBN сеть включающая в себя входной слой, скрытый слой и выходной

слой. Предположим, что входная размерность это p и количество нейронов в скрытом слое J, тогда выход сети будет следующим:

$$F(\underline{x}) = \sum_{j=1}^{J} w_j \phi_j(\underline{x}) + \Theta = \sum_{j=0}^{J} w_j \phi_j(\underline{x}) \tag{5}$$

где $\underline{x}$ представляет входной вектор, w_j представляет значение веса от

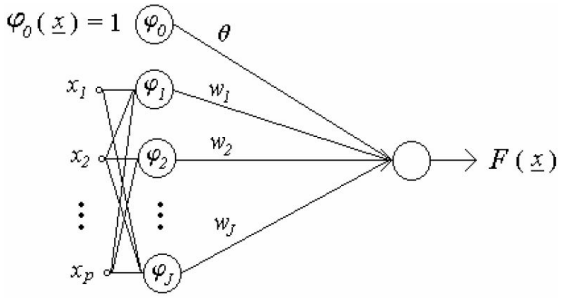


Рис. 8: Структура RBFN

j-го нейрона в скрытом слое до искусственного нейрона в выходном слое, $\Theta = w_0$ представляет регулируемый отступ, а $\phi_j(\underline{x})$ представляет базисную функцию от выходного значения j-го нейрона в скрытом слое, которая является Гауссовой функцией:

$$\phi_j(\underline{x}) = exp(-\frac{||\underline{x} - \underline{m}_j||^2}{2\sigma_j^2}) \tag{6}$$

где J представляет количество искусственных нейронов в скрытом слое, $\underline{m}_j$ представляет центральную позицию j-й Гауссовой функции, а σ_j представляет стандартное отклонение j-й Гауссовой функции.

Две фазы алгоритма включены в RBFN. Обучение узлов в скрытом слое заключается в основном в использовании "non-supervision"для настройки параметров $\underline{m}_j$ и σ_j, в то время, как выходные узлы настраиваются адаптацией "supervision"метода LMS для параметров нейросети.

$$E(n) = \frac{1}{2}(y_n - F(\underline{x}_n))^2 \tag{7}$$

где $F(\underline{x}_n))$ это значение n-го выхода, а y_n представляет ожидаемое значение n-го входа. Учитывая выбранную базисную функцию в формуле (6), настроечные выражения для $\underline{w}$, $\underline{m}_j$ и σ_j запишутся как:

$$\underline{w}(n+1) = \underline{w}(n) + \eta(y_n - F(\underline{x}_n))\underline{\phi}(\underline{x}_n) \tag{8}$$

$$\underline{m}_j(n+1) = \underline{m}_j(n) + \eta(y_n - F(\underline{x}_n))w_j(n)\underline{\phi}(\underline{x}_n)\frac{1}{\sigma_j^2}(\underline{x}_n - \underline{m}_j(n)) \quad (9)$$

$$\sigma_j(n+1) = \sigma_j(n) + \eta(y_n - F(\underline{x}_n))w_j(n)\underline{\phi}(\underline{x}_n)\frac{1}{\sigma_j^2}||\underline{x}_n - \underline{m}_j(n)||^2 \quad (10)$$

где η - это коэффициент скорости обучения, в то время, как

$$\underline{\phi}(\underline{x}_n) = [\underline{\phi}_0(\underline{x}_n), \underline{\phi}_1(\underline{x}_n), ..., \underline{\phi}_J(\underline{x}_n]^T \quad (11)$$

$$\underline{\phi}_0(\underline{x}_n) = 1 \quad (12)$$

и

$$\underline{w}(n) = [\theta(n), w_1(n), ..., w_J(n)]^T \quad (13)$$

2.4. Предлагаемое IPS решение.

2.4.1. Концепция

IPS предлагается разрабатывать с использованием Zigbee и алгоритма оценки на основе RBFN используемый для достижения требуемой точности. Прежде всего, RBFN используется для моделирования окружающей среды. Она разделена на фазы офлайн и реального времени. Офлайн фаза включает в себя установку экспериментальной окружающей среды, выбора точек отбора и измерения. Фаза реального времени заключается в последующих проверках. Архитектура системы представлена на рисунке 9.

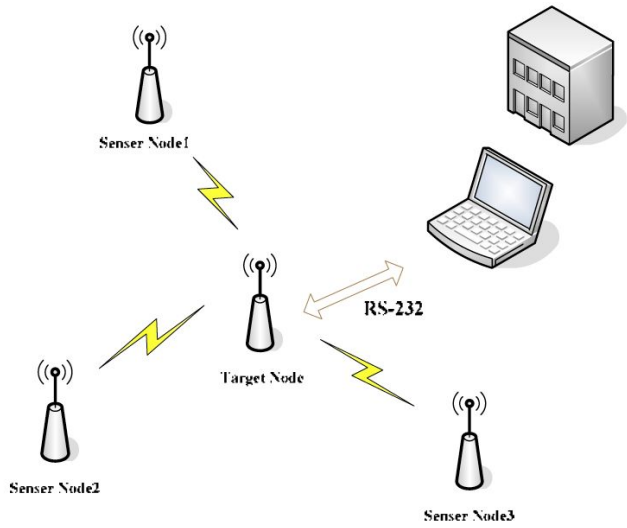


Рис. 9: Предлагаемая конфигурация локализации с помощью Zigbee

Метод позиционирования показан на рисунке 10. Он заключается в размещении трех сенсоров с известными позициями в помещении, а затем в использовании мощности сигнала, переданного с сенсора к цели (роботу) для оценки позиции робота.

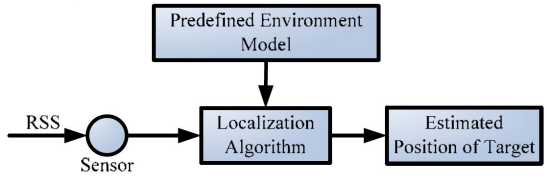


Рис. 10: Конфигурация RSS локализации

2.4.2. Препроцессинг полученного сигнала

Сигналы, переданные сенсором, должны быть получены в некоторый период времени. Иными словами, измерение сигнала восприимчиво к влиянию окружающей среды, таким, как хождение людей и т.д. Влияние таких интерференционных факторов может быть снижено адаптацией препроцессинга полученного сигнала, как показано на рисунке 11.

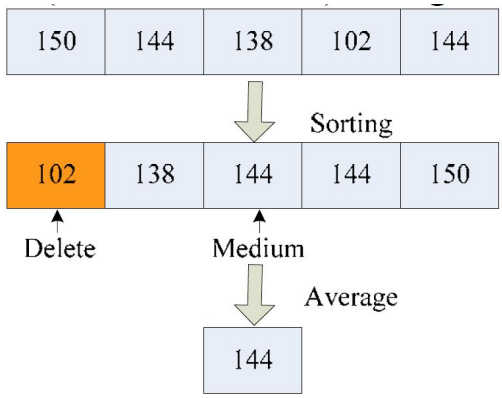


Рис. 11: Препроцессинг сигнала Zigbee

2.4.3. Общая система

Комплект для разработки Zigbee FT-6200 разработанный «Fontal» был адаптирован для измерений LQI, а оценочным алгоритмом является RBFN.

Общая система изображена на рисунке 12.

Процедура позиционирования может быть разделена на два шага.

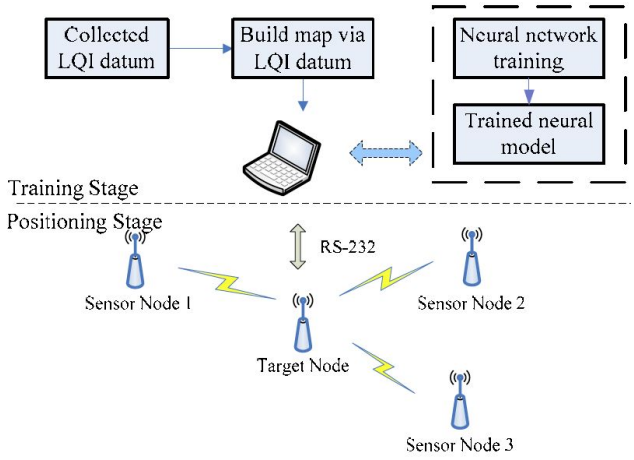


Рис. 12: Блок-схема функционирования предложенной системы

(1) Сбор средних значений LQI окружающей среды. После этого, они могут быть поданы на вход RBFN для обучения. Кластерная классификация производится для всех значений LQI в базе данных с помощью соревновательного обучения для определения параметров RBFN.

(2) Значения LQI в реальном времени будут переданы в обученную RBFN для проверки, как это показано на рисунке 13.

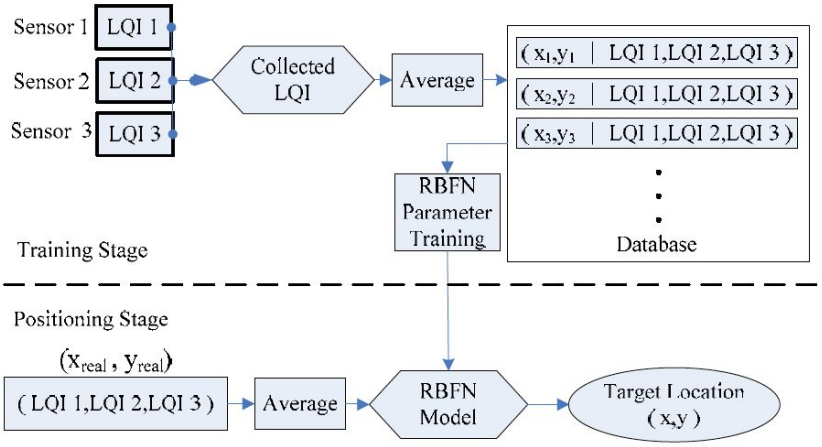


Рис. 13: Обработка данных

2.4.4. Основанный на RBFN алгоритм IPS

- Алгоритм соревновательного обучения

Соревновательное обучение нужно для нахождения похожих характеристик и отношений среди необозначенных наборов образцов для определения вектор центра кластера обучающих образцов. Входами на рисунке 13 являются $(x, y|LQI1, LQI2, LQI3)$, где (x, y) представляет координаты точки, которые необходимо измерить. Следуем шагам алгоритма в разделе 2.3. Число нейронов в наборе в скрытом слое и образец выбраны на рисунке 8, где

$$X = [LQI1, LQI2, LQI3]^T \tag{14}$$

В дополнение, вектор центра кластера начальной конфигурации выбирается из входного обучения в первый раз. Лидирующий нейрон в скрытом слое определяется по формуле (3). Весовой вектор лидирующего нейрона настраивается по формуле (4). Вычисленный вектор центра кластера $\underline{m}_j(n)$ и стандартное отклонение σ_j сохраняются затем для конфигурации RBFN.

- RBFN

Далее следует шаг обучения модели RBFN для локализации (см. рис. 8):

Шаг 1: Результат соревновательного обучения используется для инициализации вектора центра кластера $\underline{m}_j(n)$ и стандартного отклонения σ_j RBFN. Весовой вектор $\underline{w}$, скорость обучения η и условия сходимости так же установлены. Правильными инициализирующими данными для RBFN считаются 6000 итераций, 1 выход, 3 скрытых нейрона, 3 входных вектора и начальные веса -1 1.

Шаг 2: В этом шаге прямого распространения входной обучающий образец подается на вход нейронной сети в n-й итерации так, что выход этого обучения может быть вычислен.

Шаг 3: В этом шаге обратного распространения выполняется вычисление средней квадратичной погрешности по ожидаемым выходам нейронной сети.

Шаг 4: Параметры RBFN, такие, как $\underline{w}$, $\underline{m}_j(n)$, σ_j и θ настраиваются одновременно в направлении, обратном градиенту функции среднего квадрата.

Шаг 5: Проверка, были ли достигнуты условия сходимости. В противном случае, происходит возврат к шагу 2 для продолжения обучения.

Шаг 6: Вывод оценочных координат цели.

Необходимо заметить, что количество векторов центров определено количеством входов. Было выяснено, что в этом случае обучение нейронной сети будет работать лучше.

3. Вопросы практического применения

Программная и аппаратная реализация, а также экспериментальная статистика приводятся в опорной работе [1], использованной при подготовке совместного доклада на данную российскую конференцию.

В результате проведенных тайваньскими разработчиками экспериментов и последующего анализа влияние размера обучающей выборки, наряду с детальной практической реализацией, были сделаны следующие основные выводы: хотя использование большего количества RBFN повышает точность, исследование изменения значения LQI полученного в помещении все еще требует определения наиболее соответствующего количества RBFN. Использование чрезмерного количества RBFN будет занимать слишком много памяти, если не удастся достичь точности позиционирования.

4. Заключение

Тайваньскими исследователями была предложена достаточно простая, но весьма эффективная IPS, использующая Zigbee [1]. Сравнение IP алгоритма с несколькими RBFN позволило достичь большую точность, чем у других рассмотренных смежных подходов. Точность также может быть повышена, в рассматриваемых условиях, путем настройки числа скрытых слоев в определенных пределах. Данный метод может быть использован не только для IPS роботов, но и для других прикладных сервисов и разнообразных целей, предоставляя удобство локального позиционирования, практическую реализуемость, а также низкую стоимость. Более того, представленные результаты исследования возможностей модификации радиально-базисной функции при построении системы локализации в здании были применены при подготовке инициативной совместной российско-тайваньской заявки на получение поддержки МНТ-РФФИ (MOST-RFBR).

5. Благодарности

Данное исследование изначально было творческими работами и было поддержано National Science Council, ROC, Grant NSC 100-2221-E-216-007. Авторы благодарят рецензентов и редактора за их значимые комментарии и мнения о данной работе.

Перевод статьи профессора Ле Луо как доклада на данную международную конференцию был осуществлен Г. Антиохом (НИУ МАИ), использовавшим указанный подход в обзоре методов своей выпускной квалификационной работы бакалавра [2-5].

БИБЛИОГРАФИЯ

1. L.Luoh, ZigBee-based intelligent indoor positioning system soft computing, Department of Electrical Engineering, Chung Hua University, Hsin-Chu, Taiwan
2. Белавкин П.А., Федосеев С.А., Рожнов А.В., Лобанов И.А. Исследование стратегической мобильности проблемно-ориентированных систем управления и их позиционирование в условиях развития информационного пространства // Известия ЮФУ. Технические науки. 2013. Тема-

тический выпуск "Перспективные системы и задачи управления 2013, № 3. С. 211-217.
3. Барышев П.Ф., Рожнов А.В., Губин А.Н., Лобанов И.А. Обоснование информационно-аналитической системы в развитии методов и моделей согласования иерархических решений // Динамика сложных систем — XXI век. 2014. № 3. С. 43-52.
4. Рожнов А.В., Антиох Г.М. и др. Системная интеграция направлений научной деятельности в условиях формирования предынтеллектуальной инфраструктуры // Информационно-измерительные и управляющие системы. 2014. № 11. С. 59-64.
5. Патент 141445 РФ, МПК G05B23/02. Инструментально-моделирующий комплекс исследования процессов управления и диспозиции сложного динамического объекта в группе / Абросимов В.К., Захаров В.Л., Лобанов И.А., Рожнов А.В., Бажанов О.В. заявитель и патентообладатель ФГБОУ ВПО МАИ (НИУ). - №2013158438/08, заявл. 27.12.2013; опубл. 10.06.2014, бюл. №16.

CREATION METHODOLOGY OF FUNCTIONAL MODEL OF COMPUTER NETWORK OF THE REAL DIMENSION

L. I. Abrosimov[1]
[1]National research university "MPEI", Moscow, Russia

Abstract

The article discusses the original model of polynomial approximation using the formalized system of the isomorphic transformations providing communication of real devices and communication lines of the computer networks (CN) with model elements is considered. The introduced transformation rules provide algorithmization of process of construction and the decision of system of the nonlinear equations displaying communication of intensity of the movement and processing of messages in CN model as for elements both for opened and so as for the closed routes of the movement of messages. The developed algorithms of the solution of the equations allow to determine productivity, reaction time, loading coefficients for the considered CN option.

МЕТОДОЛОГИЯ ПОСТРОЕНИЯ ФУНКЦИОНАЛЬНОЙ МОДЕЛИ ВЫЧИСЛИТЕЛЬНОЙ СЕТИ РЕАЛЬНОЙ КОНФИГУРАЦИИ

Л. И. Абросимов[1]
[1]Национальный исследовательский институт "МЭИ", Москва, Россия, AbrosimovLI@mpei.ru

Аннотация

В статье рассматривается оригинальная модель полиномиальной аппроксимации, использующая формализованную систему изоморфных преобразований, обеспечивающие связь реальных устройств и линий связи вычислительных сетей (ВС) с элементами модели. Введенные правила преобразований обеспечивают алгоритмизацию процесса построения и решения системы нелинейных уравнений, отображающих связь интенсивностей движения и обработки сообщений в элементах модели ВС, как для разомкнутых, так и для замкнутых маршрутов движения сообщений. Разработанные алгоритмы решения уравнений позволяют определять производительность, время реакции, коэффициенты загрузки для рассматриваемого варианта ВС.

Ключевые слова: Вычислительные сети, модель полиномиальной аппроксимации, метод контуров, методология построения модели ВС

1. Введение

Разработка и модернизация эффективно функционирующих региональных и корпоративных сетей ЭВМ является сложным и трудоемким процессом, так как разработчику необходимо учитывать много критериев эффективности, большое количество ограничений, громадное количество устройств, каждое из которых обладает широким спектром технических характеристик, которые часто зависят от местоположения и взаимосвязей с другими устройствами.

Кроме того, особенностью современных вычислительных сетей (ВС) является рост потока заявок на обслуживание.

Разработчику приходится постоянно сталкиваться со следующими до сих пор не решенными проблемами.

- Упорядочить и постоянно обновлять исходные данные, которые могут понадобиться при разработке.
- Для обеспечения качественного обслуживания пользователей необходимо согласовать производительность ВС с постоянно возрастающими требованиями пользователей.
- Определить производительность для различных типов топологий и архитектур сетей ЭВМ.

Для успешного решения указанных проблем необходима компьютеризация не только процессов хранения и поиска данных, но и процессов построения моделей.

Для оценки производительности ВС необходимы модели, которые, во-первых, учитывают реальную размерность ВС, во-вторых, базируются на аналитических соотношениях, которые позволяют оперативно и детально оценивать предельную производительность ВС, в третьих, позволяют однозначно, с необходимой детализацией соотносить результаты, определенных для элементов моделирования, с параметрами устройств ВС.

Таким требованиям удовлетворяет модели, построенные с использованием метода контуров [1], в котором последовательно выполняются: описание топологической структуры ВС; построение функциональной структуры ВС; формализованного детализированного описания потока заявок на обслуживание; составление и решение линейных и нелинейных уравнений для определения требуемых вероятностно-временных характеристик функционирования ВС.

2. Постановка задачи

Для создания полностью автоматизированной процедуры требуется решить задачу изоморфного преобразования топологической, логической и функциональной структур ВС, а именно:

- сформулировать правила логических преобразований и формализовано записать топологическую структуру ВС в форме топологической матрицы $\mathbf{T}$;

- сформулировать правила логических преобразований матриц **T** в **L** и формализовано записать логическую структуру в форме логической матрицы **L**;
- сформулировать правила логических преобразований матриц **T** и **L** в **F** и формализовано записать функциональную структуру в форме функциональной матрицы **F**;
- определить требуемые функциональные характеристики.

3. Этапы методологии

Кратко рассмотрим состав и последовательность выполняемых действий (компьютеризированных процедур) при разработке и модернизации региональных и/или корпоративных ВС.

Подготовительные процедуры.

1. Формирование базы данных технических и программных средств (БДС), в результате которого получают:

- классификации устройств (У), линий связи (ЛС) и программного обеспечения (ПО);
- технические характеристики У, Л, и ПО;
- Экономические характеристики У, Л, и ПО.

2. Формирование базы данных трафика (БДТ), в результате которого получают:

- классификации клиентов и серверов;
- схема взаимодействия клиентов и серверов;
- информационные характеристики взаимодействия;

Процедуры, выполняемые при разработке ВС.

3. Разработка принципиальной схемы соединения устройств посредством линий связи, с указанием идентификаторов $(A.r)$, $(B.k)$ и типов ω устройств, типов линий θ связи и каналов γ связи. Разработанная принципиальная схема ВС может быть представлен в виде визуального отображения на экране монитора разработчика и в виде списка соединений устройств линиями связи.

Для обеспечения изоморфизма топологической, логической и функциональной моделей при выполнении формирования и преобразования матричных представлений моделей используется массив М метаданных, который состоит из набора таблиц 1М, 2М, 3М, 4М.

4. Формирование топологической матрицы **T** , которое выполняется по соотношению (1) и отображает все физические элементы ВС, к которым относятся устройства множества **У** (рабочие станции, серверы, коммутирующие устройства) и множества **Л** (линий связи).

$$T = ||t(A.k - B.r)|| = \begin{cases} \omega, R, & \text{если A = B;} \\ \theta, \gamma, & \text{если A} \neq \text{B.} \end{cases} \quad (1)$$

где ω - тип устройства, A и B - идентификатор, определяющий местоположение У $\in$ **У**, R - количество разъемов в устройстве У, θ - тип линии связи Л $\in$ **Л** (физическая реализация), γ - тип канала связи Л (дуплексный, полудуплексный, симплексный).

В таблице метаданных 1М каждому идентификатору $A.k-B.r$ ставится в соответствие порядковый номер i.

5. Формирование логической матрицы **L**, которая предназначена для отображения узлов, соответствующих тем устройствам, представленным в матрице T, которым в ВС присвоены логические имена, адреса и номера интерфейсов.

Множество узлов U, отображаемых в матрице L, является подмножеством устройств **У** (U $\in$ У), описываемых в топологической матрице T, поэтому каждому узлу U_A соответствует устройство Y_A

Для описания логического соединения узлов U , соответствующих ВС, используются дуги D, которые могут быть взвешены в соответствии с выбранной метрикой маршрутизации.

Основным параметром дуги D является тип γ канала связи, организованного между узлами U_A и U_B.

На основании матрицы **T**, списков $\{U\}$ и $\{D\}$ можно сформировать матрицу L логической структуры, в которой:

$$T = ||l(A-B)|| = \begin{cases} \omega, & \text{если A = B;} \\ \gamma, & \text{если A} \neq \text{B.} \end{cases} \qquad (2)$$

В таблице метаданных 2М каждому идентификатору A ставится в соответствие порядковый номер i.

6. Решение задачи выбора кратчайших маршрутов, используя данные логической матрицы **L** и выбранный критерий эффективностью. Результатом расчета является таблица маршрутизации между всеми адресуемыми устройствами ВС.

7. Формирование функциональной матрицы **F**, которая является основой математической модели ВС и предназначена для отображения функциональных элементов **Е**, соответствующих устройствам **У** и линиям **Л**, которые задерживают транзакции при обработке и оказывают существенное влияние на производительность ВС.

Каждый элемент Е описывается соотношениями моделей массового обслуживания, является ориентированным и имеет вход (1) и выход (2). Подмножество ЕУ $\subset$ Е является подмножеством элементов отображающих У и каждый элемент ЕА подмножества ЕУ имеет идентификатор А, введенный при формировании топологической структуры. Для устройств и элементов используется одна и та же система индикаторов. Подмножество ЕК $\subset$ Е является подмножеством элементов отображающих каналы К и каждый элемент ЕК подмножества ЕК имеет идентификатор $(A.k - B.r)$.

Для формирования матрицы F необходимо использовать множество ES переходов, отображающих соединение выходов (2) элементов подмножеств ЕУ с входами (1) элементов подмножества ЕК и соединение выходов (2) элементов подмножеств ЕК с входами (1) элементов подмножества ЕУ.

На основании матрицы **T**, можно сформировать матрицу **F** функциональной структуры, в которой

$$F = f\{(A.r - B.k)i, (A.r - B.k)j\} = \begin{cases} \omega, & \text{если A = B и } E \in EY; \\ \omega, \gamma, & \text{если A} \neq \text{B и } E \in EK; \\ 1, & \text{если A} \neq \text{B и } E \in ES. \end{cases} \quad (3)$$

где i и j - соответственно, строки и столбцы в матрице F, a, r и k - разъемы устройств A и B.

В таблице метаданных 3М каждому идентификатору устройства $-$ и каждому идентификатору линии связи $A.k - B.r$ ставится в соответствие порядковый номер элемента E, а каждому входу/выходу элемента E ставится в соответствие порядковый номер i.

8. Определение состава контуров $q \in Q$, каждый из которых представляет собой последовательность элементов ЕУ и ЕК, которые моделируют обслуживание узлами и линиями ВС потока заявок на обслуживание клиентом и сервером с интенсивностью λ_{iq}. Для формирования состава контура используется таблица маршрутизации между всеми адресуемыми устройствами ВС, которая дополняется элементам ЕК , которые не являются адресуемыми элементами. Для решения задачи используется модифицированная методика выбора кратчайших маршрутов [1].

9. Формирование линейных уравнений (ЛУ) состоит в использовании для каждого элемента E_i контура $q \in Q$. Решение ЛУ состоит в вычислении для каждого элемента контура коэффициентов a_{ij} базовой интенсивности, которые показывают, какая часть базовой интенсивности переходит от узла i к узлу j, т.е.:

$$\lambda_{ij} = \lambda_0 a_{ij} \quad (4)$$

10. Формирование системы нелинейных уравнений (НЛУ) использует для каждого контура $q \in Q$ соотношение (5), если элемент E_{iq} моделируется СМО $M/M/1/\infty$.

$$n_q - \sum_{E_i \in W_q} n_{iq}, \quad (5)$$

где

$n_q = \frac{\sum_{\varphi}^{d_{iq}} \frac{\lambda_{iq}}{\mu_{iq\varphi}}}{1-\gamma_i \sum_{q \in Q_i} \sum_{\varphi=1}^{d_{iq}} (\frac{\lambda_{iq}}{\mu_{iq\varphi}})}$;

W_q - маршрут потока заявок контура $q \in Q$;

Q_i - множество контуров, обслуживаемых i-м элементом;

d_{iq} - количество фаз обработки контура q i-м элементом;

Количество уравнений равно числу контуров $q \in Q$.

11. Решение системы НЛУ для ВС произвольной конфигурации целесообразно производить, используя метод тангенсов. [1] В результате решения НЛУ получаем значения базовых интенсивностей λ_{0q} для каждого контура $q \in Q$.

12. Определение функциональных характеристик моделируемой ВС. Среднее время t_q^s доставки сообщений контура q в разомкнутой сети:

$$t_q^s = \sum_i t_{iq}, i \in W_q, \tag{6}$$

Среднее время отклика t_q^d сообщений замкнутого контура q, содержащего n_q сообщений, каждое из которых генерируется i-м элементом:

$$t_{iq}^d = (n_q/\lambda_{iq}) - (1/\mu_{iq}), \tag{7}$$

Коэффициент ρ_i загрузки элемента i интенсивностями соответствующих контуров Q_i:

$$\rho_i = \sum_i \rho_{iq} = \sum_i \sum_{\varphi=1}^{d_{iq}} (\frac{\lambda_{iq}}{\mu_{iq\varphi}}) \tag{8}$$

Следует подчеркнуть, что приведенные функциональные характеристики определены для функциональной модели ВС. Однако, выполненные преобразования матриц **T**, **L**, **F** обладают изоморфизмом, т.е. каждому элементу функциональной матрицы **F** соответствует техническая реализация в виде устройства или линии связи которые занесены в базу данных БДС, поэтому для повторного расчета требуется меньше вычислительных ресурсов.

4. Пример

В качестве примера рассмотрим изображенный на рис. 1 фрагмент принципиальной схемы корпоративной ВС, который состоит из 5-ти устройств: рабочих станций $A = 1, 2; (\omega = 1)$, сервера $A = 5; (\omega = 2)$, адресуемого коммутатора $A = 3; (\omega = 3)$, неуправляемого коммутатора $A = 4(\omega = 4)$, который образует сегмент ЛВС, использующий единую среду передачи данных.

В качестве Л, соединяющей У, используется витая пара ($\theta = 1$), протокол передачи данных - дуплексный канал ($\gamma = 1$).

В качестве примера на рис. 3 изображена матрица **L** логической структуры для фрагмента корпоративной ВС, изображенного на рис.1 .

На рис. 3 изображены дуги, отображающие наличие соединения между узлами 1, 3, 5, которые соответствуют устройству 1(рабочей станции), коммутатору 3 и устройству 5 (серверу), изображенным на рис.1. Как видно

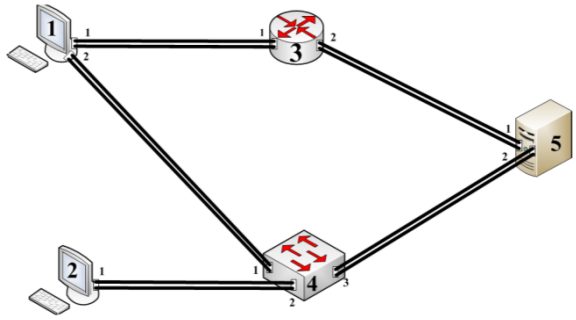


Рис. 1: Фрагмент корпоративной ВС

(A, r)	1.1	1.2	2.1	3.1	3.2	4.1	4.2	4.3	5.1	5.2
i	1	2	3	4	5	6	7	8	9	10

Таблица 1: Сравнение характеристик системы.

T=

	1	2	3	4	5	6	7	8	9	10
1	1,2			1,1						
2		1,2				1,1				
3			1,1				1,1			
4	1,1			3,2,						
5					3,2				1,1	
6		1,1				4,3				
7			1,1				4,3			
8								4,3		1,1
9					1,1				2,2	
10								1,1		2,2

Рис. 2: Матрица T топологической структуры ВС

из рис. 4, дуги отображают наличие логических соединений между узлами 1 и 2, 1 и 5, 2 и 5 организованных по нескольким (двум) линиям связи, соединенных через коммутатор 4, которые не вошли в состав логического описания ВС, так как коммутатор 4 не имеет логического адреса.

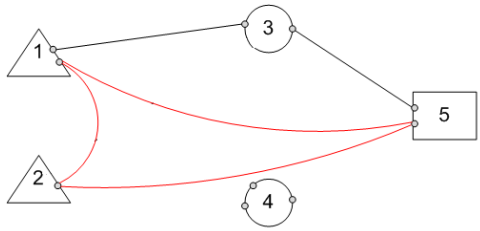
A	1	2	3	5
i	1	2	3	4

Таблица 2: 2M

L

	1	2	3	4
1	**1**	**1**	**1**	**1**
2	**1**	**1**		**1**
3	**1**		**3**	**1**
4	**1**	**1**	**1**	**2**

(а)



(б)

Рис. 3: Матрица L логической структуры ВС (а) и дуги между узлами в логической структуре ВС (б)

В качестве примера на рис.5 показана функциональная структура ВС, которая отображает функциональные элементы Е, соответствующие устройствам **У** и линиям **Л**, которые задерживают потоки заявок при обработке и оказывают существенное влияние на производительность ВС. Для фрагмента корпоративной ВС, изображенного на рис.1 на рис.6 представлена матрица F функциональной структуры.

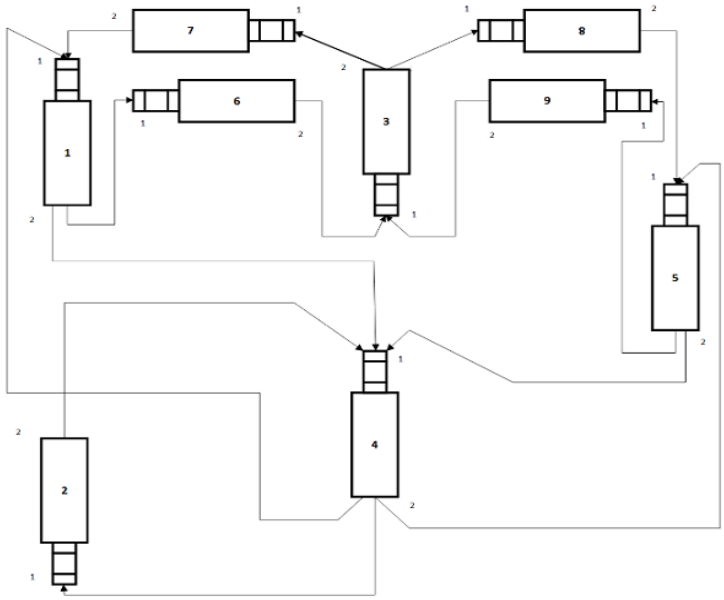


Рис. 4: Функциональная структура ВС

Для автоматизации преобразования матриц необходимо решить 2 задачи:

$(A.r - B.k)\alpha$	(1-1)1	(1-1)2	(2-2)1	(2-2)2	(3-3)1	(3-3)2
$(E)\alpha$	(1)1	(1)2	(2)1	(2)2	(3)1	(3)2
i	1	2	3	4	5	6
$(A.r - B.k)\alpha$	(4-4)1	(4-4)2	(5-5)1	(5-5)2	(1.2-3.1)1	(1.2-3.1)2
$(E)\alpha$	(4)1	(4)2	(5)1	(5)2	(6)1	(6)2
i	7	8	9	10	11	12
$(A.r - B.k)\alpha$	(3.2-1.1)1	(3.2-1.1)2	(3.2-5.1)1	(3.2-5.1)2	(5.2-3.1)1	(5.2-3.1)2
$(E)\alpha$	(7)1	(7)2	(8)1	(8)2	(9)1	(9)2
i	13	14	15	16	17	18

Таблица 3: 3М

1. Преобразовать матрицу **T** топологической структуры в матрицу **L** логической структуры.

2. Преобразовать матриц топологической **T** и логической **L** структур в матрицу **F** функциональной структуры.

F=

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1		1																
2							1				1							
3				1														
4											1							
5						3												
6													1		1			
7								4										
8	1		1						1									
9							1			2							1	
10																		
11												1						
12					1													
13														1				
14	1																	
15																1		
16																		
17																		1
18					1													

Рис. 5: Матрица функциональной структуры F

В матрице **F** в диагональные ячейки с1 по 10 записаны значения ω ; в диагональные ячейки с 11 по 18 записаны значения θ - (тип линии связи Л $\in$ **Л**). Численные значения типов соответствуют принятым для примера классификациям.

Примеры процедур 8 – 12, выполняемых при разработке ВС, приводятся в докладе на конференции DCCN 2015.

ЛИТЕРАТУРА

1. Абросимов Л.И. Базисные методы проектирования и анализа сетей ЭВМ: учебное пособие / Л.И. Абросимов. - М.: Университетская книга, 2015. - 248 с. - ISBN 978-5-98699-153-5.

DETERMINATION OF COVERAGE AREA FOR SIGNAL OF 802.11 WIRELESS NETWORK

L. I. Abrosimov[1], *M. A. Rudenkova*[2]
National research university "MPEI", Moscow, Russia

Abstract

The article formulates the problem of determining the distance at which is provided a performance, defined by the 802.11 b and 802.11 g, with variation in the transmitter power and noise. The article describes the methodology of the study, in which first, for a given interval variable parameters are calculated using the formula of calculation of the range of communications and free space path loss FSL, and then conducted a series of experiments. The results showing substantial differences between theoretical and practical values of the boundaries of coverage area.

ОПРЕДЕЛЕНИЕ ГРАНИЦ УВЕРЕННОГО ПРИЕМА СИГНАЛА БЕСПРОВОДНОЙ СЕТИ СТАНДАРТА 802.11

Л. И. Абросимов[1], *М. А. Руденкова*[1]

Национальный исследовательский институт "МЭИ", Москва, Россия,
[1]AbrosimovLI@mpei.ru, [2]mpatu@yandex.ru

Аннотация

В статье формулируется задача определения расстояния, на котором обеспечивается пропускная способность, определяемая стандартом 802.11b и 802.11g, при варьировании мощности передатчика и уровня шума. Излагается методика исследования, при которой сначала для заданного интервала изменяемых параметров производится расчет, используя формулы определения дальности связи и потерь в свободном пространстве FSL, а затем проводится серия экспериментальных измерений. Приводятся результаты, свидетельствующие о существенных различиях теоретических и практических значений границ уверенного приема.

Ключевые слова: Граница уверенного приема сигнала стандарта 802.11

1. Введение

Непрерывное развитие беспроводных технологий приводит к повышению пропускной способности беспроводных сетей и увеличению зоны их

покрытия. Быстрое развитие и появление новых технологий приводит к возникновению проблем при проектировании корпоративной беспроводной локальной сети БЛВС. Перед проектировщиком БЛВС стоит задача определить зоны покрытия, создаваемые каждой сетью. Существуют различные программы для планирования географического размещения беспроводных точек доступа и роутеров. Программы Dlink Wi-Fi Planner Pro, Ubiquiti UniFi Enterprise WiFi System не отображают реальное возможное покрытие беспроводной сетью. В данной работе показываются различия в определяемых зонах покрытия сетью с экспериментальными данными.

Под термином граница уверенного приема принимаем расстояние, на котором не будет происходить снижение пропускная способность, вызванное потерями из-за удаленности приемника от передатчика.

Для обеспечения пропускной способности 11 Мбит/с (802.11b CCK-кодирование) минимальная мощность принимаемого приемником сигнала передатчика AP со стороны приемной станции должна составлять -76 dBm, а отношение мощности от соседнего канала передатчика другой станции должно быть не менее 35 dBm. Максимальная мощность -10dBm.

Для обеспечения пропускной способности 54 Мбит/с (802.11g ERP - OFDM) минимальная мощность принимаемого приемником сигнала передатчика AP со стороны приемной станции должна составлять -76 dBm, а отношение мощности от соседнего канала передатчика другой станции должно быть не менее 3 dBm. Максимальная мощность -20dBm.

2. Задача определения границ уверенного приема сигнала беспроводной сети стандарта 802.11

Рассматривается стенд беспроводной сети стандарта 802.11 (рис. 1), состоящий из беспроводного роутера AP1(Linksys WRT 350N) и мобильной станции STA. В роли мобильной станции используется:

- мобильный компьютер с беспроводным адаптером стандарта 802.11 и утилитой InSSIDer ;
- беспроводной тестер Fluke AirCheck Wi-Fi Tester.

Мониторинг и изменения параметров состояния и мощности роутера производиться посредством веб-интерфейса управления, входящего в состав ПО . роутера.

Требуется: Определить зависимость расстояния L от параметров, указанных в табл. 1.

3. Методика исследования

3.1. Утилита для анализа частотных диапазонов беспроводных сетей стандарта 802.11. Утилита InSSIDer является инструментом для поиска и сбора информации о беспроводных сетях стандарта 802.11, которые находятся в зоне расположения компьютера. Для найденных сетей

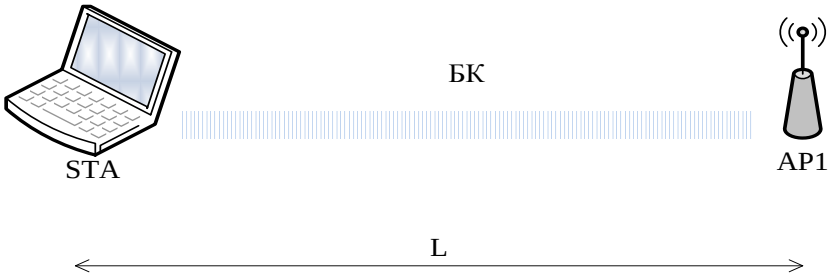


Рис. 1: Схема стенда измерения границ уверенного приема БЛВС

Расстояние L, м	Мощность передатчика мВт	Уровень шума, dBm	Стандарт БЛВС
x	a	b	802.11b
y	c	d	802.11g

Таблица 1: Зависимость расстояния от мощности передатчика и уровня шума

с помощью программы InSSIDer можно определить MAC-адрес устройства доступа из “соседней” сети, производителя данного устройства, канал, используемый данным устройством, идентификатор SSID или публичное название сети, тип безопасности. Кроме того, программа показывает уровень мощности принимаемого приемником сигнала от передатчика в dBm. Кроме того, с помощью данной программы можно посмотреть диаграмму “соседних” сетей, расположенных в диапазоне 2,4 ГГц и оценить загруженность каналов данного диапазона.

3.2. Тестер для беспроводных сетей Fluke AirCheck Wi-Fi Tester. Тестер AirCheck используется в корпоративных беспроводных сетях для проведения оперативного анализа проблем, возникающих после подключения пользователей в конкретной зоне сети, для проверки функционирования сети, для обнаружения беспроводного оборудования. В состав функций тестера входит определение параметров точек доступа таких как: частотный канал, уровень сигнала, имя или MAC-адрес точки доступа и параметры производителя оборудования.

4. Результаты

Для расчета дальности работы беспроводного канала связи в свободном пространстве, используется формула потерь в свободном пространстве FSL [дБ]:

$$FSL = 33 + 20(\log_{10} F + \log_{10} D); \quad (1)$$

где

F[МГц] - центральная частота канала работы беспроводной сети;

D[км] - расстояние между двумя точками (км).

Преобразуя (1), получаем формулу расчета дальности D_{calc} связи:

$$D_{calc} = 10^{(\frac{FSL}{20} - \frac{33}{20} - \log_{10} F)}. \quad (2)$$

FSL вычисляется по формуле:

$$FSL = Y - SOM; \quad (3)$$

где

SOM[дБ] Ц (System Operating Margin) запас в энергетике радиосвязи.

FSL определяется суммарным усилением системы Y [дБ], которое рассчитывается следующим образом:

$$Y = P_t + G_t + G_r - P_{min} - L_t - L_r; \quad (4)$$

где P_t [дБмВт] - мощность передатчика;

G_t [дБи] - коэффициент усиления передающей антенны;

G_r [дБи] - коэффициент усиления приемной антенны;

P_{min} [дБмВт] - чувствительность приемника на данной скорости;

L_t [дБ] - потери сигнала в коаксиальном кабеле и разъемах передающего тракта;

L_r [дБ] - потери сигнала в коаксиальном кабеле и разъемах приемного тракта.

Определение значения D_{STA} и D_{tester} производится экспериментальном путем: замером расстояния при уровне мощности приемного сигнала -76 dBm, например, эксперимент, приведенный на рис. 2.

Рис. 2: График зависимости расстояния от чувствительности приемника при мощности передатчика 2мВт (4dBm), уровень шума -95 dBm

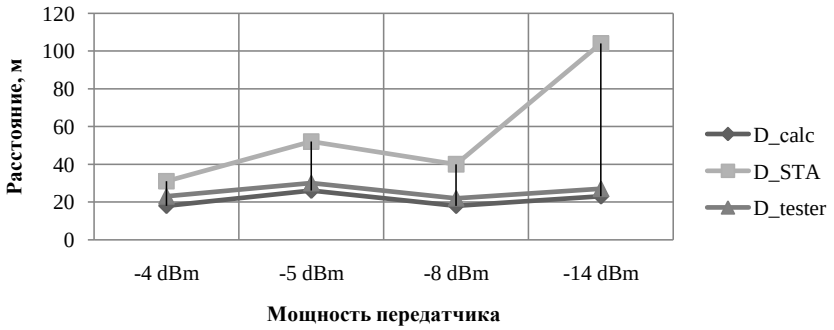


Рис. 3: График зависимости расстояния от мощности передатчика при уровне шума -96 - -91 dBm

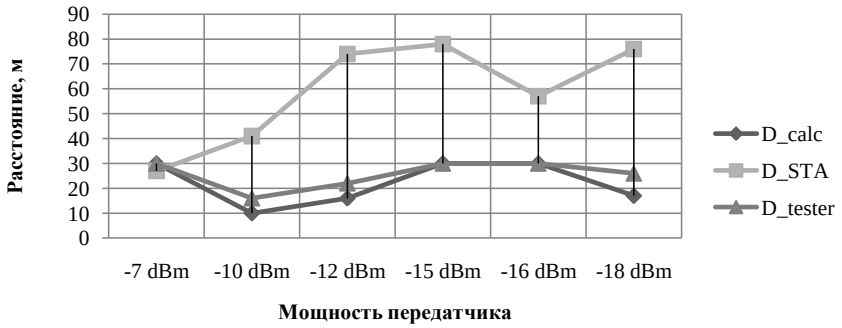


Рис. 4: График зависимости расстояния от мощности передатчика при уровне шума -90 - -89 dBm

По стандарту 802.11b и 802.11g значение границы уверенного приема(Coverage radius) составляет 38 м (см. рис. 6.), однако измеренные значения показатели отличаются от полученных данных (см. рис. 3 Ц 5).

Производители беспроводного оборудования в характеристиках своих устройств при установке специальных антенн обеспечивают дальность более 50м.

Расхождение рассчитанного значения границы уверенного сигнала с измеренным значением доказывает необходимость использования адаптивной настройки. Значение, полученное с помощью беспроводного тестера лучше, чем значение, полученное посредством STA, так как тестер обладает более чувствительным приемником. Однако, клиентам для прогнозирования границ уверенного приема важна чувствительность клиентской

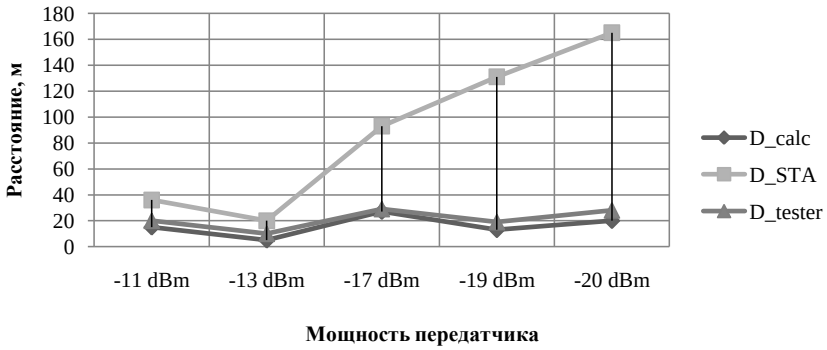


Рис. 5: График зависимости расстояния от мощности передатчика при уровне шума -88 - -82 dBm

	IEEE 802.11a	IEEE 802.11b	IEEE 802.11g	IEEE 802.11n
Frequency band	5.7 GHz	2.4 GHz	2.4 GHz	2.4 / 5 GHz
Average Theoretical speed	54 Mbps	11 Mbps	54 Mbps	600 Mbps
Modulation	OFDM	CCK modulated with QPSK	DSSS, CCK, OFDM	OFDM
Channel bandwidth	20 MHz	20 MHz	20 MHz	20 / 40 MHz
Coverage radius	35 m	38 m	38 m	75 m
Unlicensed spectrum	Yes (it depends on countries)	Yes	Yes	Yes (it depends on countries)
Radio Interference	Low	High	High	Low
Introduction cost	Medium-Low	Low	Low	High-medium
Device cost	Medium-Low	Low	Low	Medium
Mobility	Yes	Yes	Yes	Yes
Current use	Medium	High	High	High
Security	Medium	Medium	Medium	High

Рис. 6: Сравнение технологий 802.11

станции, обладающей характеристиками беспроводного адаптера аналогичными STA.

Для расширения зоны покрытия BSS, создаваемой одной точкой доступа, можно устанавливать антенны с более высоким коэффициентом усиления. При установке новой антенны требуется также провести исследование дальности уверенного приема сигнала для фиксации нового значения на карте зон покрытия корпоративной беспроводной сети.

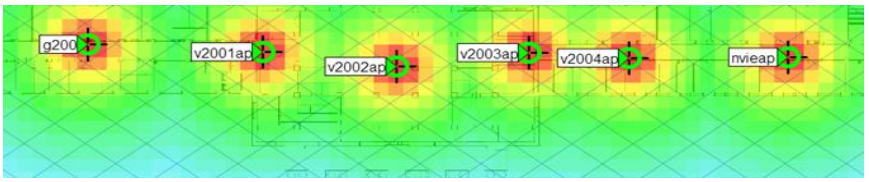


Рис. 7: Схема расположения точек доступа в корпоративной беспроводной сети НИУ МЭИ

Проведенные исследования позволили разработать схему установки точек доступа, обеспечивающую уверенный прием сигнала стандарта 802.11 в НИУ МЭИ. На рис. 6 приведен пример расстановки точек доступа с зонами покрытия в части одного из корпусов НИУ МЭИ.

ЛИТЕРАТУРА

1. Основы построения локальных сетей стандарта 802.11/ Рошан П., Лиэри Д. Пер. с англ. - М.: Изд. дом кВильямсњ. 2004. - 304 с
2. Wireless LAN Medium Access Control IEEE Standard for Information technology - Telecommunications and information exchange between systems Local and metropolitan area networks part 11: Specific requirements (MAC) and Physical Layer (PHY) Specifications / The Institute of Electrical and Electronics Engineers, Inc. - USA; New York: IEEE, 2012
3. Столлингс В. Беспроводные линии связи и сети.: Пер. с англ. - М.: Издательский домкВильямсњ, 2003. - 640 с.
4. Григорьев В.А., Лагутенко О.И., Распаев Ю.А. Сети и системы радиодоступа. - М.:Эко-Трендз, 2005. - 384 с.
5. Fluke Corporation. Спецификации: AirCheck Wi-Fi Tester. [Электронный ресурс]: Fluke Networks/Fluke Corporation. - Электрон. дан. - [2014]. - Режим доступа: http://ru.flukenetworks.com/content/datasheet-aircheck-wi-fi-tester

ABOUT CANONICAL REPRESENTATION OF THE BELLMAN FUNCTION IN AN OPTIMAL CONTROL AND MONITORING SYSTEMS FOR DYNAMIC NETWORKS

Yu. V. Solodyannikov
[1] SJC "Samara-Dialog Samara, Russia

Abstract

For dynamic network queues are posing problems of optimal control of complete and incomplete data, consider some types of functional, stated limitations on the state and management. Recorded Bellman equation for Markov network management tasks for complete data. The main result is to obtain a canonical representation of the solution of the Bellman equation for total Markov homogeneous Markov network management strategy. Examples of design problems of optimal control and optimal networking of information structures for some of modern telecommunications systems.

О КАНОНИЧЕСКОМ ПРЕДСТАВЛЕНИИ ФУНКЦИИ БЕЛЛМАНА В ЗАДАЧАХ ОПТИМАЛЬНОГО УПРАВЛЕНИЯ И НАБЛЮДЕНИЯ ДЛЯ ДИНАМИЧЕСКИХ СЕТЕЙ

Ю. В. Солодянников[1]
[1] ЗАО "Самара-Диалог", Самара, Россия,
solo-dialog@mail.ru

Аннотация

Для динамической сети очередей даются постановки задач оптимального управления по полным и неполным данным, рассмотрены некоторые типы функционалов, сформулированы ограничения на состояния и управления. Записано уравнения Беллмана для задачи управления марковской сетью по полным данным. Основным результатом является получение канонического представления решения уравнения Беллмана для общей марковской сети с однородной марковской стратегией управления. Приводятся примеры решения задач синтеза оптимальных управлений и оптимальных сетевых информационных структур для некоторых современных систем телекоммуникаций.

1. Введение

В работе содержится обзор некоторых математических методов синтеза оптимальных управлений в марковских системах (СМО) и сетях (СеМО) массового обслуживания в условиях полных и неполных данных. Основные результаты были полученны ранее и изложены в [1, 2].

Задача ставится и решается как построение совместного управления сетевыми состоянием и наблюдениями, т. е. путем ответа на вопросы: что, где, когда и как измерять на сети при динамической маршрутизации.

2. Задача оптимального управления по полным данным на марковской сети

Вероятностная модель случайных процессов в сети задается на вероятностном пространстве $\langle \Omega, \mathfrak{F}, \mathbf{P} \rangle$ с потоком σ-алгебр $\mathfrak{F} = \{\mathfrak{F}_t\}_{t \geqslant 0}$.

Входные потоки требований описываются последовательностью $\{\tau_{n_{ij}}^{ij}\}$ моментов прихода на узел i с узла j. Точечные процессы моментов прихода считаем независимыми в совокупности пуассоновскими потоками. Пусть λ_i^j - интенсивность пуассоновского потока требований, назначенных узлу j.

Каждое требование, находящееся в сети в данный момент времени t, характеризуется парой $(i, j) : i \in I, j \in I$, i - узел, на котором находится требование в данный момент, j - узел, которому требование предназначено. Требование, передающееся по каналу (i, k), до момента конца передачи считается находящимся на узле i.

X_{it}^j - случайная величина длины очереди (в числе требований) на узле i с назначением в узел j в момент времени t.

Кроме пары (i, j) для характеризации типа очереди используется индекс предыстории $\overline{p} = (p_1, \ldots, p_L)$, где p_k - номера узлов, которые проходило требование, причем $p_L = i$. Уровень запоминания предыстории L может быть различным. Если действует ограничение $L = 1$(простая предыстория), то запоминается только узел, на котором требование находится. Если $L = 2$, то запоминается узел i (на котором находится требование) и узел, с которого данное требование пришло. Если $L \leq \infty$, то запоминается вся предыстория до уровня L.

Предполагается, что траектории случайных процессов $X_{\overline{p}t}^j$ непрерывны справа и имеют конечные пределы слева (класс D). Сделанные предположения относительно $X_{\overline{p}t}^j$ означают,что $X_{\overline{p}t}^j$ - неотрицательные скачкообразные процессы с траекториями класса D.

Обслуживание в каналах описывается в терминах точечных процессов [3]. Время обслуживания требований в канале (i, j) является случайной величиной, распределенной экспоненциально со средним значением $1/\mu_{ij}$. Времена обслуживания одного требования в различных каналах независимы. Величина μ_{ij} является характеристикой канала и называется его пропускной способностью.

Задача оптимального управления ставится в классической форме. Имеется функционал вида

$$\Phi(u) = \mathbf{E}^u \left\{ c_0(T, X_T) + \int_0^T c(t, X_t, u_t)dt \right\} , \qquad (1)$$

где $\mathbf{E}^u$ - символ математического ожидания, $c_0(\cdot)$ и $c(\cdot)$ - непрерывные неотрицательные функции своих аргументов. Функция $c(\cdot)$ называется функцией цены, функционал (1) - функционалом стоимости, а время T — горизонтом оптимального управления.

Задача оптимального управления состоит в нахождении управляющего процесса u_t, реализации которого удовлетворяют некоторой системе ограничений $u_t \in U$ и который доставляет минимум (или максимум) функционалу (1).

3. Уравнение Беллмана

Функция Беллмана $V(t, x)$, называемая иногда также функцией цены, определяется как

$$V(t, x) = \min_{\{u_s \in U\}_{t \le s < T}} \mathbf{E}^u \left\{ c_0(T, X_T) + \int_t^T c(s, X_s, u_s)ds \mid X_{t-} = x \right\} . \qquad (2)$$

Вывод уравнения Беллмана для марковской СеМО имеется в [4], а само это уравнение в случае простой предыстории имеет вид:

$$-\frac{\partial V}{\partial t}(t, x) = \min_{u \in U}\{c(t, x, u) + \sum_{i \in I} \sum_{j \in I} u_{0i}^j \lambda_i^j [V(t, x + e_{(i)}^j) - V(t, x)] 1(Q_i < N) +$$

$$+ \sum_{k \in I} \sum_{i \in I, i \ne k} \sum_{j \in D^-(i), j \ne i} u_{ij}^k [V(t, x + e_j^k - e_i^k) - V(t, x)] 1((Q_j < N_j) \wedge (x_i^k > 0)) +$$

$$+ \sum_{k \in I} u_{k0}^k [V(t, x - e_k^k) - V(t, x)] 1(x_k^k > 0)\}. \qquad (3)$$

В уравнении (3) $e_{\overline{p}}^k$ - совокупность величин, индексированных соответствующим множеством индексов, причем $e_{\overline{p}}^k = 1$ для индексов $\overline{p}, k$ и $e_{\overline{p}}^k = 0$ для остальных.

Начальные условия уравнений для (3):

$$V(T, x) \equiv c_0(T, x) . \qquad (4)$$

4. О каноническом представлении функции цены

Идея о существовании некого канонического представления для функции цены, являющейся решением задачи (3)-(4), впервые изложена в работах автора [3] и [4]. Существование такого представления, во-первых, ограничивает поиск решения уравнения Беллмана некоторым узким классом функций, во-вторых, позволяет делать выводы о свойствах (в частности, асимптотических) функции цены. Во многих случаях знание канонического представления позволяет получить функцию цены в явном виде, не прибегая к непосредственному решению уравнения Беллмана.

Сформулируем результат для модели с простой предысторией и однородной стратегией управления в виде следующего утверждения.

Теорема 1. *Пусть $\hat{u}_t = \hat{u}(X_t)$ — оптимальная однородная марковская стратегия, отвечающая минимуму функционала накопленной задержки*

$$\Phi(u) = \mathbf{E}^u \int_0^T \sum_{i\in I} \sum_{j\in I} X_{it}^j dt$$

в задаче оптимального управления на марковской сети с простой предысторией. Тогда функция Беллмана, являющаяся решением задачи Коши (3)-(4), имеет представление

$$\begin{aligned} V(t,x) = \tau \sum_{i\in I}\sum_{j\in I} x_i^j + \sum_{i\in I}\sum_{j\in I} \lambda_i^j \hat{u}_{0i}^j(x)[\tau \circledast p_i^{N_i}(\tau,x)] - \\ - \sum_{j\in I} \hat{u}_{j0}^j(x)[\tau \circledast p_j^{+}(\tau,x)] \ , \end{aligned} \tag{5}$$

где $p_i^{N_i}(\eta,x) = \mathbf{P}\{\sum\limits_{j\in I} x_{i\eta}^j < N_i \mid X_0 = x,\ u_\eta = \hat{u}(x_\eta)\}$, $p_j^{+}(\eta,x) = \mathbf{P}\{x_{j\eta}^j > 0 \mid X_0 = x,\ u_\eta = \hat{u}(x_\eta)\}$, $\tau = T - t$.

Рассуждения, позволяющие получить каноническое представление функции цены для марковской сети и функционала накопленной задержки, аналогичны рассуждениям при выводе формулы Литтла для произвольной СМО [5]. Доказательство теоремы 1 имеется в [1].

Иногда представление, аналогичное (5), можно получить и для случаев, когда оптимальная марковская стратегия не является однородной. Как правило, упрощения в таких случаях определяются ограничениями, накладываемыми на управления в соответствии со спецификой задачи.

5. Жидкостная аналогия

Рассмотрим жидкостную модель СМО с управляемым выходным потоком. Пусть имеем воронку бесконечной емкости с регулятором, который управляет потоком, вытекающим из воронки. В воронку заливается

жидкость с постоянной скоростью $\frac{da(t)}{dt} = \lambda > 0$, а вытекает со скоростью $\frac{db(t)}{dt} = u(t)$. Будем предполагать, что скорость истечения ограничена интервалом $0 \leqslant u(t) \leqslant \mu$, где $\mu > 0$. Ограничимся рассмотрением "эргодического" случая с $\lambda/\mu = \rho < 1$.

Пусть в начальный момент времени $t = 0$ в воронке находился начальный объем жидкости $x_0 \geqslant 0$. Общее количество $b(t)$ вытекшей за время t жидкости не может превысить суммы поступившего количества $a(t)$ и начального количества x_0. Общее количество жидкости, содержащейся в воронке в момент t, определяется дифференциальным уравнением

$$\frac{dx(t)}{dt} = \lambda - u(t)$$

с начальным условием $x(0) = x_0$. Очевидно, $x(t) \geqslant 0$, и поскольку скорость истечения жидкости из воронки при $x(t) = 0$ не может превышать скорости притока λ, то управляемая скорость истечения должна подчиняться системе ограничений

$$u(t) \in U(x, \lambda, \mu) = \begin{cases} [0, \mu], & x(t) > 0 \ , \\ [0, \lambda], & x(t) = 0 \ . \end{cases}$$

Поставим задачу оптимального управления, состоящую в отыскании оптимальной стратегии управления $\tilde{u}(t, x(t))$, подчиняющейся указанной системе ограничений и доставляющей минимум функционалу

$$\tilde{J}_0 = \int_0^T x(t)dt \ .$$

Функция Беллмана, определяемая как

$$\tilde{V}(t, x) = \min_{u \in U(x,\lambda,\mu)} \{ \int_t^T x(s)ds \mid x(t) = x \} \ , \tag{6}$$

является решением уравнения Беллмана

$$-\frac{\partial \tilde{V}}{\partial t}(t, x) = \min_{u \in U(x,\lambda,\mu)} \{x + \lambda \frac{\partial \tilde{V}}{\partial x}(t, x) - u \frac{\partial \tilde{V}}{\partial x}(t, x)\}$$

с начальным условием

$$\tilde{V}(T, x) \equiv 0 \ .$$

Теорема 2. *При $\rho < 1$ оптимальное управление в задаче не зависит от времени и имеет представление*

$$\tilde{u}(t, x) = \lambda 1(x = 0) + \mu 1(x > 0) = \begin{cases} \lambda \ , & x = 0 \ , \\ \mu \ , & x > 0, \end{cases}$$

а функция цены имеет представление:

$$\tilde{V}(t,x) = x\tau + \frac{\lambda\tau^2}{2} - [\lambda 1(x=0) + \mu 1(x>0)]\tau \circledast \left[1 - \mathfrak{u}_{-1}\left(\tau - \frac{x}{\mu-\lambda}\right)\right] \ . \quad (7)$$

Заметим, что формула (7) является "жидкостным" вариантом канонического представления (5) и легко получается из определения функции цены (6) и расчета количеств поступившей и вытекшей жидкости. Такой подход с использованием канонического представления функции цены может быть применен и к другим самым различным объектам сетевой структуры живой и неживой природы, например управление водохранилищами, управление кровообращением и др.

6. Пример синтеза оптимального управления роутером

Рассмотрим стационарный пуассоновский поток требований интенсивности $\lambda > 0$. Требования обслуживаются пулом из n приборов, перед каждым из которых допускается неограниченная очередь. В момент поступления каждое из требований некоторым устройством с вероятностью $u_i \in [0,1]$, $\sum\limits_{i=1}^{n} u_i = 1$ направляется в очередь i-го прибора. Время обслуживания на i - м приборе экспоненциально распределенно с интенсивностью обслуживания $\mu_i > 0$.

Обозначим $X_t = (X_{1t}, \ldots, X_{nt})$ n-компонентный скачкообразный процесс длин очередей с множеством состояний $x = (x_1, \ldots, x_n)$, $x \in [\mathbb{Z}_0^{\infty}]^n$. При сделанных предположениях каждый из приборов роутера в отдельности представляет собой СМО с ограничением нагрузки, а процессы X_{it} есть управляемые марковские процессы.

Поставим задачу нахождения оптимальной марковской стратегии управления $\hat{u}(t,x) = \hat{u}(t,X_t) \mid X_t = x$, доставляющей минимум функционалу накопленной задержки.

$$J(u) = \mathbf{E}^u \int_0^T \sum_{i=1}^{n} X_{it} dt \ . \quad (8)$$

Уравнение Беллмана (3) для данного случая имеет вид

$$\begin{aligned} -\frac{\partial V}{\partial t}(t,x) = \min_{u\in U} \Bigg\{ \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} \lambda u_i [V(t,x+e_i) - V(t,x)] + \\ + \sum_{i=1}^{n} \mu_i [V(t,x-e_i) - V(t,x)] 1(x_i > 0) \Bigg\} \end{aligned} \quad (9)$$

с начальным условием $V(T,x) \equiv 0$ и множеством ограничений

$$U = \left\{ u_i \in [0,1], \sum_{i=1}^{n} u_i = 1 \right\} \ . \quad (10)$$

Выражение для оптимальной стратегии управления через функцию цены находится исходя из линейности по u_i выражения под знаком минимума в (9) и линейности системы ограничений (10). Обозначим через $J_0(t,x) = \{j : j = \arg\min\limits_{1\le i\le n} \Delta_i^+(t,x)\}$ заведомо непустое множество индексов i, для которых достигается минимальное значение разностных производных функции цены $\Delta_i^+(\tau,x) = V(t,x+e_i) - V(t,x)$ в точке (τ,x), $\tau = T-t$, а через $j_0(t,x)$ — произвольно выбранный элемент множества $J_0(t,x)$. Оптимальное управление записывается в виде

$$\hat{u}_i(t,x) = 1(i = j_0(t,x)) \ .$$

При получении явного выражения для функции цены используем ее каноническое представление определенное теоремой 1. Поскольку в векторе оптимального управления согласно (10) только один компонент (с индексом $j_0(x)$) равен 1, а остальные равны 0, то выражение для функции цены при замене переменных $\tau = T-t$, $z = \theta - t$, $y = s - t$ примет вид

$$V(t,x) = \tau \sum_{j=1}^{n} x_j + \lambda \frac{\tau^2}{2} - \sum_{j=1}^{n} \mu_j [\tau \circledast p_j^+(\tau,x)] \ ,$$

где $p_j^+(\eta,x) = \mathbf{P}\{x_{j\eta} > 0 \mid x_0 = x,\ u_\eta = \hat{u}(\eta, x_\eta)\}$ — зависящие от оптимального управления вероятности ненулевых длин очередей в момент времени η при условии, что в нулевой момент времени вектор длин очередей был равен x.

Для получения замкнутого выражения для функции цены осталось записать формулы для $p_j^+(\tau,x)$. А именно, для $j = j_0(x)$ будем иметь

$$p_j^+(\tau,x) = 1 - e^{-(\lambda+\mu_j)\tau}[\rho_j^{-x_j/2} I_{x_j}(2\tau\sqrt{\lambda\mu_j}) +$$

$$+\rho_j^{-(x_j+1)/2} I_{x_j+1}(2\tau\sqrt{\lambda\mu_j}) + (1-\rho_j) \sum_{i=x_j+2}^{\infty} \rho_j^{-i/2} I_i(2\tau\sqrt{\lambda\mu_j})] \ ,$$

где $\rho_j = \lambda/\mu_j$. Для $j \ne j_0(x)$ будем иметь

$$p_j^+(\tau,x) = 1 - \sum_{i=x_j}^{\infty} \frac{(\mu_j\tau)^i}{i!} e^{-\mu_j\tau} \ .$$

Переходя к этапу синтеза оптимального управления, запишем выражения для разностных производных функции цены $\Delta_j^+(\tau,x)$. Обозначим через $F_k(s,a,b)$ и $f_k(s,a,b)$ соответственно функцию распределения и плотность распределения обобщенного интервала занятости в СМО $M/M/1$ с параметром входного потока a и параметром времени обслуживания b. Тогда для $j = j_0(x)$ будем иметь

$$\Delta_j^+(\tau,x) = \int_0^\tau (1 - F_{x_j}(s,\lambda,\mu_j))ds \ ,$$

а для $j \neq j_0(x)$

$$\Delta_j^+(\tau, x) = \int_0^\tau (1 - F_{x_j}(s, 0, \mu_j))ds \ .$$

Плотность распределения $f_k(s, a, b)$ для $a > 0$ известна:

$$f_k(s, a, b) = e^{-(a+b)s}(k+1)s^{-1}(a/b)^{-(k+1)/2} I_{k+1}(2s\sqrt{ab}) \ ,$$

а для $a = 0$ она очевидно выражается формулой

$$f_k(s, 0, b) = \frac{(bs)^k}{k!} b e^{-bs} \ .$$

Оптимальная стратегия управления состоит в направлении пришедшего в момент времени t из внешней среды требования в очередь к прибору, для которого минимально значение разностной производной $\Delta_j^+(\tau, x)$ при условии, что вектор длин очередей в этот момент точно известен и равен x. Эта оптимальная стратегия, являясь марковской, вообще говоря, не является однородной. Однако, исходя из асимптотических свойств $\Delta_j^+(\tau, x)$, при больших τ можно показать, что оптимальные управления зависят от времени лишь вблизи горизонта оптимального управления, т. е. для значений t, близких к T.

7. Заключение

В работе рассмотрены актуальные вопросы совместного оптимального управления и наблюдения для управляемых объектов сетевой структуры. При этом учитываются особенности скачкообразных процессов как в измерениях, так и в управлении. Такая общая постановка задачи управления и наблюдения применительно к СМО и СеМО позволяет с единых позиций формулировать и решать соответствующие оптимизационные задачи.

ЛИТЕРАТУРА

1. Солодянников Ю.В. Управление и наблюдение для динамических сетей массового обслуживания. I // АиТ. 2014. Т. 75. No.3 С. 14-45.
2. Солодянников Ю.В. Управление и наблюдение для динамических сетей массового обслуживания. II // АиТ. 2014. Т. 75. No.5 С. 91-114.
3. Солодянников Ю.В. Некоторые вопросы сетеметрии//Теоретические проблемы вычислительных сетей. Научный Совет по комплексной проблеме "Кибернетика" АН СССР, Куйбыш. гос. ун-т. Куйбышев, 1986. С. 74–102.
4. Солодянников Ю.В. О статистике систем и сетей массового обслуживания//Проблемы устойчивости стохастических моделей. Тр. X Всесоюз. семинара. Куйбыш. гос. ун-т. Куйбышев, 1987. С. 101–116.
5. Клейнрок Л. Теория массового обслуживания. М.: Машиностроение, 1979.

A METHOD FOR ENHANCING THE SECURITY AND DATA STORAGE DURING INFORMATION TRANSMISSION IN TELEMETRY SYSTEMS

I.Ivanov[1], S.Vetova[2]
[1] High School of Telecommunications and Post, Sofia, Bulgaria
[2] Bulgarian Academy of Sciences, Sofia, Bulgaria
ivanivanov@hctp.acad.bg , vetova.bas@gmail.com

Abstract

The task of the proposed method in the following paper is to increase security and data storage in the information transmission in telemetry systems based on symmetric block cryptographic algorithm by increasing the number of switches and the number of internal operations (in accordance with Feistel scheme) i.e. increasing the resistance of cryptographic encrypted data against cryptographic attacks.

Keywords: cryptography, Feistel scheme, cryptographic algorithm, data storage, telemetry systems, data protection, cryptographic resistance, cryptographic attacks.

1. Introduction

Individual cryptographic protection is used in the cases when it is needed to provide personal protection of the information exchange for a separate user (network subscriber). Then, he/she is given an individual encryption tool and is able to implement secure information exchange with other subscribers of the communication network who has the same tool and corresponding encryption key interaction.

Cryptographic protection covers the entire route between the two subscriber loop and the chosen route through the network. This means that encryption is performed "end-to-end" and there is no unprotected section of the connection between two callers, which is essential advantage of this organization approach.

Basic algorithms, used for such protection, are block cryptographic information algorithms [1÷6]. They are symmetric (using the same key for encryption and decryption) and work on blocks of bits with fixed length. From these blocks, they generate encrypted information with equal, longer or shorter length. These algorithms use a secret key, which has a fixed length too.

The main properties of block algorithms are on one hand confusion - the ability of an algorithm to make the connection between the key and ciphertext as complex and mixed as possible and on the other hand diffusion - the ability of an algorithm to make the link between text and ciphertext as complex and scattered as possible too. These properties are achieved by using the scheme used for the encryption algorithm.

2. Method

The task of the proposed method is to increase security and data storage in the information transmission in telemetry systems based on symmetric block cryptographic algorithm by increasing the number of switches and the number of internal operations (in accordance with Feistel scheme). That is, increasing the

resistance of cryptographic encrypted data against cryptographic attacks without significant reduction of the algorithm rate.

In accordance with the method, these problems can be solved through a series of operations at a certain information flow formatting in the following way:

- The actual use of 256-bit master key is entered;
- Two 48-bit and two 32-bit sub-keys at each cycle are used;
- The total number of sub-keys are 64;
- An additional block of the encryption function in each scheme cycle is added;
- An additional XOR adder in each cycle is added.

2.1. Method description

The data encryption process is performed in the following manner. First the manifest information (encryption information) is divided into blocks of 64-bits each (Figure 1).

In the initial phase and final transposition 64-bit blocks are submitted and blocks of the same size are generated as the displacement is a process of changing the bits positions without changing the values, which is 1-1 reversal. Reversal is based on Table 1 and Table 2. The obtained bit sequence is divided into two sequences left L(0) and right R(0), each of which containing 32 bits. Then, the encryption process performs using the function F(.). In the proposed method this function is implemented on both parts. The output 32-bit sequences of both F(.) functions are applied to the first inputs of two adders XOR. The second inputs of these adders are fed first with two 32-bit sub-keys. The 32-bit output sequences of the two XOR adders, change places. That is left becomes right and vice versa, which in its turn forms the input sequence for the next cycle. These procedures are carried out totally sixteen times, on the basis of the internal cycles.

58	50	42	34	26	18	10	2
60	52	44	36	28	20	12	4
62	54	46	38	30	22	14	6
64	56	48	40	32	24	16	8
57	49	41	33	25	17	9	1
59	51	43	35	27	19	11	3
61	53	45	37	29	21	13	5
63	55	47	39	31	23	15	7

Table 1. Primary transposition

40	8	48	16	56	24	64	32
39	7	47	15	55	23	63	31
38	6	46	14	54	22	62	30
37	5	45	13	53	21	61	29
36	4	44	12	52	20	60	28
35	3	43	11	51	19	59	27
34	2	42	10	50	18	58	26
33	1	41	9	49	17	57	25

Table 2. Final transposicion

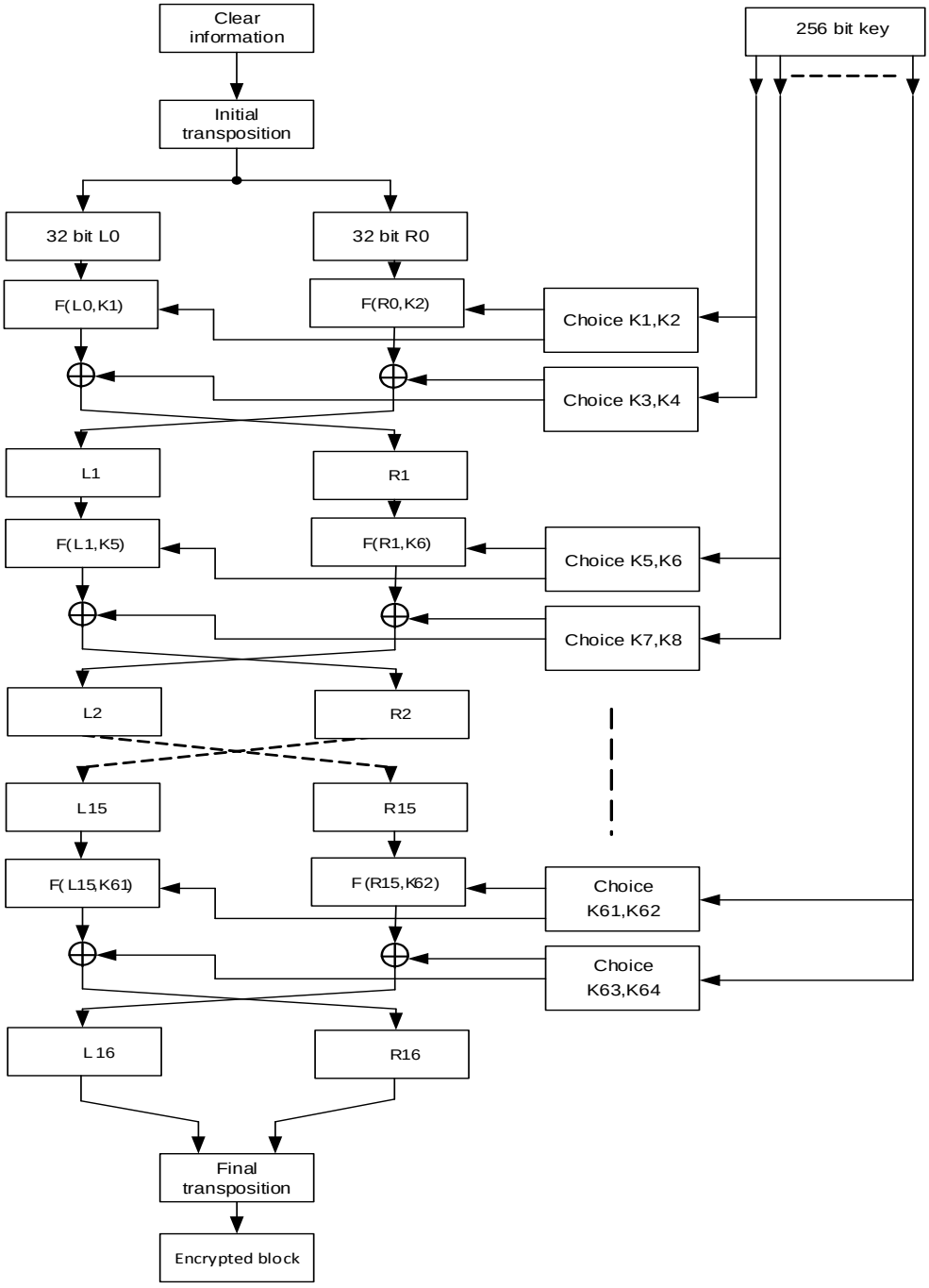


Fig. 1. Overall block diagram of the proposed encryption method

2.2. Realization of the function F (.)

The realization of the function F(.) (Figure 2) is performed in the following order.

Initially, the right side of R_{i-1} of the data block is expanded by 32 to 48 bits by repeating the bits with number 1, 4, 5, 8, 9, 12, ..., 24, 28, 32 (the border bits of the groups containing 4 consecutive bits). The new bits of the extension are joined cyclically to the eight formed adjacent structures by 4 bits according to the scheme:

$$r_{32}\, r_1\, r_2\, r_3\, r_4\, r_5,\ r_4\, r_5\, r_6\, r_7\, r_8\, r_9,\ \ldots\ldots\ldots,\ r_{28}\, r_{29}\, r_{30}\, r_{31}\, r_{32}\, r_1 \qquad (1)$$

Then, the formed part R_{i-1} with the length of 48 bits is multiplied by mod 2 48-bit key K_i element by element and divided into 8 successive structures with the length of 6 bits. These structures are submitted to the S - boxes, performing nonlinear procedure - a choice of 4 outputs (bits) from 6 inputs (bits). Then, transposition P implements in a specific pattern and produces the final result of the cryptographic processing F(.) with the length of 32 bits.

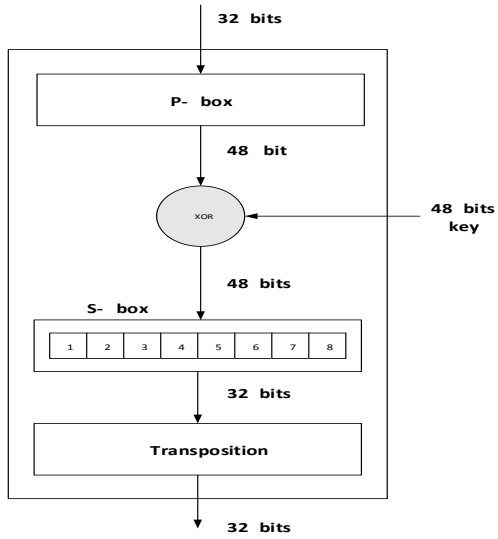


Fig.2. F – function

2.3. Keys generator.

The process of generating the 64 keys $\{K_j\}$ with the length of 48 bits and 32 bits runs according to the standard algorithm, shown in the right part of Fig.1. Key $\{K_j\}$ is entered in the cryptographic scheme with the length of 256 bits.

The keys involved in each cycle are generated from the basic one in the following manner. Consistently required number of bits are taken for the sub-keys, as follows: the first 48 bits of the main switch are taken as key K_1 and fed at the entrance of the function F (L_0, K_1), the next 48 bits are taken as key K_2 at the input of F (R_0, K_2), the next 32 bits are taken as K_3 and fed the input of the first adder XOR of the first cycle, the next 32 bits are taken as key K_4 and fed to the input of the second adder XOR of the first cycle, then the same procedure implements to generate the quadruple keys for the rest cycles up to the sixteenth one. After reaching the last byte of the basic key, shift of 25 bits is performed linearly to the left in the 256-bit sequence as many times

as necessary to obtain the required number of sub-keys.

Based on the proposed scheme in Figure 1, the cryptographic resistance of the encrypted data is significantly enhanced using 256-bit master key, implementation of function F(.) on the left and on the right of the data and using sixty four sub-keys for the sixteen internal cycles.

The algorithm rate is comparable with the used nowadays triple DES, since the encryption is performed only once, while triple DES encryption is performed three times, both according to one and the same scheme.

3. Conclusion

The method of individual encryption of transmitted information, which is used in telemetry systems with special purpose, is especially useful for protection of any kind of data exchanging between certain users. Protected data can be exchanged either in real time or sent after any period of time.

Based on the proposed method, the cryptographic resistance of the encrypted data is significantly enhanced by the use of 256-bit master key, implementation of the encryption function on the left and on the right of the data and the use of sixty four sub-keys for the sixteen internal cycles.

By using cryptographic protection in telemetry systems with special purpose, data can be sent in real time and reliably stored in an appropriate buffer (memory) and delivered only to the recipient as intended: insurance, road assistance, fire-brigade, ambulance and others.

REFERENCES

[1] Stallings W. Cryptography and Network Security: Principles and Practice (6th Edition), Hardcover, 2013.

[2] Schneier B., Appliet Cryptography Protocols, Algorithms, and Source Code in C, Wiley, 2013.

[3] Ivanov I. Laboratory experiments on security and protection of information and administration and protection of communication and computer networks. HS CTP, Sofia 2013.

[4] Ferguson, N., Schneier, B., Kohno, T. Cryptography Engineering: Design Principles and Practical Applications, Wiley, 2010.

[5] Sokolov V., Shangin F. Information protection razpredelenyh corporate networks and systems. DMK Press, Moskva, 2002.

[6] Katz J., Lindell Y. Introduction to Modern Cryptography, Second Edition (Chapman & Hall/CRC Cryptography and Network Security Series), CRC Press, 2014.

EFFECTIVE BANDWIDTH ESTIMATION IN THE REGENERATIVE NETWORKS

K. Kalinina[1], *E. Morozov*[1,2]

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

[2] Petrozavodsk State University, Petrozavodsk, Russia

emorozov@karelia.ru, smesharikk@mail.ru

Abstract

In this work, we study the accuracy of the estimation of the effective bandwidth (EB) of the nodes in a computer network. The basic process describing the network is assumed to be (quasi) regenerative. This approach is based on the fact that the regenerative input stays regenerative crossing acyclic network with no saturated nodes. Thus each node of such a network can be analyzed as an isolated system. We study the property of a regeneration-based estimator of the EB and show that it ensures a predefined QoS requirement, however with an overestimation. We discuss the problem when this overestimation is acceptable for highly reliable networks. Moreover, we touch upon the problem of the accelerated estimation because regeneration-based estimator uses some facts from theory of large deviations.

Keywords: regeneration, effective bandwidth, fast simulation

1. Introduction

Consider an one-server queueing system with loses with constant service rate C, finite buffer b and stationary workload process W (i. e. the unfinished work in steady state mode of the system). The capacity C can be chosen according to some quality of service (QoS) parameters (such as stationary waiting time W, stationary loss probability, stationary overflow probability, etc.). We consider the overflow probability P_b of the stationary workload as a QoS parameter satisfying a requirement:

$$P_b := \mathbb{P}(W > b) \leq \Gamma, \tag{1}$$

where bound Γ is predefined. If the service rate C satisfies condition (1), then C is called the *effective bandwidth* (EB).

We consider a slotted time scale, and let v_i be the workload arrived in the system in time interval $[i, i-1)$, $i = 0, 1, \ldots$ Assume the following limit exists

$$\Lambda_n(\theta) = \frac{1}{n} \ln e^{\theta \sum_{i=0}^{n-1} v_i} \to \Lambda(\theta), \ n \to \infty, \tag{2}$$

where free parameter $\theta > 0$. This limit $\Lambda(\theta)$ is called (scaling) *limit logarithmic moment generating function* of the input process [1]. Then EB C can be found

as follows (for details, see [1, 2, 3]):

$$C := \frac{\Lambda(\theta^*)}{\theta^*}, \quad \theta^* = -\ln \Gamma / b > 0. \tag{3}$$

In this paper, we describe regenerative estimator of the effective bandwidth C of a node being a component of a computer network. It is assumed that the basic process describing the network is regenerative or quasi-regenerative [3]. Also we discuss properties of the corresponding estimator of C.

2. Regenerative estimator

According to (3), to calculate the effective bandwidth C, we need to find function $\Lambda(\theta)$. It is possible to get the analytical expression of $\Lambda(\theta)$ only in a few simplest cases when random variables (r.v.) v_i are i.i.d. Otherwise it is necessary to use the strongly consistent standard sample mean estimator,

$$\Lambda_k(\theta) := \ln \frac{1}{k} \sum_{i=0}^{k-1} e^{\theta v_i} \to \Lambda(\theta), \ k \to \infty. \tag{4}$$

As it has been discussed earlier (see [2, 3]), the preferable way to estimate $\Lambda(\theta)$ is to apply the regeneration. Assume the input sequence $\{v_i\}$ is regenerative with regeneration points $\{\beta_n\}$. Then

$$X_k := \sum_{i=\beta_k}^{\beta_{k+1}-1} v_i, \ k \geq 0, \ \beta_0 := 0,$$

are i.i.d. blocks (with generic element X). It allows to construct the following regenerative estimator of function $\Lambda(\theta)$ [4]:

$$\Lambda_k(\theta^*) := \frac{k}{\beta_k} \ln \frac{1}{k} \sum_{i=1}^{k} e^{\theta^* X_i} \to \frac{\ln \mathbb{E} e^{\theta^* X}}{\mathbb{E}\beta} =: \Lambda(\theta^*), k \to \infty, \tag{5}$$

with the following moment assumptions:

1) $\mathbb{E}\beta^2 < \infty$;
2) there exists $\theta_0 > 0$ such that $\mathbb{E} e^{\theta X} < \infty, \ \theta \in (0, \theta_0)$.

Then the upper bound of the EB C can be obtained as the limit

$$C_k(\theta^*) := \frac{\Lambda_k(\theta^*)}{\theta^*} \to \frac{\ln \mathbb{E} e^{\theta^* X}}{\theta^* \, \mathbb{E}\beta}. \tag{6}$$

Table 1 shows the simulation results of regeneration-based EB of a node of a tandem network with $M = 5$ nodes and the input rate λ. The results demonstrate that the regenerative technique gives some overestimation Δ(about 15-18%) of the required QoS guarantee (1). The overestimation Δ

λ	$\hat{D}$	$Var\hat{D}$	$\hat{C}$	$\hat{\Gamma}$	Δ	$\hat{\Gamma}_\Delta$
0.3	3.93	9.17	1.80	$3.26 \cdot 10^{-6}$	0.11	$8.90 \cdot 10^{-6}$
0.3	3.93	9.17	1.80	$3.26 \cdot 10^{-6}$	0.15	$2.71 \cdot 10^{-4}$
0.9	6.59	18.22	2.85	$3.36 \cdot 10^{-7}$	0.15	$3.79 \cdot 10^{-6}$
0.9	6.59	18.22	2.85	$3.36 \cdot 10^{-7}$	0.18	$2.94 \cdot 10^{-4}$

Table 1: Tandem network, M=5, $\Gamma = 10^{-5}$

is discussed in more details in [2]. Simulation results show that Δ increases together with regeneration period and its variance. Because we deal with very small probabilities, well-known Monte-Carlo simulation technique is extremely time-consuming to obtain the reliable estimates in acceptable time. It is necessary to speed up the simulation, and by this reason we discuss some acceleration methods of rare event simulation. In particular, we focus on variance reduction techniques such as Importance Sampling and Multilevel Splitting (see [1, 2, 5]).

3. Conclusion

In this paper, we discuss regeneration-based estimator for the EB, the overestimation that it gives and some ways of speeding up simulation to investigate possible dependence between overestimation Δ and length of regeneration cycles and its variance $Var\hat{D}$.

4. Acknowledgments

This work is supported by Russian Foundation for Basic research, projects 15–07–02341 A, 15–07–02354 A,15–07–02360 A and the Program of strategic development of Petrozavodsk State University.

REFERENCES

1. Lewis J. T., Russell R. An introduction to large deviation for teletraffic engineers. DIAS Technical Report DIAS-STP 97-16, 1997.
2. Morozov E., Kalinina K. On the effective bandwidth estimation in communication network // 29th European Conference on Modelling and Simulation ECMS 2015: Proceedings.
3. Borodina A., Kalinina K., Morozov E. On the accuracy of the effective bandwidth regenerative estimation, ICUMT, 2014, Saint-Petersburg, 652-656.
4. Dyudenko I., Morozov E., Pagano M. Regenerative estimator for effective bandwidth // Mathematical methods for analysis and optimization of information telecommunication networks: Proceedings of the International Conference. — Minsk: Belarusian State University, 2009. P. 58–60.
5. Miretskiy D.I., Scheinhardt W.R.W., Mandjes R.H. On efficiency of multilevel splitting // Communications in statistics — Simulation and Computation, 41: 890-904, 2012.

ESTIMATION OF MULTISERVER QUEUES BASED ON REGENERATIVE ENVELOPES

E. Morozov[1], *R. Nekrasova*[1], *I. Peshkova*[2]

[1] Institute of Applied mathematical research Karelian Research Centre RAS Center, Petrozavodsk, Russia

[2] Petrozavodsk State University, Petrozavodsk, Russia

emorozov@karelia.ru, ruslana.nekrasova@mail.ru, iaminova@psu.karelia.ru

Abstract

We consider an infinite buffer FCFS m-server queueing system $GI/G/m$. The regenerative simulation method can be used to estimate important QoS parameters of presented system, such as the mean stationary waiting time or mean queue size. In case classic regeneration is absent, we suggest a novel method called *regenerative envelopes*. The method is based on a coupling and monotonicity property of the workload process. We construct two auxiliary regenerative systems which give upper and lower bounds for the workload and queue size processes in the original non-regenerative system. Estimation based on simulation of a multiserver FCFS system $GI/G/m$ illustrates this approach.

Keywords: regenerative method, confidence estimation, infinite buffer systems, monotonicity

We consider an infinite buffer FCFS m-server queueing system $GI/G/m$ with the i.i.d. inter-arrival times $\{\tau_n := t_{n+1} - t_n,\, n \geq 0\}$ and i.i.d. service times $\{S_n \;\; n \geq 0\}$ with rates $\lambda := 1/\mathsf{E}\tau$ and $\mu := 1/\mathsf{E}S$, respectively. (We denote by τ and S the corresponding generic elements.) Let ν_n define the number of customers in the system before the nth arrival (at instant t_n^-). Under well-known conditions $\rho := \lambda/\mu < m$ and $\mathsf{P}(\tau > S) > 0$ classical regenerations of the system (and, in particular, of the process $\{\nu_n,\, n \geq 0\}$) exist. Regenerations of the process $\{\nu_n\}$ and some other processes related to the system are defined as follows:

$$\beta_{n+1} = \inf_k (k > \beta_n : \nu_k = 0), \qquad n \geq 0. \tag{1}$$

Note, that regeneration periods $\alpha_n = \beta_{n+1} - \beta_n$ are i.i.d. and define the generic element by α. If the regenerative process $\{\nu_n\}$ is positive recurrent ($\mathsf{E}\alpha < \infty$) and aperiodic ($\mathsf{P}(\alpha = 1) > 0$), then the following weak limit exists [1]

$$\nu_n \Rightarrow \nu, \qquad n \to \infty. \tag{2}$$

(The same holds for other regenerative processes describing the system.) The limit ν is an important QoS parameter. The regenerative method allows to construct confidence intervals for performance measures such like $\mathsf{E}\nu$ [1].

Under the opposite condition $\mathsf{P}(\tau > S) = 0$ the classical regenerations are excluded and the regenerative confidence estimation is not applicable. In this case we present another method based on the so-called *regenerative envelopes*. We use a coupling technique to construct an auxiliary infinite buffer system (denoted by Σ) in the following way. The system Σ has the same number of servers (m) and the same input flow as original non-regenerative queuing system. Define by $S_k^{(i)}$ the remaining service time in the ith server at instant t_k^-. For fixed integer $q \leq 0$ and real $0 \leq a_i \leq b_i < \infty$ we define the following arrival moment in the system Σ:

$$\gamma_1 = \inf_k \{k : \nu = q, S_k^{(i)} \in [a_i,\, b_i],\, i = 1, \ldots, m\}. \tag{3}$$

Our goal is the following: if $k = \gamma_1$, we replace all remaining service times $S_k^{(i)}$, $i = 1, \ldots, m$ by its upper bound b_i. Thus, we increase the corresponding service times in comparison with service times in the original system.

Define by d_n the departure instant of the nth arrival in system Σ. (Note that in general the order of the departures is not the same as order of the arrivals.) Consider the set $\mathcal{M}_n = \{k : t_k \leq t_n < d_k\}$ which defines the numbers of customers in Σ at instant t_n^+. Now we are able to present the following recursion

$$\gamma_{n+1} = \inf\{k > \gamma_n : \nu_k = q,\, S_k^{(i)} \in [a_i,\, b_i], \\ \mathcal{M}_{\gamma_n} \bigcap \mathcal{M}_k = \varnothing, \quad i = 1, \ldots, m\}, \;\; n \geq 1. \tag{4}$$

Note, that at each instant γ_n, $n \geq 2$, we replace all remaining service times by the corresponding upper bounds b_i, and then replace all service times of the customers waiting in the *queue* (if any) by the i.i.d. random variables $\{S_n^{(0)}\}$ distributed as the generic service time S. The system Σ regenerates at instants $\{\gamma_n,\, n \geq 1\}$ and, in particular, the regenerative confidence estimation is applicable. Moreover, we synchronize the original and auxiliary systems as follows. Define by $\{\tilde{\tau}_n\}$ and $\{\tilde{S}_n\}$ the arrival epochs and service times in Σ, respectively. Then, we obtain the following relations

$$\tau_n = \tilde{\tau}_n,\; S_n \leq \tilde{S}_n, \qquad n \geq 0. \tag{5}$$

By the well-known monotonicity property inequalities (5) allow to establish the corresponding relations between the mean queue size and the mean workload in both systems [2]. As a result, the auxiliary system Σ majorates the original non-regenerative system. Such approach allows to construct an upper bound for mail QoS parameters (mean queue siza and mean workload) based on regenerative method. In the similar way we construct a minorant regenerative system (at instants γ_n, $n \geq 1$ replacing the remaining service times to the corresponding lower bounds a_i). The minorant system provides lower bound of the required QoS parameters.

Acknowledgment

This work is supported by Russian Foundation for Basic research, projects No 15–07–02341, 15–07–02354,15–07–02360, and by the Program of strategic development of Petrozavodsk State University.

REFERENCES

1. Asmussen S. Applied Probability and Queues, (2nd ed.) New York: Springer-Verlag, 2003.
2. Morozov E. V. Coupling and stochastic monotonicity of queueing process Petrozavodsk: PetrSU, 2013.

COMPUTER SIMULATION OF THE THROUGHPUT OF CROSSBAR SWITCH WITH MODIFIED CHANG'S MODEL FOR LOAD TRAFFIC

T. Tashev[1], V. Monov[1], M. Marinov[2]
[1] Institute of information and communication technologies – B.A.S., Sofia, Bolgaria
[2] Technical University – Sofia, Sofia, Bolgaria

Abstract: *Achieving maximum throughput (THR) of crossbar switch node is obtained by calculating the non-conflict schedule to switch incoming packets. In this paper the results from a numerical simulation of the THR obtained using the grid-cluster of IICT-BAS (ww.grid.bas.bg) are presented. Our simulation employs MiMa-algorithm specified by apparatus of Generalized Nets. We utilize family of patterns of our own design for non-symmetric traffic (based on the Chang-model). The main result includes determining of the upper bound of the THR of the MiMa-algorithm as the maximal possible throughput value equal to 100%.*

Key words: *Computer and Communication Networks, Crossbar Switch Node, Throughput modeling, Generalized Nets, Algorithms.*

КОМПЬЮТЕРНАЯ СИМУЛЯЦИЯ ПРОПУСКНОЙ СПОСОБНОСТИ КОММУТАТОРА С МАТРИЧНЫМ ПЕРЕКЛЮЧАТЕЛЕМ ДЛЯ ВХОДЯЩЕГО ТРАФИКА ТИПА МОДИФИЦИРОВАННОЙ МОДЕЛИ ЧАНГ-А

Т. Ташев[1], В. Монов[1], М. Маринов[2]
[1] Институт информационных и коммуникационных технологий Болгарской Академии Наук, София, Болгария
Технический университет – София, София, Болгария

Резюме: Достижение максимальной пропускной способности (ПС) коммутационного узла с матричным переключателем получается путем вычисления бесконфликтного расписания для коммутации входящих пакетов. В данной работе представлены результаты численного моделирования ПС, полученые с использованием вычислительной сети IICT-BAS (ww.grid.bas.bg). Наше моделирование использует MiMa-алгоритм, специфицированный аппаратом Обобщенных сетей. Для несимметричного входящего трафика мы применили семейство шаблонов на основе Чанг-модели. Основной результат состоит в определения максимальной ПС при MiMa-алгоритме как стремящееся к значению равному 100%.

Ключевые слова: Компьютерные и коммуникационные сети, Пакетный коммутатор, Оценка производительности, Обобщенные сети, Алгоритмы.

1. Introduction

Crossbar switch node is a device which maximizes the speed of data transfer using parallel existing paths between the input and output lines in the commutation field of the node (as shown in Figure 1) [1]. This is obtained by means of a non-conflict commutation schedule calculated by the control block (Scheduler) of the switch node (as shown in Figure 2) [2]. From a mathematical point of view the calculation of such a schedule is NP-complete [3]. The existing classical solutions (PIM-algorithm [4], iSLIP-algorithm [5]) partly solved the problem. New more effective algorithms for schedule calculation are needed and they have to be checked for efficiency. As examples we can mention the CTC(N) [6] and RR/LQF [2] algorithms.

The efficiency of the switch performance is firstly evaluated by the throughput (THR) provided by the node for uniform and non-uniform load traffic [1, 2, 6]. The modelling of the THR should be performed for large area of the switch field (from 2x2 to at least 128x128 input/output lines). Such simulations and the necessary computations are typically carried out by using grid-computer structures. In our previous works we used this approach [7].

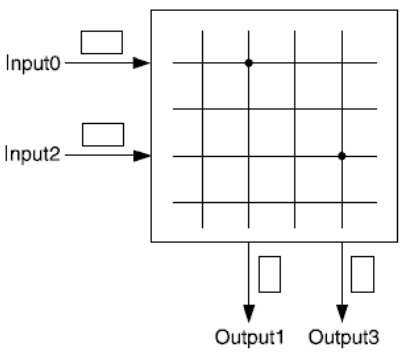


Fig.1 Crossbar switching field.

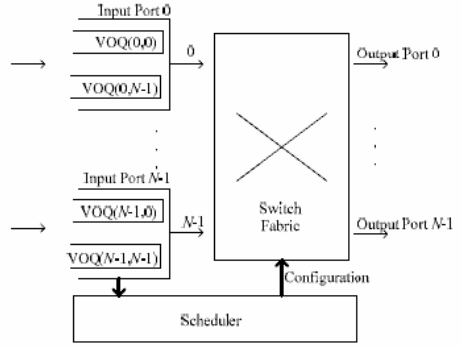


Fig.2 An input-queued switch with a scheduler.

In this paper we determine the THR of a crossbar switch (with input buffering and Virtual Output Queues – Figure 2) by means of a computer simulation up to 130x130 lines. We utilize family of patterns of our own design for non-symmetric traffic (based on the Chang-model [8]). Our simulation employs MiMa-algorithm (developed by the first author) specified by apparatus of Generalized Nets (GN). Generalized Nets [9] are a formal tools allowing a representation of connections in serial and parallel processes [10, 11].

The main results of the paper include determining of an upper bound of the throughput for MiMa-algorithm under the specified family of patterns. The bound of the throughput of the MiMa-algorithm approaches the maximal possible throughput value equal to 100%.

2. MiMa-algorithm for computing a conflict-free schedule

The requests for packet transmission through switching $n \times n$ line switch node is presented by an $n \times n$ matrix $\boldsymbol{T}$, named traffic matrix (n is integer). Matrix $\boldsymbol{T}$ describes unidirectional packet flow - from input lines to output lines. Every element t_{ij} ($t_{ij} \in \{0, 1, 2, \ldots\}$) of the traffic matrix represents a request for a packet from input i to output j. For example $t_{ij} = p$ means that p packets from the i−th input line have to be send to j−th output line of the switch node [5].

A conflict situation arises when in any row of the matrix $\boldsymbol{T}$ the number of requests is more than one. This corresponds to a case when source declares packets for more than one receiver. If any column of a matrix $\boldsymbol{T}$ contain more than one element different from zero, this indicates also a conflict situation. Avoiding conflicts is closely related to the effectiveness of the switch node. In order to obtain the non-conflict schedule is necessary to compute the sequence of conflict-free matrices Q_1 , . . . , Q_m such that their sum will be equal to the traffic matrix $\boldsymbol{T}$. Each column and row of matrices Q_i, $i = 1,2,...,m$ has no more than 1 element equal to one and the rest of elements are equal to 0.

We will give a brief description of the MiMa-algorithm (compute the matrix Q_1).

Step 1. Initially, matrix $\boldsymbol{T}$ is introduced ($t_{ij} : I,j \in \{ 1, 2, \ldots, n\}$, n=const).

Step 2. A vector-column, which consists of the number of conflicts in each row (row conflict weights) is calculated.

Step 3. If there is no requests (vector-column contain only 0-elements) then go to **Step 9**. Else - continue.

Step 4. A vector-row, which consists of the number of conflicts in each column (column conflict weights), is calculated too.

Step 5. In the vector-row we choose the maximal element which determines the column with the most conflicts.

Step 6. In the vector-column we choose the maximal element which determines the row with the most conflicts.

Step 7. If there is a request in the place of intersection of the column and row with most conflicts,

then: we take this request as an element of the non-conflict matrix Q_1. Temporary records zero(0)-weight for these input and output lines. Go to **Step 2**.

else: (If there is no request) we choose the element in the vector-column which is closest in value to the maximal element. The element in the vector row remains the same.

Step 8. We check if there is a request in the intersection, etc. (like **Step 7**, we omit details).

As a result for the chosen column of $\boldsymbol{T}$ we will have a request selected for commutation (if such a request exists at all). The row and column containing the selected request are excluded from the computation of Q_1. Go to **Step 2** (The next elements of Q_1 are computed by repeating the above procedure).

Step 9. Stop.

As a result the first matrix Q_1 may consists of elements (requests) with maximal weight of conflicts in T. The next non-conflict matrices $Q_2, \ldots, Q_m$ are computed analogously. The last matrix Q_m will contain only the non-conflict requests in matrix $\boldsymbol{T}$.

Our model is developed for packet switch node with an equal number of inputs and outputs. Its graphic form is shown on Figure 3. At the first moment of the current modeling time, one token enters into place l_1 (start). This token represents requests for sending a packet (all packets have the same size). It has an initial characteristic : "$ch_0 = (pr_1ch_0, pr_2ch_0) = < n, \boldsymbol{T} >$" (the number of the input/output lines, noted by n and the traffic matrix $\boldsymbol{T}$). The end of the MiMa-algorithm is indicated by receiving a token in the place l_{22} (stop). At this moment the place l_{20} contains the tokens of the final non-conflict schedule (the tokens who represent the solutions $Q_1, Q_2, \ldots, Q_m$).

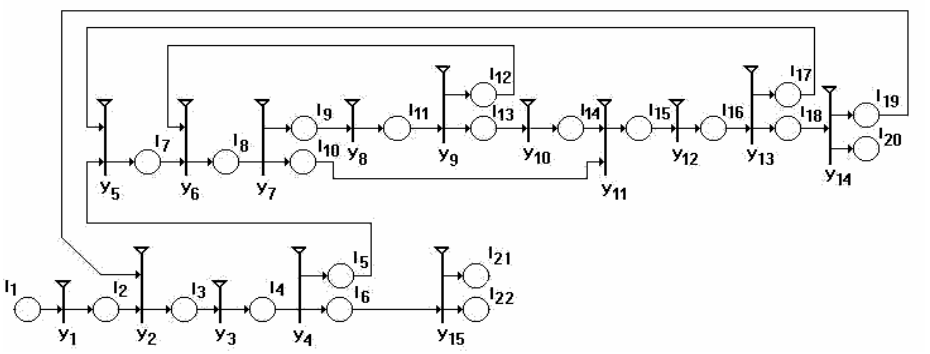


Fig.3 Graphical form of GM-model of MiMa-algorithm

The model has possibilities to provide information about the number of switching in crossbar matrix, as well as about the average number of packets transmitted by one switch. Analysis of the model proves receiving a non-conflict schedule.

3. Models for load traffic

For simulations we will use a modified pattern model for Chang's [8] load traffic. Our basic model includes a family of patterns denoted below as Cg-i , i=1,2,… Requests represent packets with the same size. The index i shows values of element in the traffic matrix. The optimal schedule for Cg-1 requires $k_{opt} = 1.(n-1)$ switchings of the crossbar matrix for n x n switch. The optimal schedule for Cg-i requires $k_{opt} = i.(n-1)$ switchings of the crossbar matrix for n x n switch. The throughput is computed by dividing the result of optimal solution by the result of the simulated solution: if the Mima-algorithm gives the schedule with k_r switchings, then the throughput will be k_{opt} / k_r. This model ensures 100% load intensity of the input lines and 100% workflow of the output lines. The traffic matrices for Cg-1 and traffic matrices for Cg-i are shown in Figure 4. The generation of these patterns does not depend on the type of hardware and software simulation tools.

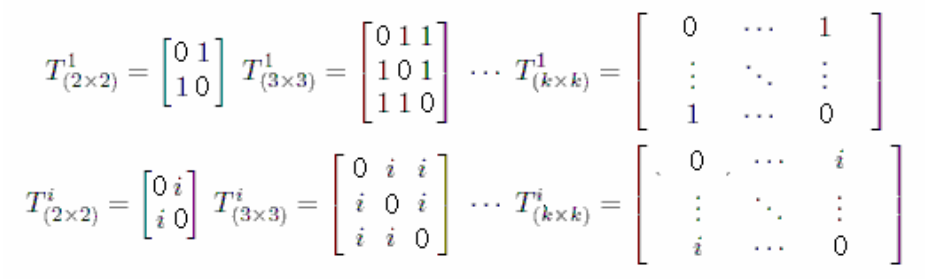


Fig.4 The basic family of patterns for Chang's load traffic: Cg-1 and Cg-i

The MiMa-algorithm is deterministic. For one matrix ***T(kxk)*** it gives one solution $Q_1, Q_2, \ldots, Q_m$ and THR receives quantized values for k from 3 to n. To achieve more precision of simulations in this paper we propose a modification of the family of patterns $Chao_i$, as it is shown below : for $n = 3$ we have three traffic matrices $\boldsymbol{T}^1$ (see Figure 5 for model Cm-1), for $n = k$ we have k traffic matrices $\boldsymbol{T}^i$ (see Figure 5 for Cm-i). The resulting throughput is the average for n runs for each size $n \times n$.

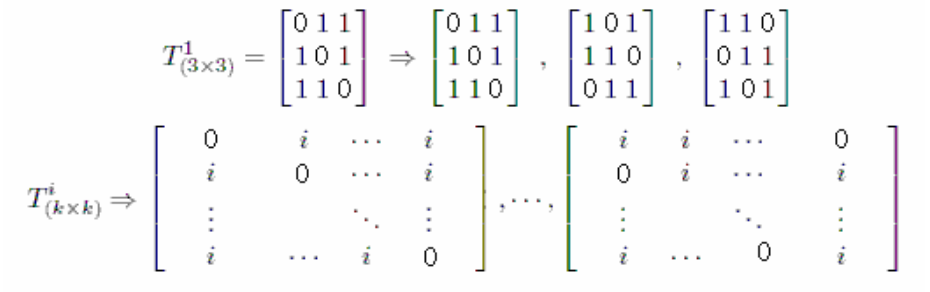


Fig.5 The family of patterns for modified Chang's load traffic: Cm-1 and Cm-i

4. Computer simulation

The transition from the formal model to the executive program is carried out as in [12]. The source codes have been created using the program packet Vfort [13] - free access from Institute of mathematical modeling of Russian Academy of Sciences. The source codes have been compiled with means of the grid-cluster BG01-IPP of the Institute IICT- BAS (www.grid.bas.bg). The resulting code is executed locally in the grid-cluster. The operation system is Scientific Linux release 6.5 (Carbon), kernel 2.6.32-431.20.3.e16.x86 64. We used the following grid-resources: up to 16 CPU (2 blades), 32 threads, 2GB RAM. The main restriction is the time for execution (< 72 hours).

Figure 6 shows the results from computer simulation of the MiMa-algorithm with input data by basic Chang and modified Chang pattern. Sizes of the crossbar matrix used for these simulations range from 3x3 to 130x130. Smoothing of the THR is clearly visible. The curve of calculation time is not only smoothed, but also the time is

significantly reduced. Thus, we conclude that our modification of the basic family of patterns enables us to obtain more precise results in the simulations of MiMa-algorithm with respect to the THR and the time for execution.

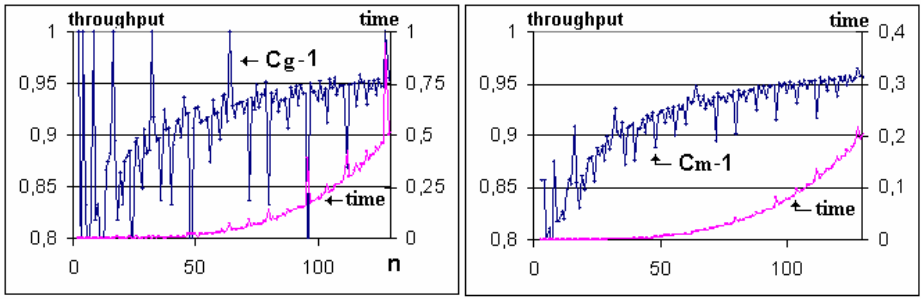


Fig.6 THR and time for basic Chang's load traffic Cg-1 and modified Cm-1

Figure 7 shows the results from computer simulation of the MiMa-algorithm with input data Cm-1, Cm-10, Cm-100. Sizes of the crossbar matrix used for these simulations range from 3x3 to 130x130. To simplify the notations in the figures, the modified pattern is denoted as Cm-j for j=1,10,100… It is shown that the time of execution increases linearly with increasing of the pattern index j. The THR also increases.

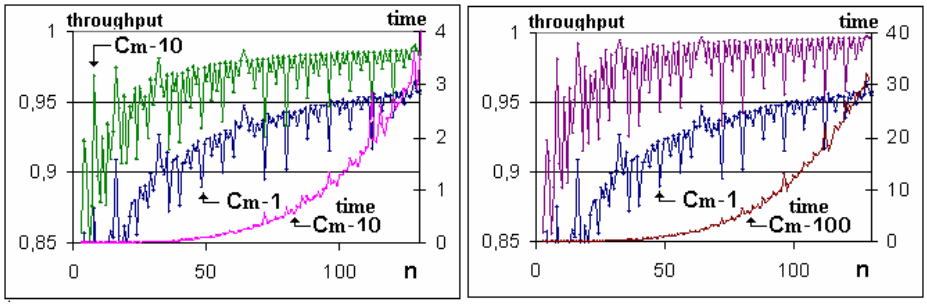


Fig.7 THR and time for modified Chang's load traffic Cm-10 and Cm-100.

This rise the question of whether the 100% THR is a bound for the family of curves for Cm-j when j tends to infinity. In order to answer this question, further simulations are needed.

Figure 8 shows results from computer simulation of the MiMa-algorithm with input patterns Cm-1000 and Cm-10000. A main restriction is the time for execution. The grid-time used for this simulation varies from 58 hours up to 70 hours for Pattern j=10000. The dimension n varies from 3×3 to 100×100 and n simulations (runs) for each size (n×n) of pattern Cm-10000 are executed.

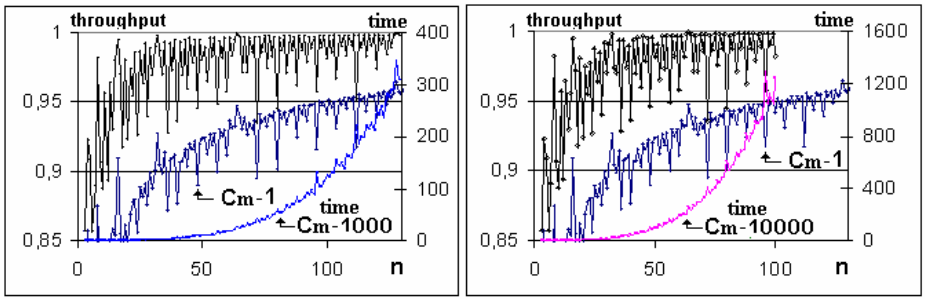


Fig.8 THR and time for modified Chang's load traffic Cm-1000 and Cm-10000.

It is seen that values of throughput for Cm-j are approaching the value of maximum of 100%, when $j \geq 1000$ and $n \geq 64$. As a whole the results obtained are consistent with the expectation that the THR approaches the maximum value of 100%.

5. Conclusion

The main results of the paper include determining of an upper bound of the throughput for MiMa-algorithm under a modified family of patterns. The bound of the throughput of the algorithm approaches the maximal possible throughput value equal to 100%. This is achieved at the expense of an increased time of the algorithm execution. The further investigations should be orientated to optimize the time of execution through parallel computation of operations in MiMa-algorithm.

Acknowledgement

The work reported in the paper were supported by the project "Advanced Computing for Innovation" (AComIn), grant No. 316087, funded by the Capacity Programme (Research Potential of Convergence Regions) FP7.

References

1 Chao, H., B. Lui, "High performance switches and routers". John Wiley & Sons, 2007.

2 Hu, B., K. Yeung, C. He. On Iterative Scheduling for Input-queued Switches with a Speedup of 2-1/N. Proceedings of 15th IEEE Int. Conf. HPSR 2014, July 1-4, 2014, Vancouver, Canada, pp.26-31.

3 Chen, T., J. Mavor, Ph, Denyer, D. Renshaw. Traffic routing algorithm for serial superchip system customisation. IEE Proc.- E.part, vol. 137, no.1. pp.65-73, Jan 1990.

4 Anderson, T., S. Owicki, J. Saxe, and C.Thacker. High speed switch scheduling for local area networks. ACM Trans. Comput. Syst., vol. 11, no.4, P.319-352, Nov. 1993.

5 McKeown, N., "The iSLIP Scheduling Algorithm for Input-Queue Switches," IEEE/ACM Transactions on Networking, vol. 7, no. 2, pp. 188-200, April 1999.

6 Chang, H.J., G. Qu, S.Q. Zheng, “Performance of CTC(N) Switch under Various Traffic Models,” In: Recent Advances in Computer Science and Information Engineering. Springer Berlin Heidelberg, pp. 785-793, 2012.

7 Tashev, T., V. Monov. Modeling of the Hotspot Load Traffic for Crossbar Switch Node by Means of Generalized Nets. Proceedings of 6th IEEE International Conference Intelligent Systems (IS)” 6-8 September 2012, Sofia, Bulgaria. pp.187-191.

8 Chang C-S., D-S. Lee, Y-S. Jou. Load Balanced Birkhoff-von Neumann Switches. Proceedings of IEEE Workshop on HPSR 2001, May 29-31, Dallas, USA, 2001. pp. 276-280.

9 Atanassov, K. Generalized Nets and System Theory. Acad. Press “Prof. M. Drinov”, Sofia, Bulgaria, 1997.

10 Gochev, V., P. Gocheva, “.NET implementation of generalized nets. Object-oriented approach”. Sofia: College of Telecommunications and Post Publishing House, 2012 (in Bulgarian).

11 Shahpazov, G., L. Doukovska. Generalized net model of internal financial structural unit’s functionality with intuitionistic fuzzy estimations. Proc. of the 17th International Conference on Intuitionistic Fuzzy Sets, 2013, Sofia, Bulgaria. Notes on Intuitionistic Fuzzy Sets (NIFS), vol. 19, №3, pp. 111-117.

12 Tashev, T. An estimation of time required for modeling of an algorithm calculate a non-conflict schedule for crossbar switch node by means of grid-structure. Int. J. “Information Theories and Applications”, vol. 19, no 2, pp.149-154, 2012.

13 Vabishchevich, P. VFort. http://www.nomoz.org/site/629615/vfort.html (last checked Apil 14, 2015).

MULTIPLE QUEUING SYSTEM WITH CONTROLLABLE NUMBER OF SERVERS: STATIONARY CASE

A. Mandel, I. Barladyan, A. Tokmakova
V.A.Trapeznikov ICS RAS, Moscow, Russia

Abstract

Controllable multiple queuing system is considered with periodically done control actions: channels may be switch off or switch on. Poisson input flow has variable random intensity which is govern with homogeneous Markovian chain. It is supposed that the length of control period is sufficient for a set-up of probabilistically stationary mode. The criteria is planning horizon maximum profit. Optimal switching strategy computing algorithm has been constructed. The examples are given.

МНОГОЛИНЕЙНАЯ СМО С ИЗМЕНЕНИЕМ ЧИСЛА РАБОЧИХ КАНАЛОВ: СТАЦИОНАРНЫЙ СЛУЧАЙ

А. Мандель, И. Барладян, А. Токмакова
ИПУ РАН им. В.А. Трапезникова, г. Москва, РФ
almandel@yandex.ru, irigina@gmail.com, forabt@ipu.ru

Аннотация

Рассматривается управляемая многолинейная СМО, в которой управление, осуществляемое с постоянным временным шагом, сводится к подключению резервных каналов обслуживания или отключению рабочих. В СМО поступает простейший поток требований со случайной интенсивностью, величина которой изменяется в соответствии с однородной марковской цепью. Предполагается, что за время шага в СМО успевает установиться стационарный режим. Построен алгоритм определения оптимальной по критерию максимума прибыли СМО в интервале планирования стратегии переключений. Приводятся примеры. Ключевые слова: многолинейная СМО, управляемая СМО, однородная марковская цепь, стационарный режим, стратегия переключений.

Ключевые слова: многолинейная СМО, управляемая СМО, однородная марковская цепь, стационарный режим, стратегия переключений

1. Введение

Исследуется модель управляемых систем массового обслуживания (далее СМО), которые имеют широкие применения в сфере анализа деятельности таких систем социально-экономического назначения, как системы продажи различного рода билетов, бронирования мест в гостиницах, торгово-производственные, транспортные и многие другие системы. При этом в качестве динамически управляемого и в данном случае оптимизируемого параметра выступает число устройств обслуживания в многолинейной системе массового обслуживания, которое выбирается как функция от изменяющейся по времени случайной интенсивности входящего потока требований.

Для решения таких задач необходимо располагать достаточно точной моделью подобных управляемых систем, которая строится с применением классических методов теории массового обслуживания [1, 2] и адекватным аппаратом решения соответствующих задач оптимизации (в рассматриваемом случае - динамической). Интерес к управляемым системам массового обслуживания возникает в 70-е годы прошлого столетия. В связи с этим отметим фундаментальную работу В.В. Рыкова [3], в которой была приведена достаточно общая постановка задачи управления системами массового обслуживания и предложены подходы к ее решению. Отмечено, что управлять в системах массового обслуживания можно дисциплиной обслуживания требований, выбором их структуры, воздействиями на входящий поток требований, изменением интенсивности обслуживания требований и рядом других параметров. Оригинальным вкладом в исследование возможности постановки и решения задачи управления так называемыми конфликтными системами массового обслуживания стал труд Ю.И. Неймарка [4].

Интерес к управляемым системам массового обслуживания к концу 80-х заметно вырос в связи с тем, что обрел широчайшее распространение принципиально новый класс прикладных систем, которые можно было интерпретировать и анализировать как системы массового обслуживания - вычислительные и информационные компьютерные сети. Именно поэтому усилия специалистов в области теории массового обслуживания оказались направлены на разработку специфических моделей (в том числе, и сетевых) массового обслуживания, рассчитанных на применение для указанного класса прикладных систем. Особенно следует отметить вклад ученых - представителей школ профессоров В.В. Рыкова, В.М. Вишневского и В.А. Каштанова [5, 6, 7].

Таким образом, в этом отрезке времени, который исчисляется примерно 20 годами (с 1985 по 2005 гг.), практически без внимания большинства специалистов по теории массового обслуживания оставались такие модели управляемых систем массового обслуживания, которые были бы пригодны для описания, анализа и оптимизации функционирования объектов социально-экономической природы. Специфика объектов социально-

экономической природы заключается, прежде всего, в том, что при их описании огромное значение имеет чисто эконометрический анализ их функционирования, выражающийся в тщательном и корректном учете всех экономических показателей, которые необходимо принимать во внимание при оценке качества функционирования таких объектов.

Первая попытка такого исследования была предпринята в работах [8, 9], в которых рассматривалась управляемая в указанном выше смысле СМО, действующая в стационарном (вероятностно). В настоящем докладе рассматривается развитие полученных в [8, 9] результатов.

Рассматриваемые в докладе СМО характеризует то, что входящий поток является пуассоновским потоком с переменной в соответствии с некоторой дискретной по времени однородной марковской цепью интенсивностью. Управление такой системой массового обслуживания сводится к выработке стратегии подключения резервных устройств обслуживания или отключения основных устройств обслуживания (перевода их в резервные).

2. Модели управляемых систем массового обслуживания для описания социально-экономических систем

Будем считать, что наша система массового обслуживания представляет собой многолинейную СМО с набором подключаемых резервных устройств обслуживания.

Все приборы обслуживания имеют экспоненциальные взаимно независимые времена обслуживания с интенсивностью обслуживания μ. Зададим компоненты затрат.

Пусть

c_1 – стоимость эксплуатации одного основного (рабочего) устройства обслуживания в ед. времени;

c_2 – стоимость содержания одного резервного устройства обслуживания в ед. времени (естественно, что $c_1 > c_2$, нередко $c_2 = 0$);

A_1 – цена переключения одного устройства из числа резервных в число основных («включения»);

A_2 – цена переключения одного устройства из числа основных в число резервных («отключения»);

D – стоимость единицы времени пребывания одного требования в очереди на обслуживание;

H – доход, связанный с окончанием обслуживания одного требования.

m_1 – число рабочих каналов обслуживания;

m_2 – число резервных каналов обслуживания.

Требуется максимизировать суммарную среднюю прибыль в системе за время ее функционирования в периоде планирования $[0, T]$. Здесь $T = N\tau$, где N - достаточно большое натуральное число, а τ - интервал времени между последовательными моментами принятия решений о подключении новых рабочих приборов или об отключении части работающих приборов. Интервал τ будем называть «шагом».

Пусть интенсивность входящего простейшего потока в фиксированные моменты времени $t_n = n\tau$, $n = 1, 2 \ldots N$, претерпевает скачкообразные изменения от λ_i к λ_j с вероятностями $p_{ij}(m_1)$, $i, j = 1, 2, \ldots, L$ [1], (в принципе возможен случай, когда переходные вероятности зависят также и от n, тогда вместо $p_{ij}(m_1)$ следует писать $p_{ij}(m_1, n)$). Таким образом, в каждом из интервалов длительности интенсивность входящего потока требований постоянна и описывается моделью случайной величины, принимающей значения из конечного множества $\Lambda = \lambda_1, \lambda_2, \ldots, \lambda_L$ (считаем, что $\lambda_1 < \lambda_2 < \cdots < \lambda_L$). Будем также считать, что $M\mu > \lambda_L$, где - это интенсивность обслуживания на одном основном (рабочем) устройстве, а $M\mu = m_1 + m_2$.

В моменты времени t_n имеется возможность переводить устройства обслуживания из числа резервных в основные и обратно, т.е. осуществлять переключения резервных устройств в основные (включение) или обратные переключения: из основных - в резервные (отключение). Рассматриваемая модель массового обслуживания по своей идее весьма близка к предложенной в 60-е гг. прошлого столетия модели надежности, получившей название динамического резервирования [10, 11], хотя, как станет ясно в дальнейшем ее алгоритмическое решение качественно гораздо ближе к схеме двухуровневых стратегий управления запасами [12].

Будем рассматривать случай установления стационарных режимов в системе массового обслуживания, т.е. случай, когда в каждом из интервалов длительности τ успевает поступить (и пройти обслуживание) такое число требований, которое достаточно велико для того, чтобы в рассматриваемой СМО на каждом шаге устанавливался стационарный в вероятностном смысле режим функционирования. Практически для этого достаточно, чтобы:

$$\mu\tau >> 1 \tag{1}$$

Разумеется, в этом случае справедливо и $\lambda\tau >> 1$.

Предположение стационарности накладывает достаточно жесткое ограничение на рассматриваемую модель.

Во–первых, должны быть выполнены соотношения $\rho_i = \lambda_i/\mu M < 1, \quad \lambda_i \in \Lambda$. Таким образом, система может устойчиво работать только с таким потоком, для которого выполнено условие:

$$\lambda_i < \mu M \quad \forall \lambda_i \in \Lambda. \tag{2}$$

[1] Введение зависимости вероятностей скачков интенсивности входящего потока от числа рабочих устройств m_1 связано с необходимостью учета эффекта влияния качества процесса обслуживания (в данном случае обусловленного временем ожидания клиентами начала обслуживания) на потребительский спрос, который и описывает рассматриваемая интенсивность входящего потока.

Пусть λ_{max} максимальный элемент множества Λ (фактически, это значение λ_L). Тогда неравенство (2) можно переписать в виде

$$\lambda_{max} < \mu M \tag{3}$$

Во–вторых, на каждом шаге существует нижняя граница допустимого числа рабочих устройств $\underline{m}_{1p}^{(i)}$, на которое можно переводить систему с расчетом на ее устойчивое функционирование (в смысле существования стационарного режима), и эта граница зависит от номера состояния i на соответствующем шаге. Действительно, в состоянии i (по значению интенсивности входящего потока) должно выполняться условие стационарности $\rho_i = \lambda_i/\mu m_{1p}^{(i)} < 1$, откуда получаем что:

$$\underline{m}_{1p}^{(i)} > \frac{\lambda_i}{\mu} \Rightarrow \underline{m}_{1p}^{(i)} = \frac{\lambda_i}{\mu} + 1, \forall i = \overline{1, N} \tag{4}$$

Таким образом, число работающих в состоянии i устройств может принимать значения от $\underline{m}_{1p}^{(i)}$ до M.

3. Постановка и решение задачи для стационарной модели

Обозначим через $C^{(1)}(\lambda_i, m_1, m_{1p})$ среднее значение прибыли на одном шаге (длительности τ), если в начале шага устанавливается значение интенсивности входящего потока, равное λ_i; число основных (рабочих) устройств, с которыми СМО подходит к этому шагу, равно m_1, и принимается управляющее решение о введение в действие в момент начала шага m_1 основных устройств. Что означает сохранение прежнего числа работающих устройств обслуживания, если $m_1 = m_1$; необходимость отключения $m_1 - m_1$ устройств, если $m_1 - m_1 > 0$; и необходимость включения $m_1 - m_1$ устройств, если $m_1 - m_1 < 0$.

Очевидно, что $C^{(1)}(\lambda_i, m_1, m_{1p})$ = средний доход за один шаг – средние одношаговые затраты.

Из предположения о стационарности следует, что выполняются все входящие требования, а значит, в любом случае средний доход за один шаг составляет $h\lambda_i$.

Средние затраты складываются из затрат на включение резервных (отключение основных) устройств; затрат, связанных с пребыванием требований в очереди; и затрат на эксплуатацию основных и резервных устройств. Затраты на включение (отключение) равны:

$$З_{\text{переключения}}(m_1, m_1) = \begin{cases} A_1(m_1 - m_1) & \text{если } m_1 - m_1 > 0, \\ 0 & \text{если } jm_1 = m_1, \\ A_1(m_1 - m_1) & \text{если } m_1 - m_1 < 0. \end{cases} \tag{5}$$

Затраты, связанные с пребыванием требований в очереди на шаге длительности τ, равны

$$З_{\text{на очередь}}(\lambda_i, m_{1p}) = \mathbf{E}\int\limits_0^\tau dk(t, m_{1p}, \lambda_i)dt =$$
$$= \mathbf{E}d\int\limits_0^\tau k(t, m_{1p}, \lambda_i)dt = d\overline{k}(m_{1p}, \lambda_i)\tau, \tag{6}$$

где $\mathbf{E}$ - символ оператора математического ожидания, $k\lfloor t, m_{1p}, \lambda_i\rfloor$- число требований в очереди к СМО с параметрами m_1 и λ_i в момент времени t, а $\overline{(m_{1p}, \lambda_i)}$ - средняя длина очереди в СМО с параметрами m_1 и λ_i. Интегрирование в силу предположения о стационарности ведется не в текущем интервале $[t_n, t_{n+1}]$, а в интервале $[0, \tau]$. Для стационарного случая имеем [1] :

$$\overline{k}(m_{1p}, \lambda_i) = \pi_0 \frac{(m_{1p}\rho_i)^{m_{1p}}\rho_i}{m_{1p}!(1-\rho_i)^2}, \tag{7}$$

где $\rho_i = \frac{\lambda_i}{\mu m_{1p}}$, а $\pi_0 = \left[\sum_{i=0}^{m_{1p}} \frac{(m_{1p}\rho_i)^i}{i!} + \frac{(m_{1p}\rho_i)^{m_{1p}}}{m_{1p}!(1-\rho_i)}\right]^{-1}$ – стационарная вероятность того, что в системе нет требований.

Параметр d характеризует косвенные издержки рассматриваемой СМО, связанные с тем, что увеличение длительности времени пребывания требований в очереди снижает предпочтения клиентов по обращению в рассматриваемую СМО. Можно построить математическую модель, которая связывала бы уровень утраты предпочтений со средним временем пребывания требований в очереди в случае его увеличения, и на основе этой модели сформировать обоснованную оценку значения коэффициента d. В настоящей работе задача построения подобной модели не рассматривается. Ограничимся замечанием, что в большинстве случаев значение параметра d на 1,5 – 2 порядка меньше значения введенного выше параметра h.

Затраты на эксплуатацию основных и резервных устройств за один шаг составляют:

$$З_{\text{эксплуатации}}(m_{1p}) = c_1 m_{1p} + c_2(M - m_{1p}). \tag{8}$$

Таким образом, средняя чистая прибыль за шаг равна:

$$C^{(1)}(\lambda_i, m_1, m_{1p}) = h\lambda_i - З_{\text{переключения}} - З_{\text{на очередь}} - З_{\text{эксплуатации}} \tag{9}$$

Пусть $П_s^*(\lambda_i, m_1)$ - максимальное значение средней прибыли на интервале, который начинается за s шагов до конца периода планирования $[0, T]$, $s+n = N$, при значении λ_i интенсивности входящего потока и m_1 включенных (до принятия управляющего решения о включении m_{1p} устройств) основных устройствах. Ниже выводятся уравнения динамического программирования для функционала $П_s^*(\lambda_i, m_1)$ с учетом условия (4).

Очевидно, что за один шаг до конца периода планирования при случайном значении интенсивности входящего потока λ_i значение введенного выше функционала $\Pi_1^*(\lambda_i, m_1)$ запишется как

$$\Pi_1^*(\lambda_i, m_1) = \max_{\underline{m_{1p}^{(1)}} \le m_{1p} \le M} C^{(1)}(\lambda_i, m_1, m_{1p}) \tag{10}$$

где$\lambda_i \in \Lambda$.

За s шагов до конца периода планирования имеем:

$$\Pi_s^*(\lambda_i, m_1) = \max_{\underline{m_{1p}^{(1)}} \le m_{1p} \le M} \left\{ C^{(1)}(\lambda_i, m_1, m_{1p}) + \sum_{j=1}^{L} p_{ij}(m_{1p}) \Pi_{s-1}^*(\lambda_j, m_{1p}) \right., \tag{11}$$

где $\forall s \in \overline{2, N-1}, \quad \lambda_i, \lambda_j \in \Lambda$, – значения интенсивностей входящего потока в моменты времени за s и за $s-1$ шагов до конца периода планирования соответственно.

Уравнения (10) и (11) используются для расчета программы оптимального управления $m_{1p}^*(i, s)$, где $1 \le s \le N-1, \quad i = 1, 2, \ldots, L$.

4. Компьютерное моделирование

Следует отметить, что в реальных прикладных проблемах задачах входящий поток требований редко бывает чисто пуассоновским. При этом нередко можно отследить четкие закономерности характера скачкообразных изменений интенсивности входящего потока требований. Перечислим наиболее распространенные из них:

- **«Плато».** Частота изменений интенсивности входящего потока требований невелика. Поэтому достаточно долгое время рассматриваемая СМО функционирует как чисто пуассоновская, вступая при этом в стационарный (вероятностно) режим.
- **«Пик».** Входящий поток почти все время стационарен, однако имеют место периоды пиковых скачков нагрузки. Например, системы биллинга испытывают пиковые нагрузки утром и вечером, и на фоне этого еще более резкие скачки в праздничные дни. Утренние и вечерние пики также характерны для парков общественного транспорта (днем нагрузка сравнительно невысока).

Исследуемая имитационная модель и семантически, и функционально учитывает отмеченные особенности поведения входящего потока.

При моделировании рассматривались два класса систем:

Класс 1 ($c_1 > \mu h; A_2 >> c_1$) :

- высокая стоимость содержания рабочих устройств обслуживания;
- низкая стоимость перевода резервных устройств обслуживания в рабочий режим;

- высокая стоимость отключения рабочих устройств.

Класс 2 ($c_1 << \mu h; A_2 \approx A_1$) :

- низкая стоимость содержания рабочих устройств обслуживания;
- издержки на включение и отключение устройств обслуживания невысоки и примерно равны между собой.

Исследуем зависимость решений от параметров систем при отмеченных выше двух типах входящего потока. Для каждого эксперимента представим построенные в результате моделирования в среде MATLAB 7.0 графики зависимостей $m_{1p}(t) - m^*_{1p}(t)$.

Значения N, τ, μ, и c_2 задаются для всей серии экспериментов. При выборе значения μ следует учитывать ограничение $\mu >> 1$, которое связано с предположением о стационарности режима, устанавливающегося в СМО, а с учетом предположения (3) и того, что, как правило, $c_1 >> c_2$, положим $c_2 = 0$. Для удобства будем считать, что $\tau = 1$.

Для задания «плато» используется матрица переходных вероятностей вида

$$\mathbf{Prob}_{\text{плато}} = \left\| \begin{array}{cccc} 0,9 & 0,09 & 0,09 & 0,9 \\ 0,01 & 0 & 0,01 & 0 \\ 0 & 0,05 & 0 & 0,09 \\ 0,9 & 0,05 & 0,01 & 0,9 \end{array} \right\| \tag{12}$$

Наибольший интерес представляют компоненты затрат c_1, A_1, A_2. Именно они и изменяются в ходе проведенных экспериментов. В процессе моделирования множество возможных значений интенсивностей входящего потока задавалось как $\Lambda = \{100, 150, 270, 350\}$, $\mu = 20$, $\tau = 1$, $c_2 = 0$.

На приводимых ниже рисунках зависимость $m_{1p}(t)$ изображена точечной линией, а $m^*_{1p}(t)$ – сплошной.

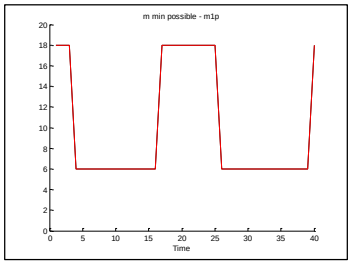


Рис. 1: (Обе линии совпадают). Класс 2. $A_1 = A_2 = 10, c_1 = 5, h = 1, d = 0,15$.

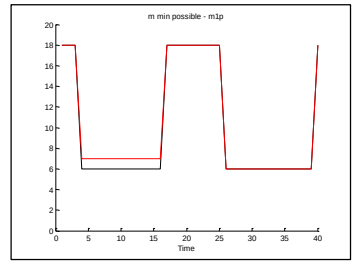


Рис. 2: Класс 2.$A_1 = A_2 = 30, c_1 = 5, h = 1, d = 0,15$

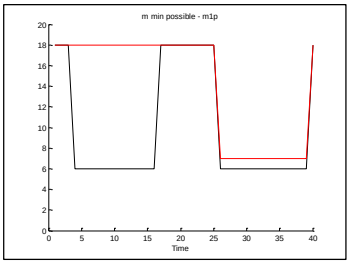


Рис. 3: Класс 1. $A_1 = 2, A_2 = 190, c_1 = 15, h = 1, 3, d = 0, 15.$

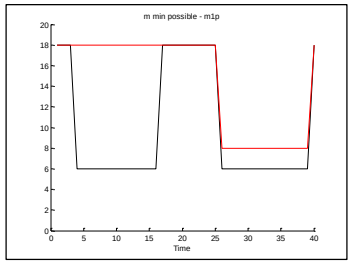


Рис. 4: Класс 1.$A_1 = 2, A_2 = 202, c_1 = 5, h = 1, d = 0, 15$

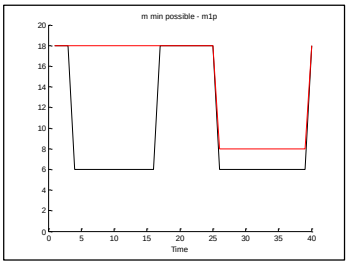


Рис. 5: (Класс 1. $A_1 = 2, A_2 = 204, c_1 = 15, h = 1, 3, d = 0, 15.$

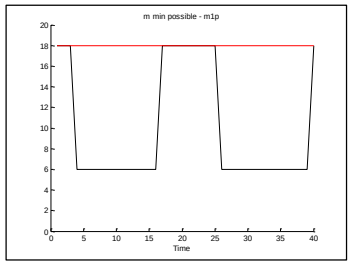


Рис. 6: Класс 2.$A_1 = 2, A_2 = 202, c_1 = 5, h = 1, 3, d = 0, 15$

В данном случае оба класса систем характеризуются схожей динамикой. С ростом компонент затрат, связанных с переключением устройств обслуживания, решение постепенно перестает «отзываться» на резкие спады интенсивности. Для класса 1 это начинает наблюдаться при $c_1 : A_2 \approx 1 : 12$, для класса 2 - при $c_1 : A_{1,2} \approx 1 : 6$ (рис. 2). При дальнейшем увеличении A_2 относительно c_1 эта тенденция сохраняется (рис. 3; 5; 6). При соотношении $c_1 : A_2 \approx 1 : 13$ для класса 1 и $c_1 : A_{1,2} \approx 1 : 7$ для класса 2 переключение устройств обслуживания становится невыгодным. Оптимальным решением является включение в рабочий режим числа устройств обслуживания, равного максимальной компоненте вектора $\underline{m}_{1p}$.

5. Заключение

Предложена модель управляемой многолинейной системой массового обслуживания, с помощью которой можно проанализировать и оптимизировать функционирование различных классов объектов социально-эконо-

мической природы: систем продажи и бронирования билетов и мест в гостиницах, торгово-закупочных систем, систем общественного транспорта, билинговых систем и ряда других. Интенсивность простейшего входящего потока изменяется в соответствии с однородной марковской цепью. В предположении об установлении в таких системах стационарного (вероятностно) режима построены алгоритмы формирования оптимальных стратегий переключения каналов из резервных в основные и наоборот.

ЛИТЕРАТУРА

1. Гнеденко Б.В., Коваленко И.Н. Введение в теорию массового обслуживания.// – М.: Наука, 1966.
2. Саати Т.Л. Элементы теории массового обслуживания и ее приложения. //– М.: Сов. Радио, 1971.
3. Рыков В.В. Управляемые системы массового обслуживания // Теория вероятностей. Математическая статистика. Теоретическая кибернетика. Том 12, 1975. – С. 43–153.
4. Неймарк Ю.И. Динамические системы и управляемые процессы. М.: Наука, 1978.
5. Голышева Н.М., Федоткин М.А. Циклическое управление конфликтными потоками в условиях гибели и рождения очередей критических размеров // Автоматика и телемеханика, №4, 1990.– С.68–75.
6. Вишневский В.М. Теоретические основы проектирования компьютерных сетей. М.: Техносфера. 2003.
7. Захаров П.П. Разработка автоматизированного программного комплекса для исследования качества и эффективности функционирования моделей технических систем и управляемых систем массового обслуживания // Диссертация на соискание ученой степени кандидата технических наук. М.: МГИЭМ, 2006.
8. Кузнецов А.В., Мандель А.С., Токмакова А.Б. Об одной модели управляемой системы массового обслуживания // Проблемы управления. 2007, №6. - С. 39-43.
9. Барладян И.И., Кузнецов А.В., Мандель А.С. Анализ критических значений параметров и моделирование управляемой системы массового обслуживания // Проблемы управления. 2007, №6. - С. 21–25.
10. Райкин А.Л. Элементы теории надежности для проектирования технических систем. – М.,: Сов. Радио, 1967. – 256 с.
11. Мандель А.С., Райкин А.Л. Формирование оптимального плана включения запасных элементов// Автоматика и телемеханика, №5, 1967. С. 55–63.
12. Хедли Дж., Уайтин Т. Анализ систем управления запасами.// – М.: Наука, 1969. 512 с.

MULTIPLE QUEUING SYSTEM WITH CONTROLLABLE NUMBER OF SERVERS: NON-STATIONARY CASE

A. Mandel [1], *V. Makhukova* [2]
[1] V. Trapeznikov ICS RAS, Moscow, Russia
[2] M. Lomonosov MSU, Moscow, Russia

Controllable multiple queuing system is considered with periodically done control actions: channels may be switch off or switch on. Poisson input flow has variable intensity which is a given piecewise constant function of time. The criteria is planning horizon maximum profit. Optimal switching strategy computing algorithm has been constructed.

МНОГОЛИНЕЙНАЯ СМО С ИЗМЕНЕНИЕМ ЧИСЛА РАБОЧИХ КАНАЛОВ: НЕСТАЦИОНАРНЫЙ СЛУЧАЙ

А. Мандель[1], *В. Махукова*[2]
[1] ИПУ РАН им. В.А. Трапезникова, г. Москва, РФ
[2] МГУ им. М.В. Ломоносова, г. Москва, РФ,
[1]almandel@yandex.ru, [2]makhukova.vladilena@physics.msu.ru

Аннотация

Рассматривается управляемая многолинейная СМО, в которой управление сводится к подключению резервных каналов обслуживания или отключению рабочих. В СМО поступает простейший поток требований с интенсивностью, величина которой меняется в соответствии с заданной кусочно постоянной функцией времени. Построен алгоритм определения оптимальной по критерию максимума прибыли СМО в интервале планирования стратегии переключений.

Ключевые слова: многолинейная СМО, управляемая СМО, простейший поток с переменной интенсивностью, стратегия переключений

1. Введение

Рассматривается модель управляемой системы массового обслуживания (далее СМО) с периодическим контролем за интенсивностью входящего потока, изменяющейся скачкообразно в соответствии с некоторой заданной кусочно постоянной функцией времени. При этом в качестве динамиче-

ски управляемого, оптимизируемого параметра выступает число устройств обслуживания в многолинейной системе массового обслуживания.

Настоящая работа является попыткой обобщения исследования, предпринятого в работах [1], [2]], в которых рассматривалась управляемая в указанном выше смысле СМО с входящим потоком случайной интенсивности.

Управление такой системой массового обслуживания сводится к выработке стратегии подключения резервных устройств обслуживания или отключения основных устройств обслуживания (перевода их в резервные).

2. Модель управляемой системы массового обслуживания

Итак, будем считать, что наша система массового обслуживания представляет собой многолинейную СМО с набором подключаемых резервных устройств обслуживания, схематично представленную на рис. 1.

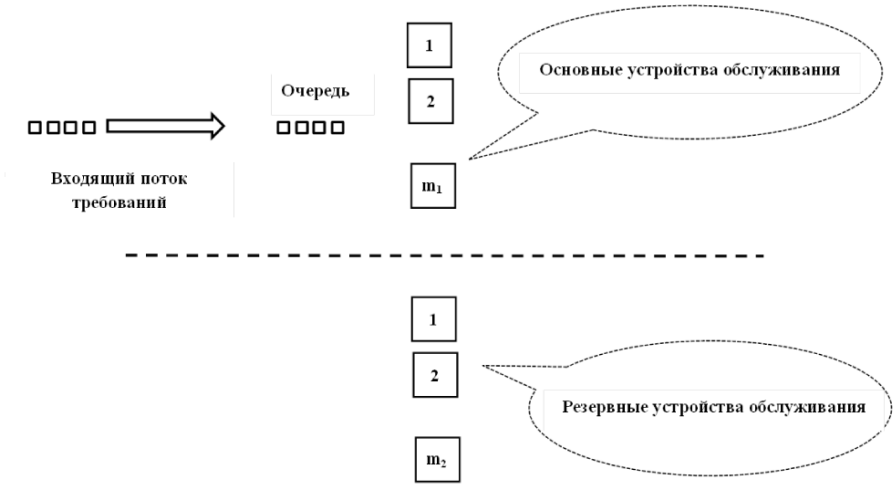


Рис. 1: Структура управляемой системы массового обслуживания.

Все приборы обслуживания имеют экспоненциальные взаимно независимые времена обслуживания с интенсивностью обслуживания μ . Зададим компоненты затрат.

Пусть

c_1 – стоимость эксплуатации одного основного (рабочего) устройства обслуживания в ед. времени;

c_2 – стоимость содержания одного резервного устройства обслуживания в ед. времени (естественно, что $c_1 > c_2$, нередко $c_2 = 0$);

A_1 – цена переключения одного устройства из числа резервных в число основных («включения»);

A_2 – цена переключения одного устройства из числа основных в число резервных («отключения»);

D – стоимость единицы времени пребывания одного требования в очереди на обслуживание;

H – доход, связанный с окончанием обслуживания одного требования.

m_1 – число рабочих каналов обслуживания;

m_2 – число резервных каналов обслуживания.

Требуется максимизировать суммарную среднюю прибыль в системе за время ее функционирования в периоде планирования $[0, T]$. Здесь $T = N\tau$, где N - достаточно большое натуральное число, а τ - интервал времени (шаг) между последовательными моментами принятия решений о подключении новых рабочих приборов или об отключении части работающих приборов.

Рассматривается случай нестационарной модели системы массового обслуживания, т.е. случай, когда каждый из интервалов длительности τ недостаточен для того, чтобы в рассматриваемой СМО возник стационарный в вероятностном смысле режим.

Для упрощения модели предположим, что изменения величины интенсивности входящего потока возможны только в моменты контроля состояния СМО[1]. Во избежание переусложнения модели допустим также, что, какие бы управляющие решения ни принимались, они не должны выводить СМО в состояния, в которых коэффициент загрузки СМО $\rho \geq 1$.

Будем также считать, что на каждом шаге $n = 1, 2, \ldots, N$, в СМО поступает простейший входящий поток интенсивности λ_n , где множество $\Lambda_{\text{возм}}$ возможных значений интенсивностей λ_n представляет собой конечное множество вида $\Lambda_{\text{возм}} = \{\lambda^{(1)}, \lambda^{(2)}, \ldots, \lambda^{(M)}\}$. При этом считается, что график изменения интенсивности входящего потока по шагам дискретного времени n задан в форме:

$$\lambda(t) = \lambda^{(j)}, t \in [t_{N-n+1}, t_N - n], n = 1, 2, \ldots, N, \tag{1}$$

где в обозначении индексов временных границ шага использована конструкция «обратного» времени, отсчитываемого (в шагах) от конца периода планирования $[0, T]$. По сути, график изменения интенсивности входящего потока задается в форме набора индексов $J = \{j_n\}_{(n=1)}^{N}$.

Теперь состояние системы в каждый из моментов t_n описывается числом требований i, который в этот момент времени находятся в СМО.

Посмотрим, что означает предположение $\rho \geq 1$.

Это с необходимостью означает, что при $t \in [t_n, t_{n-1}]$ должны быть выполнены соотношения $\rho^{(j_m)} = \lambda^{(j_m)}/\mu M < 1, \lambda^{(j_m)} \in \Lambda_{\text{возм}}$. Таким образом, система может устойчиво работать только тогда, когда

$$\lambda^{(j)} < \mu M, \forall \lambda^{(j)} \in \Lambda_{\text{возм}} \tag{2}$$

[1]Для этого нужно выбрать величину шага τ достаточно малой.

Пусть λ_{max} – максимальный элемент множества Λ (фактически, это значение $\lambda^{(M)}$). Тогда неравенство (2) можно переписать в виде

$$\lambda_{max} < \mu M. \tag{3}$$

Во-вторых, на каждом шаге существует нижняя граница допустимого числа рабочих устройств $\underline{m}_{1p}^{(i)}$, на которое можно переводить систему с расчетом на ее устойчивое функционирование, и эта граница зависит от номера состояния i на соответствующем шаге.

Действительно, в состоянии j_n (по значению интенсивности входящего потока), как отмечено выше, должно выполняться условие стационарности $\rho^{(j_m)} = \lambda^{(j_m)}/\mu M < 1, \lambda^{(j_m)} \in \Lambda_{\text{возм}}$, откуда получаем что:

$$\underline{m}_{1p}^{(j_m)} > \frac{\lambda^{(j_m)}}{\mu} \Rightarrow \underline{m}_{1p}^{(j_m)} = \frac{\lambda^{(j_m)}}{\mu} + 1, \forall \lambda^{(j_m)} \in \Lambda_{\text{возм}} \tag{4}$$

Таким образом, число работающих в состоянии I устройств может принимать значения от $\underline{m}_{1p}^{(j_m)}$ до M.

3. Постановка и решение задачи

Введем величину $C^{(1)}(\lambda^{(j)}, i, m_1, m_{1p})$, равную среднему значению прибыли на одном шаге (длительности τ), если в начале шага устанавливается значение интенсивности входящего потока, равное $\lambda^{(j)}$; в это время в СМО имеется i требований, число основных (рабочих) устройств, с которыми СМО подходит к этому шагу, равно m_1, и принимается управляющее решение о введение в действие в момент начала шага m_{1p} основных устройств. При этом формулы (5) и (8) из работы [3] сохранятся, а формулы (6), (9), (10) и (11) из той же работы [3] подлежат корректировке.

В самом деле, стационарного режима на каждом шаге теперь не возникает и, стало быть, не работают формулы для расчета характеристик стационарного режима, включая формулы для числа обработанных на шаге требований и для средней длины очереди. Ключевым инструментом расчета становятся вероятности состояний СМО на очередном шаге, то есть вероятности $Pi(t)$ того, что в момент t (напомним, что $t \in [t_{N-n+1}, t_{N-n}]$) в СМО имеется i требований. Запишем систему дифференциальных уравнений для вероятностей $Pi(t)$ [4]:

$$\begin{aligned}
&\text{для } i = 0: \quad \frac{dP_0(t)}{dt} = -\lambda P_0(t) + \mu P_1(t).\\
&\text{для } 1 \le i < m_{1p}: \frac{dP_i(t)}{dt} = \lambda P_{(i-1)}(t) - (\lambda + i\mu) P_i(t) + (i+1)\mu P_{(i+1)}(t).\\
&\text{для } i \ge m_{1p}: \frac{dP_i(t)}{dt} = \lambda P_{(i-1)}(t) - (\lambda + m_{1p}\mu) P_i(t) + m_{1p}\mu P_{(i+1)}(t),
\end{aligned} \tag{5}$$

с начальным условием

$$P_j(t) = \begin{cases} 1 & \text{если } j = i, \\ 0 & \text{если } j \neq i \end{cases}. \tag{6}$$

Чтобы не пытаться решать бесконечномерную систему дифференциальных уравнений (5)–(6), заменим ее аппроксимирующей системой конечной размерности I:

$$\begin{aligned}
&\text{для } i = 0: \quad \frac{dP_0^I(t)}{dt} = -\lambda P_0^I(t) + \mu P_1^I(t). \\
&\text{для } 1 \leq i < m_{1p}: \frac{dP_i^I(t)}{dt} = \lambda P_{(i-1)}^I(t) - (\lambda + i\mu) P_i^I(t) + (i+1)\mu P_{(i+1)}^I(t). \\
&\text{для } m_{1p} \leq i \leq I: \frac{dP_i^I(t)}{dt} = \lambda P_{(i-1)}^I(t) - (\lambda + m_{1p}\mu) P_i^I(t) + m_{1p}\mu P_{(i+1)}^I(t),
\end{aligned} \tag{7}$$

с начальным условием

$$P_j^{(I)}(t) = \begin{cases} 1 & \text{если } j = i, \\ 0 & \text{если } j \neq i \end{cases}. \tag{8}$$

Представим себе, что мы располагаем «решателем» системы (7)–(8). Обозначим решение системы (7), имея в виду начальное условие (8) $P^{(I)}(i, \tau)$. Второй аргумент τ «напоминает» о том, что решение системы (7)–(8) строится в интервале $[0, \tau]$. Итак, считая, что мы располагаем приближенным решением системы (5)–(6) в форме вероятностей $P_i^{(I)}(t), i = 1, 2, \ldots, I$, $t \in [0, \tau]$, перейдем к расчету недостающих характеристик.

В дальнейшем расчет будет вестись в обратном времени s, так что теперь график изменения интенсивности входящего потока будет представлен значениями $\lambda^{(N-s)}, s = 1, 2, \ldots, N$. Переобозначим эти интенсивности $\lambda^{(N-s)} = \lambda^{(s)}_{\text{модиф}}$. Пусть $\Pi_s^*(\lambda, i, m_1)$ - максимальное значение средней прибыли на интервале, который начинается за s шагов до конца периода планирования $[0, T]$, $s + n = N$, при значении λ (естественно, что $\lambda = \lambda^{(s)}_{\text{модиф}}$) интенсивности входящего потока, числе требований в системе i и m_1 включенных (до принятия управляющего решения о включении $m_1 p$ устройств) основных устройствах. Ниже выводятся уравнения динамического программирования для функционала $\Pi_s^*(\lambda, i, m_1)$, но уже без учета условия (4).

Очевидно, что за один шаг до конца периода планирования при случайном значении интенсивности входящего потока λ значение введенного выше функционала $\Pi_1^*(\lambda, i, m_1)$ запишется как

$$\Pi_1^*(\lambda, i, m_1) = \max_{1 \leq m_{1p} \leq M} C^{(1)}(\lambda, i, m_1, m_1 p), \tag{9}$$

где $\lambda \in \lambda_{\text{возм}}$

Перейдем теперь к корректировке формулы (6) для затрат на очередь из работы [3]. Средняя длина очереди $k(t, m_{1p}, \lambda)$ в момент времени $t \in [t_s, t_{s-1}]$, где s – обратное время, может быть представлена в следующем виде:

$$k(t, m_{1p}, \lambda) = \sum_{i=m_{1p}+1}^{I} (i - m_{1p}) P_i^{(I)}(t). \tag{10}$$

Тогда для затрат на требования, пребывающие в очереди, имеем

$$З_{\text{на очередь}}(\lambda, m_{1p}) = d \int_{t_s}^{t_{s-1}} (t, m_{1p}, \lambda) dt. \tag{11}$$

Доходы от обслуживания требований в интервале $[t_s, t_{s-1}]$ оценим приближенно, используя тождество Вальда [5]. Будем оценивать среднее число требований, обслуженных в интервале $[t_s, t_{s-1}]$, интегралом от мгновенной интенсивности обслуживания в момент t, которая равна

$$\mu(t) = \mu \sum_{i=0}^{m_{1p}-1} P_i^{(I)}(t) i + m_{1p} \mu \sum_{i=m_{1p}}^{I} P_i^{(I)}(t) \tag{12}$$

Таким образом, доход от обслуживания требований в интервале $[t_s, t_{s-1}]$ оценим величиной $h \int_{t_s}^{t_{s-1}} \mu(t) dt$. Это выражение позволяет переписать формулу (9) для средней чистой прибыли за 1 шаг из работы [3] в виде

$$C^1(\lambda, i, m_1, m_{1p}) = h \int_{t_s}^{t_{s-1}} \mu(t) dt - З_{\text{переключения}} - З_{\text{на очередь}} - З_{\text{эксплуатации}}, \tag{13}$$

где три компоненты затрат в правой части (13) вычисляются по формулам (5), (18) и (8) из работы [3].

Теперь за s шагов до конца периода планирования имеем:

$$\begin{gathered} П_s^*(\lambda_{\text{модиф}}^{(s)}, i, m_1) = \\ = \max_{1 \le m_{1p} \le M} \{ C^{(1)}(\lambda_{\text{модиф}}^{(s)}, i, m_1, m_1) + \sum_{j=o}^{I} P_j^{(I)}(t) П_{s-1}^*(\lambda_{\text{модиф}}^{(s-1)}, j, m_{1p}) \} \end{gathered} \tag{14}$$

где $\forall s \in \overline{2, N-1}$.

4. Заключение

Предложена новая модель управления многолинейной системой массового обслуживания, с помощью которой можно проанализировать и оптимизировать функционирование различных классов объектов социально-экономической природы: систем продажи и бронирования билетов и мест в гостиницах, торгово-закупочных систем, систем общественного транспорта, билинговых систем и ряда других. Обозначены перспективы развития решения поставленной проблемы. В частности, очевидно, что решение системы дифференциальных уравнений (7)–(8) требует использования самых современных вычислительных систем – многопроцессорных кластеров, Это подразумевает необходимость распараллеливания предложенных в работе алгоритмов.

ЛИТЕРАТУРА

1. Кузнецов А. В., Мандель А. С., Токмакова А. Б. Об одной модели управляемой системы массового обслуживания // Проблемы управления. 2007, є6. - С. 39-43.
2. Барладян И. И., Кузнецов А. В., Мандель А. С. Анализ критических значений параметров и моделирование управляемой системы массового обслуживания // Проблемы управления. 2007, є6. - С. 21-25.
3. Мандель А. С., Барладян И.И., Токмакова А.Б. Многолинейная СМО с изменением числа рабочих каналов: стационарный случай. / Труды настоящей конференции. – С.??-??.
4. Гнеденко Б. В., Коваленко И. Н. Введение в теорию массового обслуживания. // - М.: Наука, 1966.
5. Саати Т. Л. Элементы теории массового обслуживания и ее приложения // - М.: Сов. Радио, 1971.

THE SYNTHESIS OF SERVICE DISCIPLINES IN SYSTEMS WITH LIMITS

T.Aliev
ITMO University, St.Petersburg, Russia
aliev@cs.ifmo.ru

Abstract

The problem of synthesis of the disciplines of service in the class of subjects with mixed priorities in the presence of combined restrictions on the times of stay in the system for different classes of applications. The algorithm of allocation priorities, providing fulfilling given constraints with minimal system performance proposed.

Keywords: service discipline, mixed priorities, delay time, limits, device performance.

1. Introduction

One of the main characteristics of the functioning of computer systems of different classes and computer networks is the delay time of data processing and transmission, which is named as residence time of the system in queuing systems used as models. The requirements for the quality of the functioning of such systems are formulated in a variety of limits on the residence times in the different classes of applications. For example, in data-processing systems, restrictions apply to the mean residence time of the queries in the system (the average limits), the excess of these limits does not lead to critical consequences. At the same time, in information and control systems, existing in a circuit of automatic control of technological equipment or mobile units, restrictions are imposed on the probability of exceeding the set value of residence time (probabilistic limit) which may lead to a sharp deterioration in the functioning of the system or even to the exit its failure. In general, the quality requirements of the system can be formulated as a combination of limits, where some classes have an average constraints, while others - probabilistic constraints. Applying limits ensuring correct application of the priority strategies of process control and data processing. Queuing system with one serving device and unlimited storage capacity, which receives the inhomogeneous flow of requests to be processed with a given capacity are widely used as models of such systems [1]. In this case, the problem of synthesis service disciplines comes to the distribution of priority to provide the desired limits on the residence time in the different classes of applications with minimal system performance.

2. Statement of a problem

Load parameters for solving the problem of synthesis to a system with a non-uniform flow of applications used: the number of classes of applications are: the

number of classes of applications - H, coming in from the intensities $\lambda_1, \ldots, \lambda_H$ and forms a simple flows; resource-processing applications each class, asked for at least three factors: average values $\theta_1, \ldots, \theta_H$ (commands or instructions executed by the processing of the application of the corresponding class); coefficients of variation $\nu_1, \ldots, \nu_H$ and third initial moments $\theta_1^{(3)}, \ldots, \theta_H^{(3)}$.

Suppose that the classes of applications are grouped so that the restrictions on the residence times $\tau_{u_1}, \ldots, \tau_{u_H}$ in the system for the first H_1 classes are defined in a probability, and for other grades - as mean limited to:

$$\Pr\left(\tau_{u_h} > u_h^*\right) \le \delta_h^*, \qquad (h = \overline{1, H_1}), \tag{1}$$

$$u_h \le u_h^*, \qquad (h = \overline{H_1 + 1, H}) \tag{2}$$

where u_h^* – allowable value of residence time in the application class h; δ_h^* – allowable probability of exceeding a specified limit u_h^*; $\Pr\left(\tau_{u_h} > u_h^*\right)$ – probability that the residence time in the application class h (h-request) τ_{u_h} exceeds the allowable value u_h^*; $u_h = M[\tau_{u_h}]$ – average value (expectation) of residence time of h-request in system.

Resident times $\tau_{u_1}, \ldots, \tau_{u_H}$ depend on the system performance V and service discipline. Obviously, for any restrictions to (1) and (2) can be performed at the expense of the system performance. At the same time the best solution will be the discipline which can perform with minimal restrictions performance.

The synthesis problem is solved service discipline in the classroom disciplines with mixed priorities, which are described by a matrix of priorities $Q = [q_{ij} \ (i, j = 1, \ldots, H)]$, where q_{ij} set the priority of i-requests against j-requests: 0 – no priority, 1 –relative priority (RP) 2 – absolute priority (AP) [1].

Thus, the problem of synthesis to systems with constraints is formulated as follows: disciplines to find the discipline of service requests with mixed priority, which limits (1) and (2) are performed with minimal performance V.

3. Calculated identities

Resident time of h-requests ($h = \overline{1, H}$) in system τ_{u_h} in common consists of the waiting time of service start τ_{x_h} and processing time τ_{z_h}, including the waiting time in the interrupted state: $\tau_{u_h} = \tau_{x_h} + \tau_{z_h}$. Then the expectation u_h and the second moment $u_h^{(2)}$ of resident time of h-requests:

$$u_h = x_h + z_h; u_h^{(2)} = x_h^{(2)} + 2x_h z_h + z_h^{(2)}. \tag{3}$$

Values x_h, z_h and $x_h^{(2)}$, $z_h^{(2)}$ are counted [1]:

$$\left.\begin{array}{l} x_h^{!} = \frac{\sum_{i=1}^{H} r_4(i,k)\,\lambda_i\, b_i^{(2)}}{2(1-R_h^{(2)})\,(1-R_h^{(3)})}; \qquad z_h^{!} = \frac{b_h}{1-R_h^{(1)}}; \\ x_h^{(2)} = \frac{\sum_{i=1}^{H} r_4(i,k)\,\lambda_i\, b_i^{(3)}}{3(1-R_h^{(2)})^2\,(1-R_h^{(3)})} + \frac{\sum_{i=1}^{H} r_3(i,k)\,\lambda_i\, b_i^{(2)}\ \sum_{i=1}^{H} r_4(i,k)\,\lambda_i\, b_i^{(2)}}{2(1-R_h^{(2)})^2\,(1-R_h^{(3)})^2} + \\ + \frac{\sum_{i=1}^{H} r_2(i,k)\,\lambda_i\, b_i^{(2)}\ \sum_{i=1}^{H} r_4(i,k)\,\lambda_i\, b_i^{(2)}}{2(1-R_h^{(2)})^3\,(1-R_h^{(3)})}; \\ z_h^{(2)} = \frac{b_h^{(2)}}{(1-R_h^{(1)})^2} + \frac{\sum_{i=1}^{H} r_1(i,k)\,\lambda_i\, b_i^{(2)}}{(1-R_h^{(1)})^3}, \end{array}\right\} \quad (4)$$

where $b_i^{(n)} = \theta_i^{(n)}/V^n$ – n starting moment of service time of i-request ($i = \overline{1,H}$;$n = 1, 2, 3$); V – system capacity; $R_h^{(g)} = \sum_{i=1}^{H} r_g(i,h)\,\lambda_i b_i$ – partial total loads; $r_g(i,h)$ – coefficients with values 0 or 1, depending on the values of the elements q_{ih} and q_{hi} of priority matrix and allow applications to allocate classes i and h , with same priority class: $r_1(i,h) = 0,5\,q_{ih}(q_{ih}-1)$ – takes the value 1, if i-requests have absolute priority against h-requests; $r_2(i,h) = 0,5\,q_{ih}(3-q_{ih})$ – takes the value 1, if i-requests have relative priority or absolute priority against h-requests; $r_3(i,h) = 1-0,5\,q_{hi}(3-2q_{ih}+q_{hi})$ – takes the value 1, if i-requests have no priority or relative priority or absolute priority against h-requests; $r_4(i,h) = 1+0,5\,q_{hi}(1-q_{hi}+q_{ih})$ – takes the value 0 only if h-requests have absolute priority against i-requests.

4. The algorithm synthesis service discipline

The problem of synthesis of service discipline is reduced to the determination of values q_{ij} of the priority matrix, at which the following restrictions (1) for all classes of applications. The analytical solution of the system of inequalities (1) is not possible because the number of different joint venture to even a small number of classes of applications is significant. So, when $H = 5$ there are more than 4500 correct disciplines and when $H = 10$ – over 100 millions. In addition, the formation of priority matrix you must abide by the rules, allowing to build the so-called correct before [1]. A sequential scan of all possible matrixes leads to time-consuming.

In this case it is necessary to develop an effective and well-formalized synthesis algorithm for priority service disciplines based on a heuristic approach based on targeted iterating multiple disciplines of service with mixed priorities in general form, which allows to automate the process of setting priorities between different classes of requests.

The problem of synthesis of service disciplines is decided in two stages.

In the first stage, the problem reduces to the problem of the distribution of priorities for systems with medium restrictions by converting probabilistic constraints (1) in the middle limits.

To do this, considering the limitation (1) at the border, we write it as follows: $1 - U_h(u_h^*) = \delta_h^*$, where $U_h(\tau)$ – residence time distribution function τ_{u_h} in system of h-requests. Solving the resulting equation for the expectation, we

find the estimate of restrictions on the mean residence time equivalent probabilistic limitation, as a function of the allowable of residence time u_h^* and the probability of exceeding it δ_h^*: $\tilde{u}_h = \varphi(u_h^*, \delta_h^*)$. Now probabilistic constraints (1) may be replaced by averages limitations:

$$u_h \le \tilde{u}_h, \qquad (h = \overline{1, H_1}). \tag{5}$$

In implementing the described approach, despite its simplicity, a problem arises in that the distribution function $U_h(\tau)$ depends on a service discipline which during synthesis are constantly changing. For evaluation of $\tilde{u}_h$ we assume that the distribution function $U_h(\tau)$ is exponential. Then $1 - U_h(u_h^*) = \mathrm{e}^{-u_h^*/\tilde{u}_h} = \delta_h^*$, where we get: $\tilde{u}_h = -u_h^*/\ln \delta_h^*$. Suppose that δ_h^* is given as $\delta_h^* = 10^{-n}$, then $\tilde{u}_h = 0,435\ u_h^*/n \quad (h = \overline{1, H_1})$. As a result of this conversion problem of synthesis service discipline combined with restrictions reduced to the problem with average limitations include the following items.

1 As the initial service discipline assigned no priority service discipline Q, for which the value of performance $V_{\max}^Q$ is determined, which is necessary for performing the specified constraints (5) and (2): $V_{\max}^Q = \max\{V_1^Q, \ldots, V_H^Q\}$ where V_h^Q – minimum value of performance, where $u_h = \tilde{u}_h$ for $h = \overline{1, H_1}$ and $u_h = u_h^*$ for $h = \overline{H_1 + 1, H}$ with service discipline with mixed priority Q.

2. For service discipline Q we rearrange class numbers $h_1, \ldots, h_H$ by counted performance values in descending order: $V_{h_1}^Q \ge \ldots \ge V_{h_H}^Q$, $h := h_H$.

3. For h_1 and h classes we try to change the priority by increasing the priority of h_1 class against h taking into account the requirements of discipline correctness.

4. If the change succeeded, then for new service discipline $Q^{'}$ the value of the performance is determined as $V_{\max}^{Q^{'}} = \max\{V_1^{Q^{'}}, \ldots, V_H^{Q^{'}}\}$. If $V_{\max}^{Q^{'}} < V_{\max}^Q$, then $Q := Q^{'}$, $V_{\max}^Q := V_{\max}^{Q^{'}}$ and we continue with item 2.

5. If we cant change the priority between h_1 and h or for $Q^{'}$ performance is $V_{\max}^{Q^{'}} \ge V_{\max}^Q$, then we need repeat items 3-4 for h=h_{H-1},h_{H-2}, …,h_2.

6. If the search for all combinations of pairs of classes for this scheme did not lead to a change in the matrix of priorities, the process will stop the search service discipline and as a result received last priority matrix Q.

On the second stage its nessesary to check the implementation of probabilistic constraints(1), because the distribution function $U_h(\tau)$ may differs from exponential. These restrictions can not be fulfilled, if the coefficient of variation of of residence time $\alpha_h = \sqrt{u_h^{(2)} - u_h^2} \Big/ u_h$ is more than 1. In this case, to clarify the evaluation $\tilde{u}_h$ we perform the approximation $U_h(\tau)$ by two moments u_h and $u_h^{(2)}$, calculated using the formula (3) and (4). As we use the two-phase distribution approximating representation hyperexponential distribution [2], for that we count $q_h \le 2/(1 + \alpha_h^2)$ and expectations exponential phases $t_h^{'} = [1 + \sqrt{(1 - q_h)(\alpha_h^2 - 1)/(2\, q_h)}]\, u_h$ and $t_h^{"} = \{1 - \sqrt{q_h(\alpha_h^2 - 1)/[2\,(1 - q_h)]}\}\, u_h$.

We assume that the parameter distribution hyperexponential is $q_h = 2/(1+\alpha_h^2)$. Then, by analogy with the exponential distribution we obtain a new estimate of the average limit:

$$\tilde{u}_h = -\frac{2\,u^*}{(1+\alpha_h^2)\ln\{\delta_h^*(1+\alpha_h^2)/2\}}.$$

After assuming changes in (5) for h-class requests of $\tilde{u}_h$, obtained under the assumption of exponential distribution, a new estimate obtained for hyperexponential distribution, you must rerun the items 2-6 of the first stage of the service discipline synthesis problem.

5. Conclusion

The proposed approach to the problem of the synthesis of service discipline with mixed priorities for systems with complex limits on the residence times in the different classes of applications allows for a specified quality of the system with a minimum performance.

REFERENCES

1. Aliev T.I., Maharevs E. Queueing disciplines based on priority matrix // Scientific and Technical Journal of Information Technologies, Mechanics and Optics. 2014, №6 (94). – P. 91-97.
2. Aliev T.I. Approximation of probability distributions in queuing systems // Scientific and Technical Journal of Information Technologies, Mechanics and Optics. 2013, №2 (84). – P. 88-93.

ИТЕРАЦИОННЫЙ АЛГОРИТМ ПРОЕКТИРОВАНИЯ ШИРОКОПОЛОСНЫХ БЕСПРОВОДНЫХ СЕТЕЙ ВДОЛЬ ПРОТЯЖЕННЫХ ТРАНСПОРТНЫХ МАГИСТРАЛЕЙ[1]

С.Н. Васильев[1], *В.М. Вишневский*[2]

Институт проблем управления им. В.А. Трапезникова РАН, Москва

[1]snv@ipu.ru, [2]vishn@inbox.ru

Аннотация

Рассмотрен новый итерационный алгоритм проектирования архитектуры широкополосных беспроводных сетей вдоль протяженных транспортных магистралей. Сформулированы и решены математические задачи каждого этапа итерационного алгоритма. Разработанный итерационный метод проектирования эффективно применялся для проектирования и реализации крупномасштабного проекта создания беспроводных сетей вдоль автодорог Республики Татарстан (Россия).

1. Введение

Создание современной инфраструктуры передачи мультимедийной информации (голос, данные, видео) вдоль протяженных магистралей является одной из важнейших проблем при разработке новых и функционировании существующих транспортных магистралей. Особенно актуально решение этой проблемы для стран с протяженными автомобильными и железнодорожными магистралями. Создание такой инфраструктуры связи позволяет: обеспечивать оперативный контроль за техническими параметрами трассы путем высокоскоростной передачи информации с датчиков и сенсоров в центр управления; обеспечение контроля безопасности за участками трассы и стратегически важными объектами с использованием информации систем видеонаблюдения; обеспечение голосовой связи (IP-телефония) и передачи мультимедийной информации между стационарными и мобильными объектами на протяженных магистралях, а также связь с Центром управления и т.д.

Учитывая высокие требования к безопасности, использование сетей общего пользования (типа Интернет) в системах связи вдоль протяженных трасс обычно не допускается, тем более, что протяженные магистрали часто проходят по малонаселенным, труднодоступным территориям, где

[1]Работа выполнена при финансовой поддержке Министерства Образования и Науки РФ в рамках проведения прикладных научных исследований №14.613.21.0020 от 22.10.2014 (RFMEFI61314X0020).

доступ к сети Интернет или сотовой связи отсутствует. Создание выделенных оптоволоконных сетей вдоль протяженных магистралей или радиорелейных линий требует огромных материальных затрат. То же касается и использования спутниковых каналов и сетей связи. В то же время стоимость широкополосной высокоскоростной беспроводной связи на аппаратно-программных средствах, реализующих международный стандарт IEEE 802.11-2012, на порядок ниже. Стандарт регламентирует создание высокоскоростных каналов связи и беспроводных сетей, функционирующих под управлением протоколов IEEE 802.11n и IEEE 802.11s, на базе которых могут эффективно реализовываться беспроводные сети с линейной топологией вдоль протяженных транспортных магистралей. Указанные сети обеспечивают создание не только магистральной скоростной передачи мультимедийной информации путем расположения базовых станций на высотных зданиях и вышках вдоль транспортных магистралей, но и оперативную связь со стационарными и мобильными абонентами (автомобили, железнодорожные поезда, дорожные знаки, пункты весового контроля и контроля ПДД и т.д.).

Развертывание и развитие сетей беспроводной связи вдоль протяженных магистралей требует решения ряда научных и сложных организационно-технических задач в условиях жестких ограничений на использование частотных, экономических и аппаратных ресурсов. В связи с этим возрастает актуальность решения проблемы оптимального размещения базовых станций вдоль транспортных магистралей, которая является одной из важнейших при проектировании широкополосных беспроводных сетей этого класса. Ее решение направлено как на реализацию высокоскоростной магистральной сети, так и максимальное телекоммуникационное покрытие трассы с целью обеспечения подключения мобильных пользователей, а также минимизацию интерференции и временных задержек при передаче мультимедийной информации по сети. Исследованию проблемы проектирования сетей с линейной топологией посвящены многочисленные публикации [1–5]. Однако в большинстве этих работ решаются частные задачи синтеза архитектуры сетей с линейной топологией. В значительной мере это связано с высокой сложностью и размерностью решаемой задачи и соответствующими вычислительными трудностями.

С учетом такой специфики этой задачи и в отличие от известных работ, в настоящем проекте предлагается новый итерационный подход поэтапного учета всего комплекса критериев качества и ограничений для проектируемой беспроводной сети. Оптимизация проекта сети, помимо архитектурных характеристик, должна учитывать: ограничения на параметры частотно-территориального плана вдоль протяженной магистрали; возможность реализации базовых станций на различных типах аппаратно-программных средств, отличающихся дальностью радиорелейной связи для взаимодействия с соседними станциями и областью телекоммуникационного покрытия трассы для подключения мобильных пользователей; возможную ин-

терференцию и помехи при трансляции мультимедийной информации; ограничения на надежность, включая автоматическую реконфигурируемость сети при отказах компонент для временного поддержания связи хотя бы с потерей ее качества; ограничения на время трансляции пакетов, суммарную стоимость сети и т.д.

На каждом этапе итерационного процесса проектирования сети возникает и решается ряд новых, не исследованных в мировой литературе задач, включая: задачи оптимального размещения базовых станций, сформулированную в терминах нелинейного математического программирования; задачи исследования характеристик стохастической сети массового обслуживания для оценки времени трансляции пакетов, надежности и других характеристик производительности беспроводной сети с линейной топологией. В результате каждого поэтапного решения задач формируются наборы альтернативных решений, позволяющие синтезировать комплексные решения для полной задачи проектирования сети и многокритериального выбора наиболее предпочтительного проекта [6, 7]. Разработан также комплекс имитационных моделей для уточнения сравнительного анализа и выбора типа сетевого протокола, под управлением которого функционирует беспроводная сеть.

Предлагаемый итерационный подход и пакет прикладных программ позволяют повысить качество и сократить время оптимального проектирования беспроводных сетей вдоль протяженных транспортных магистралей. Указанный подход эффективно применялся при создании пилотного проекта широкополосной беспроводной сети вдоль автодорог Республики Татарстан (Россия), первая очередь которого (сеть вдоль окружной дороги г. Казань) разработана и реализована под руководством и при участии авторов настоящей статьи.

2. Итерационный алгоритм синтеза архитектуры беспроводных сетей с линейной топологией

Новый итерационный подход проектирования беспроводных сетей с линейной топологией позволяет учитывать весь комплекс ограничений, необходимых для практической реализации и выбора оптимальной архитектуры беспроводной сети. На первом этапе сходящегося итерационного процесса исследуется задача нелинейного программирования оптимального размещения базовых станций дискретном варианте. В рамках решения этой задачи определяется как топология высокоскоростной магистральной сети, так и архитектура клиентской сети для подключения мобильных пользователей. Определяется также оптимальный состав аппаратно-программных средств комплектации каждой базовой станции из ограниченного набора типов возможного оборудования, отличающихся характеристиками мощности излучения (дальности действия), производительности, надежности, стоимости и т.д. В качестве ограничений при решении этой задачи наряду с ограничениями возможностей радиорелейной аппаратуры используются

характеристики частотно-территориального плана, включая координаты возможных мест расположения базовых станций, карту местности вдоль протяженной трассы и т.д. В тоже время, учитывая многокритериальность и высокую размерность решаемой задачи, на первом этапе не используются важные показатели функционирования сети как: среднее время доставки пакетов; надежность - стационарная вероятность безотказной работы сети в целом и отдельных компонент; параметры сетевых протоколов и т.д., расчет каждого из которых представляет сложную в теоретическом и вычислительном планах задачу, которые решаются на следующих этапах проектирования беспроводной сети. На втором этапе осуществляется расчет надежности сети, топология которой была определена на первом этапе. Разработана теоретико-вероятностная модель для оценки стационарной вероятности безотказной работы сети. Если в результате расчетов надежность сети удовлетворяет требованиям технического задания (ТЗ), то осуществляется переход к следующему этапу проектирования. В противном случае решается задача повышения надежности за счет резервирования отдельных участков сети и соответствующей возможности автоматической реконфигурации сети при отказе отдельных компонент [8] с целью достижения требований ТЗ. Для оценки времени задержки пакетов на третьем этапе необходима формулировка и решение новой, слабо исследованной в мировой литературе, задачи оценки характеристик открытой сети массового обслуживания с коррелированными входными потоками и матрицей переходных вероятностей, определяемой топологией беспроводной сети, полученной на предыдущем этапе. В работе рассмотрен частный случай такой открытой сети - многофазная система массового обслуживания типа $MAP/PH/1/N \rightarrow \ldots \rightarrow MAP/PH/1/N$. Если рассчитанные характеристики производительности (в частности, время задержки) удовлетворяют требованиям ТЗ, то осуществляется переход к следующему этапу. В противном случае осуществляется возврат к первому этапу проектирования. При этом в рамках решения задачи синтеза топологии выбираются лишь аппаратно-программные средства комплектации базовых станций, обладающие более высокими характеристиками производительности по сравнению с первой итерацией. Указанный итерационный процесс продолжается до полного удовлетворения требованиям ТЗ. На последнем этапе разрабатывается комплекс имитационных моделей, в которых учитываются детальные характеристики различных сетевых протоколов, под управлением которых может быть реализована беспроводная сеть. Проводится их сравнительный анализ и окончательный выбор сетевого протокола (например, IEEE 802.11n или MESH-сеть). Пакет прикладных программ, реализующий предлагаемый подход, обеспечивает повышение качества и ускорение процесса проектирования беспроводных сетей вдоль протяженных транспортных магистралей.

3. Формулировка математических задач этапов итерационного алгоритма

Как уже отмечалось в предыдущем разделе, рассматриваемый итерационный подход позволяет учитывать весь комплекс ограничений, необходимых для практической реализации и выбора оптимальной архитектуры беспроводной сети. На первом этапе алгоритма формируется и решается задача нелинейного программирования оптимального выбора мест размещения базовых станций беспроводной сети и типа оборудования, на базе которого реализуется каждая базовая станция. Предполагается, что известны координаты $Z_i\,(i = \overline{1, M})$ высотных зданий и вышек вдоль транспортной магистрали, количество возможных типов оборудования N, отличающихся шириной области телекоммуникационного покрытия $D_k\,(K = \overline{1, N})$ транспортной магистрали (углом раскрытия антенн), стоимостью $C_k\,(K = \overline{1, N})$, производительностью и надежностью, уровнем затухания сигнала (дальностью действия $R_k\,(K = \overline{1, N})$.

Введем переменные

$$X_{ij} = \begin{cases} 1, & \text{если на } i\text{-месте возможного расположения установлена} \\ & \quad \text{базовая станция, оборудованная } j\text{-м типом аппаратуры.} \\ 0, & \text{в противном случае.} \end{cases}$$

Тогда задача состоит в минимизации суммарных затрат на создание сети

$$\sum_{i=1}^{M}\sum_{j=1}^{N} C_{ij}X_{ij} \to \min,$$

при ограничениях на суммарную область телекоммуникационного покрытия транспортной магистрали и связность беспроводной сети. Для формулировки ограничений на связность введем булевские переменные

$$U_{ij} = \begin{cases} 1, & \text{если } i\text{-я станция связана } j\text{-ой,} \\ 0, & \text{в противном случае,} \end{cases}$$

$$Y_i = \begin{cases} 1, & \text{если в } i\text{-ой точке установлена базовая станция,} \\ 0, & \text{в противном случае.} \end{cases}$$

Очевидно, что

$$y_i = \sum_{j=1}^{N} X_{ij},$$

Ограничение на связность сети представляется в виде

$$U_{ij}(Z_i y_i - Z_j y_j) \le R, \text{ где}$$

$$U_{ij} = U_{ji}, \quad \sum_{j=1}^{i-1} U_{ij} = y_i, \quad \sum_{j=i+1}^{M} U_{ij} = y_i.$$

Ограничение на суммарное телекоммуникационное покрытие магистрали представляется в виде

$$\sum_{i=1}^{M} D_{ij} X_{ij} \geq D.$$

Учитывая, что нелинейные переменные в виде произведения булевых переменных легко преобразуются к линейному виду, для решения сформулированной задачи может быть использован метод ветвей и границ, программно-реализованный, например, в пакете программ GLPK (GNV Linear Programing Kit). С использованием кроссплатформенного пакета GLPK удается получать точные решения для сети с небольшим количеством базовых станций. Для сетей большой размерности разработан эвристический алгоритм.

В постановке и решении предыдущей задачи отсутствует ограничение на один из важнейших показателей производительности беспроводной сети — время доставки пакетов. Расчет этого параметра представляет собой сложную, слабо исследованную в мировой литературе задачу — оценку характеристик многофазной стохастической системы с входящим MAP-потоком (Marcovian Arrival Process) и кросс-трафиком, поступающим от мобильных абонентов. Для решения этой задачи используется оригинальный подход, предложенный в [9-11]. Указанный подход базируется на теореме о том, что выходной поток системы массового обслуживания $MAP/PH/1/$ (обозначения Кендалла), адекватно моделирующей базовую станцию беспроводной сети, является также MAP-потоком.

На третьем этапе итерационного процесса проверяется ограничение на надежность беспроводной сети. Обозначим через P_i — стационарную вероятность безотказной работы i-го ($i = \overline{1, M}$) базовой станции и исходящего беспроводного канала связи. Надежность сети должна быть не ниже заданной, т.е.

$$\prod_{i=1}^{M} P_i \geq P.$$

Если это ограничение не выполняется, необходимо решение следующей оптимизационной задачи.

Обозначим через P_{ij} — стационарную вероятность безотказной работы аппаратуры j-го типа ($j = \overline{1, N}$) базовой станции I ($i = \overline{1, M}$). Полагая, что каждая базовая станция сети может быть реализована только на одном типе аппаратуры (без резервирования), задача максимизации надежности

сети может быть сформулирована в виде

$$\prod_{i=1}^{M}\sum_{j=1}^{N} P_{ij}X_{ij} \to \max,$$

при ограничениях на суммарную стоимость сети

$$\sum_{i=1}^{M}\sum_{j=1}^{N} C_{ij}X_{ij} \le C.$$

Решение этой задачи целочисленного линейного программирования осуществляется с использованием существующих пакетов прикладных программ. Отметим, что при необходимости повышения надежности беспроводной сети путем резервирования отдельных каналов связи или участков сети сформулированная задача значительно усложняется в связи с необходимостью исследования соответствующего многомерного альтернирующего марковского процесса и не рассматривается в настоящей статье.

На последнем этапе итерационного процесса проектирования разработаны детальные имитационные модели беспроводных сетей функционирующих под управлением протоколов IEEE 802.11n и MESH, для проведения сравнительного анализа сетевых протоколов и окончательного выбора архитектуры беспроводной сети с линейной топологией.

ЛИТЕРАТУРА

1. M.B. Brahim, W. Drira, F Filali. Roadside units placement within city-scaled area in vehicular ad-hoc networks // 3rd International Conference on Connected Vehicles and Expo (ICCVE 2014). — Vienna, Austria: 3–7 Nov 2014.
2. 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 — 2014. — Vol. 7. — P. 91–108.
3. Reis A.B., Sargento S., Neves F., Tonguz O.K. Deploying Roadside Units in Sparse Vehicular Networks: What Really Works and What Does Not // IEEE Transactions on Vehicular Technology. 2014. V. 63. P. 2794–2806.
4. J. H. Lee, C.M. Kim. A Roadside Unit Placement Scheme for Vehicular Telematics Networks // Lecture Notes in Computer Science. 2010. V. 6059. P. 196–202.
5. T.-J. Wu, W. Liao, C.-J. Chang. A Cost-Effective Strategy for Road-Side Unit Placement in Vehicular Networks // IEEE Transactions on Communications. 2012. V. 60. P. 2295 – 2303.
6. Васильев С.Н., Котлов Ю.В. Методы и алгоритмы многокритериальной оптимизации на основе нестрогих ранжировок альтернатив по частным критериям и опыт компьютерной реализации // Проблемы управления и информатики, Киев, 2006, №1-2, с. 28–38.

7. Васильев С.Н., Батурин В.А., Баянова Т.О. Многокритериальное принятие решений, основанное на получении оценочной функции в виде полинома третьего порядка // Управление большими системами, 2008, вып. 22, с. 5–20.
8. Васильев С.Н., Жерлов А.К., Федосов Е.А., Федунов Б.Е. Интеллектное управление динамическими системами. — М.: Физматлит, 2000. — 352 с.
9. Klimenok V., Dudin A., Vishnevsky V. On the stationary distribution of tandem queue consisting of a finite number of stations // Communications in Computer and Information Science. 2012, V.291. P. 383–392. Springer, Heidelberg.
10. Vishnevsky V.M., Semenova O.V., Dudin A.N., Klimenok V.I. Approximate Method to Study M/G/1-Type Polling System // Quality Technology & Quantitative Management (QTQM). 2012. Vol.9, №2. P. 211–228.
11. Klimenok V., Dudin A. Vishnevsky V. Tandem queueing system with correlated input and cross-traffic // Communications in Computer and Information Science. 2013. CCIS 370. P. 416–425. Springer, Heidelberg.

AN APPROACH OF DESIGNING VOICE MAN-MACHINE INTERFACE

Abramenkov A.N., Farkhadov M.P., Petukhova N.V., Vaskovsky S.V.
Institute of Control Sciences of Russian Academy of Sciences, Moscow, Russia

Abstract

In this report the problems of voice interface are investigated. We propose some methods to design man-machine interface. The methods utilize voice recognition technology. We specify rules how to design effective interface. The rules are based on our experience in implementing voice application.

Keywords: speech recognition, speech interface, man-machine interface design, dialogue management.

ОБ ОДНОМ ПОДХОДЕ ПРОЕКТИРОВАНИЯ РЕЧЕВОГО ЧЕЛОВЕКО-МАШИННОГО ИНТЕРФЕЙСА

А.А. Абраменков, М.П. Фархадов, Н.В. Петухова , С.В. Васьковский
Институт проблем управления им. В.А. Трапезникова РАН, Москва
aabramenkov@asmon.ru, mais@ipu.ru, nvpet@ipu.ru, v63v@yandex.ru

Аннотация

Работа посвящена исследованию проблематики речевых интерфейсов, предложены методы решения возникающих проблем при проектировании человеко-машинного интерфейса с применением речевых технологий, сформулированы правила создания эффективных интерфейсов, основанные на проведенных исследованиях и опыте реализации речевых приложений.

Ключевые слова: распознавание речи, речевой интерфейс, проектирование человеко-машинного интерфейса, управление диалогом.

1. Введение

Огромную роль в современном обществе играют новейшие информационные технологии. Поэтому при построении и реализации информационных систем необходимо использовать последние достижения в этой области [1–5]. Важную роль во взаимодействии человека с информационной системой может играть речевой интерфейс. Несмотря на большое количество работ по проектированию речевого интерфейса [6–16] в них мало внимания уделяется реализации результатов исследований в этой области.

Это объясняется отсутствием легко измеряемых критериев, сильной контекстной зависимостью решений, плохой формализуемостью задач, а так же значительной трудоемкостью этих работ. Ряд факторов определяют привлекательность речи для создания интерфейса. Так как речь является наиболее естественным средством коммуникации, она может быть очень эффективным способом взаимодействия. Однако при разработке речевого интерфейса следует учитывать свойства речи, которые будут затруднять ее применение в приложениях. К таким свойствам можно отнести ограниченность слуховой памяти человека, что не позволяет пользователям воспринимать слишком длинные и информационно насыщенные сообщения машины. Необходимо обратить внимание, что речь есть достаточно медленная среда передачи информации, и в сочетании с ограниченностью слуховой памяти удлиняется диалог с системой. Речь часто содержит слова и звуки, «засоряющие» диалог, например, покашливания, междометия, вводные слова, повторы и т.п. Как правило, человек склонен распространять при общении с машиной свои навыки межличностного общения, и поэтому разработчик интерфейса должен хорошо представлять себе, что именно из привычного людям стиля диалога может и должно быть реализовано в человеко-машинном диалоге. Следует отметить, что при нынешнем состоянии дел в области распознавания речи многие элементы естественного речевого интерфейса реализованы быть не могут. Рассмотрев некоторые особенности речи человека, перейдем к реализации проектирования человеко-машинных интерфейсов.

2. Общие принципы проектирования человеко-машинных интерфейсов

Остановимся на общих принципах проектирования человеко-машинных интерфейсов любого типа [6]. Необходимо помнить, что с точки зрения потребителя, именно интерфейс является конечным продуктом и олицетворяет собой всю систему, и его необходимо проектировать уже на ранних стадиях разработки информационных комплексов. При этом интерфейс должен быть ориентирован на человека, т.е. отвечать его нуждам. При разработке интерфейса необходимо сначала учесть общие факторы, свойственные всем людям, а потом уже переходить к учету факторов индивидуальных. Речевой интерфейс, подобно графическому или любому другому человеко-машинному интерфейсу, должен выполнять следующие основные функции: предоставление клиенту исходной информации о сервисах, ввода данных со стороны клиента, вывода информации для клиента, а так же исправление ошибок. Последняя функция является принципиально важной для речевого интерфейса, поскольку его отличает от других типов интерфейсов, и в частности от графического, то обстоятельство, что источником ошибок при вводе исходных данных являются не только возможные ошибки человека, но и ошибки распознавателя. Поэтому более подробно остановимся на некоторых особенностях реализации речевого интерфей-

са. Остановимся на общих принципах проектирования человеко-машинных интерфейсов любого типа [6]. Необходимо помнить, что с точки зрения потребителя, именно интерфейс является конечным продуктом и олицетворяет собой всю систему, и его необходимо проектировать уже на ранних стадиях разработки информационных комплексов. При этом интерфейс должен быть ориентирован на человека, т.е. отвечать его нуждам. При разработке интерфейса необходимо сначала учесть общие факторы, свойственные всем людям, а потом уже переходить к учету факторов индивидуальных. Речевой интерфейс, подобно графическому или любому другому человеко-машинному интерфейсу, должен выполнять следующие основные функции: предоставление клиенту исходной информации о сервисах, ввода данных со стороны клиента, вывода информации для клиента, а так же исправление ошибок. Последняя функция является принципиально важной для речевого интерфейса, поскольку его отличает от других типов интерфейсов, и в частности от графического, то обстоятельство, что источником ошибок при вводе исходных данных являются не только возможные ошибки человека, но и ошибки распознавателя. Поэтому более подробно остановимся на некоторых особенностях реализации речевого интерфейса.

3. Когнитивные особенности проектирования эффективного речевого интерфейса

Главной отличительной чертой речевого интерфейса является *проблема невидимости*. При этом зрение не участвует в процессе взаимодействия с машинной стороной, и восприятие вопросов и информации производится клиентом только лишь на слух. Аудио информация сообщается последовательно. Следует отметить, что эта информация не сохраняется на экране дисплея или на бумаге, к ней нельзя вернуться немедленно или в любой момент. Исключение зрения из процесса взаимодействия с машиной приводит к значительному возрастанию когнитивной нагрузки на человека и к необходимости учета серьезных ограничений на восприятие информации при конструировании диалога. Под *когнитивной нагрузкой* будем понимать умственную нагрузку, связанную с получением, обработкой, запоминанием информации. Все интерфейсы создают когнитивную нагрузку на пользователя. Пользователь должен понять правила работы с системой, усвоить предлагаемые термины и основные положения, запоминать информацию, поэтому речевые интерфейсы предъявляют к человеку особенно высокие требования. Имеются *три когнитивные составляющие*, которым необходимо уделять особое внимание в процессе дизайна речевого интерфейса.

Нагрузка на память. Ограниченность слуховой памяти требует от разработчика ограничивать количество сообщаемой клиенту информации.

Удержание внимания. Разработчик должен уметь оценивать уровень, на котором на том или ином шаге диалога может находиться внимание пользователя, предусмотреть меры по удержанию его внимания, уметь вос-

становить диалог, если внимание пользователя будет отвлечено внешними событиями.

Понятийная сложность. Создавая интерфейс, разработчик должен представлять себе, насколько сложны будут для пользователей предлагаемые им термины и понятия, и как добиться, чтобы конструируемый диалог согласовывался с опытом потенциальных пользователей.

Разработчику можно предложить следующие *средства для решения задачи построения эффективного и продуктивного речевого интерфейса:* выбор стратегии диалога и управление им; дизайн промптов (текстов, произносимых системой); построение грамматик. Рассмотрим последовательно перечисленные проблемы.

3.1. Снижение нагрузки на память. Нагрузка на память — едва ли не самая существенная проблема для речевого интерфейса. Размер кратковременной слуховой памяти ограничивается запоминанием 4-7 концепций. Следует отметить, что лучше всего запоминаются самые последние слова, которые человек слышит. Известны так же и временные характеристики сохранения информации в кратковременной памяти [17–20]. На основании этих данных разработаны и экспериментально проверены [21] принципы конструирования меню и формулирования промптов (вопросов и реплик машинной стороны) и подсказок, которые позволяют сократить нагрузку на память.

3.2. Удержание внимания. Проблема удержания внимания клиента становится наиболее актуальной на этапе предоставления ему аудио информации, особенно если по запросу клиента из базы данных получен достаточно ее большой объем. Ограниченность кратковременной памяти и потеря внимания имеют следствием высокую вероятность необходимости повтора. Кроме того, с большой вероятностью может оказаться, что далеко не вся эта информация одинаково важна для клиента. Деление выводимой информации на фрагменты смягчает эти проблемы, но требует затрат времени на диалог по поводу вывода этих фрагментов. Оптимальное число фрагментов зависит от контекста конкретного приложения и «делимости» выводимой информации. Также следует помнить, что фрагмент должен включать ограниченное число элементов данных для запоминания. Эта граница определяется способностями человека к запоминанию информации на слух и составляет в среднем 3-4 элемента.

3.3. Снижение понятийной сложности. Понятийная или концептуальная сложность прямо пропорциональна сложности усвоения пользователем услышанных им понятий и определений и сложности самих этих понятий и определений. Сложность усвоения определяется способностями человека к пониманию и обучению. Были предприняты исследования, позволившие дать ряд рекомендаций, которые будут способствовать снижению нагрузки на пользователей с точки зрения усвоения ими концептуальных понятий. К этим рекомендациям можно отнести *использование*

универсальных команд навигации, всегда доступных пользователю вне зависимости от приложения и контекста. Ряд организаций по разработке стандартов рассматривали вопрос об универсальных командах, и к настоящему времени составлены рекомендации по этому вопросу [22]. Исследования применительно к российским информационным системам позволили создать свой список рекомендуемых универсальных команд, который в основном совпадает с рекомендациями европейских комитетов по телекоммуникациям, но несколько короче и включает 5 команд [23]. На рис. 1 показана относительная частота использования универсальных команд. Из рисунка следует, что команды «конец» и «до свидания» практически не использовались, поэтому список универсальных команд, обязательных для отработки на любом шаге диалога, рекомендуется ограничить командами «оператор», «помощь», «повторить», «назад», «начало».

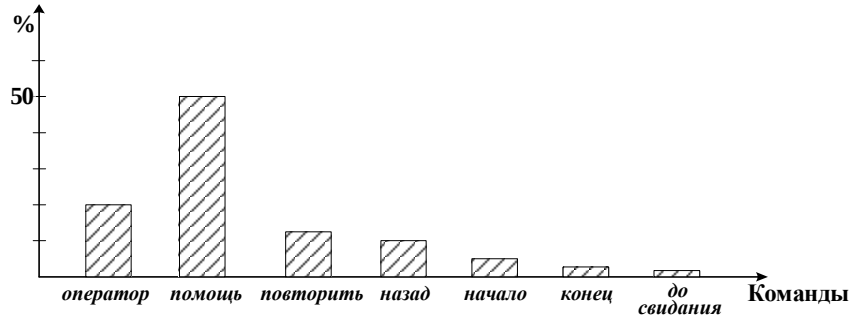


Рис. 1: Относительная частота использования универсальных команд.

При построении человеко-машинного интерфейса важно реализовывать *унификацию действий и терминологию*. Идея состоит в том, чтобы позволить клиенту выполнять похожие действия похожим образом, а также обеспечить на всем протяжении диалога постоянство терминологии. Экспериментально установлено, что на каждом шаге диалога пользователь укрепляет уже полученные знания о системе и обновляет свою ментальную модель системы, которая включает его знания о функциях системы, доступных командах, способах произнесения своих ответов на вопросы системы и т.д. Ментальная модель формируется быстрее и получается гораздо более ясной и понятной для пользователя, если он поймет, что похожие действия выполняются по одной и той же схеме. Однако на этом пути есть и свои опасности, главная из которых — монотонность, которая приведет к снижению внимания и ошибкам, а также может раздражать пользователя. При реализации необходимо так же приводить *пояснения и примеры*. Психологические исследования показывают [23], что человек легче понимает задачу и запоминает информацию, если пояснить ему, что представляет собой система, почему выполняются те или иные действия и дать при-

мер ответа. Использование таких пояснений и примеров помогает клиентам правильно формулировать ответы на вопросы системы, что приведет к снижению числа ошибок, значительно улучшит удовлетворенность пользователей и повысит эффективность системы. Были проведены исследования действующих разработок с целью выявить влияние примеров на правильность формулировки клиентами ответов на вопросы системы [21–23]. На рис. 2. показано влияние использования примера на число ошибок при вводе даты подачи машины в системе заказа такси. Когда вместо вопроса «Назовите дату подачи такси» стал использоваться промпт «Назовите дату подачи такси: например сегодня, завтра, послезавтра или пятнадцатого мая», число ошибок значительно сократилось.

Рис. 2: Влияние примера на снижение числа ошибок.

Таким образом, рассмотрев общие принципы и особенности проектирования речевого интерфейса можно рекомендовать использовать полученные результаты для успешного создания модуля речевого взаимодействия с системой электронного государства [24, 25] и разработки модуля доступа к услугам корпоративных информационных сетей [26]. Создание совершенного человеко-машинного интерфейса позволит предоставлять услуги из различных подсистем и организовывать диалоги с серверами самообслуживания, сократить число операторов, а также снизит расходы на организацию инфраструктуры. Однако в процессе реализации современного человеко-машинного интерфейса приходится решать трудоемкие задачи программирования и отладки распределенных иерархических систем реального времени [27].

4. Лингвистические особенности роль лингвистических знаний в создании речевых систем

При проектировании эффективного речевого интерфейса требуется качественная система распознавания речи, для создания которой необходи-

мо учитывать лингвистические знания. Лингвистические понятия играют важную функцию в реализации речевых систем и имеют иерархическую структуру по следующим уровням: фонетический, фонологический, морфологический, лексический, синтаксический и семантический. Изучение лингвистической структуры речи, выделение из речи отдельных лингвистических элементов, таких как фонемы, морфемы, слоги, слова, предложения имеют непосредственное отношение к исследованию систем распознавания речи.

4.1. Назначение лингвистического процессора. Основная задача лингвистического процессора (ЛП) в системах распознавания речи — обеспечивать перевод текста из орфографической формы записи в фонематическую [28]. Среди вспомогательных функций ЛП можно назвать нормализацию текста и автоматическую генерацию форм слова. С точки зрения решаемых при помощи ЛП задач в системе анализа неструктурированной речевой информации можно выделить два основных направления:

- транскрибирование текстов (под транскрибированием понимается перевод текстов из последовательности букв, специальных символов и цифр в последовательность графических символов, обозначающих фонемы — минимальные звуковые единицы языка, из которых строятся словоформы), входящих в базы данных, используемые при акустическом моделировании и обучении верификатора;
- транскрибирование слов и словосочетаний, вводимых пользователем при поиске ключевых слов в речевых данных.

4.2. Существующие подходы к анализу естественного языка. Обработку текста следует рассматривать как подраздел группы методов обработки естественного языка [29]. Данная группа исследований охватывает множество направлений, таких как автоматический перевод текстов с языка на язык, системы понимания текста и ведения диалога, автоматическое распознавание речи, распознавание текста, представленного в виде графического изображения и т.п. В современной науке выделяют два основных подхода к анализу естественного языка: глубокий (основанный на лингвистических знаниях) и поверхностный (основанный на методах машинного обучения). Первая группа методов основывается на создаваемых экспертами правилах и моделях порождения языковых явлений, вторая — на математических и статистических методах, позволяющих автоматически формировать правила с использованием обучающих корпусов лингвистических данных. Представленное выше деление характерно и для лингвистического процессора. В настоящее время наибольшей популярностью пользуются статистические методы [30, 31]. Яркий пример использования статистических алгоритмов при обработке текста — это автоматическое определение частей речи [32]. Данные алгоритмы позволяют построить вероятностные модели на основе больших корпусов данных, содержащих

ручную разметку, проведенную экспертами-лингвистами. Примерами таких корпусов являются Брауновский корпус для английского языка [33] и Национальный корпус русского языка [34]. Благодаря подобным корпусам становится возможным обучение статистических контекстных правил, которые позволяют автоматически определять часть речи слов в зависимости от окружающих их словоформ и при этом принимать не «жесткое» решение, а вероятностное. Наиболее распространенный метод основан на применении динамического программирования [35]. Статистические методы характеризуются высокой скоростью работы и достаточно высоким качеством производимого ими анализа. При этом их качество во многом зависит от объема корпуса, используемого при обучении, а также варьируется от языка к языку. Исторически первыми являются методы анализа, основанного на экспертных правилах, подобные методы восходят к трансформационным (генеративным) грамматикам Н.Хомского [36]. Хомский предлагает анализировать текст путем последовательного выделения в выражении глубинной структуры — взаимосвязанных синтаксических единиц. В результате каждое предложение получает описание в виде дерева, «листьями» которого являются слова, а "ветвями" — элементы грамматики (например, NP (Noun Phrase), именная фраза) и VP (Verb Phrase, глагольная фраза). Ключевое отличие подобных систем от статистических состоит в том, что грамматические структуры создаются экспертами на основе их представления о структуре языка, в то время как статистические методы полностью игнорируют данные знания и основываются лишь на данных, извлекаемых из размеченного корпуса. При этом представительность корпуса и диапазон анализа определяют асимптоту по качеству и универсальности статистических методов. Разрабатываемые современными лингвистами методы глубинного анализа текста снимают эти ограничения, но вносят при этом свои, в частности, вычислительную неэффективность. В настоящий момент интерес к экспертным системам сохраняется [37], а наибольших успехов добились исследователи направления HPSG (грамматическая структура высказывания, основанная на знаниях) [38]. Данный метод структурно включает в себя не только правила грамматики, но и словарь, содержащий помимо собственно лексем, фонологическую, синтаксическую и семантическую информацию.

4.3. Роль лингвистики в модуле. В современных системах с речевыми технологиями роль лингвистического процессора обычно выполняет словарь транскрипций, или лексикон. Лексикон включает список вида (слово; транскрипция). Такой подход позволяет достаточно быстро и эффективно получить транскрипции слов, однако обладает рядом недостатков:

- не позволяет транскрибировать слова, которых нет в словаре, что снижает уровень автоматизации;
- не дает возможности корректно транскрибировать слова в случае омонимии;

- не позволяет учесть контекстное взаимовлияние транскрипций смежных слов (и, соответственно, является причиной ошибок транскрибирования пограничных фонем);
- не предполагает возможности транскрипционного моделирования.

5. Выводы и основные направления развития

Рассмотрев исследования когнитивных и лингвистических аспектов, необходимые для создания речевого интерфейса программ, комфортного для человека, следует отметить, что лингвистический процессор является важным компонентом системы автоматического анализа неструктурированной речевой информации. Редуцированный характер лингвистического процессора в современных системах распознавания и связанные с таким подходом имеет ряд недостатки. В дальнейшем, необходимо исследовать расширенный вариант использования лингвистического процессора, что позволит не только устранить выявленные недостатки, но и существенно упростить процесс подготовки данных для различных модулей системы анализа речи и речевой информации.

ЛИТЕРАТУРА

1. Билик Р.В., Жожикашвили В.А., Петухова Н.В., Фархадов М.П. Анализ речевого интерфейса в интерактивных сервисных системах I. / Автоматика и телемеханика. 2009. є2. С. 80–89.
2. Билик Р.В., Жожикашвили В.А., Петухова Н.В., Фархадов М.П. Анализ речевого интерфейса в интерактивных сервисных системах II. / Автоматика и телемеханика. 2009. є3. С. 97–113.
3. А.Л. Ронжин, А.А. Карпов, "Многомодальные интерфейсы: основные принципы и когнитивные аспекты", Тр. СПИИРАН, 3:1 (2006), 300–319
4. Фархадов М.П. Распознавание речи в системах массового обслуживания населения // Труды СПИИРАН. 2011. Вып. 19. С. 65–86.
5. Фархадов М.П., Васьковский С.В., Ревонченкова И.Ф. Построение интеллектуальных информационных контакт-центров // Автоматизация и современные технологии. 2011. 4. С. 14-23.
6. Джефф Раскин Интерфейс: новые направления в проектировании компьютерных систем. The Human Interface. New Directions for Designing Interactive Systems. / Издательство: Символ-Плюс, 2005 г. 272 С.
7. Потапова Р.К. Речевое управление роботами, лингвистика и современные автоматизированные системы. Изд.2. М.: КомКнига. 2005. 328 с.
8. Michael Harris Cohen,Michael H. Cohen,James P. Giangola,Jennifer Balogh Voice User Interface Design. Nuance Communication Inc. 2004. 368 p.
9. Randy Allen Harris Voice Interaction Design: Crafting the New Conversational Speech Systems (Morgan Kaufmann Series in Interactive Technologies). // Morgan Kaufman Publishers an mprint of Elsevier Science. 2004. 598 P.

10. Daryle Gardner-Bonneau, Harry E. Blanchard Human Factors and Voice Interactive Systems (Signals and Communication Technology). Second edition. // 2008 Springer Science+Business Media LLC. 469 P.
11. Bruce Balentine, David P. Morgan How to Build a Speech Recognition Application: Second Edition: A Style Guide for Telephony Dialogues. // Enterprise Integration Group, Inc. California. (Dec 31, 2001). 319 P.
12. Bruce Balentine and Leslie Degler It's Better to Be a Good Machine Than a Bad Person: Speech Recognition and Other Exotic User Interfaces at the Twilight of the Jetsonian Age. (Mar 26, 2007).
13. Emily Yellin. Your Call Is (Not That) Important to Us: Customer Service and What It Reveals About Our World and Our Lives. // Free Press. A division of Simon & Schuster, Inc. New York, London, Toronto, Sydney. Copyright 2009 by Emily Yellin. 292 P.
14. Susan Weinschenk and Dean T. Barker Designing Effective Speech Interfaces. // WILEY COMPUTER PULISHING. WILEY John Wiley & Sons, Inc. New York. (Feb. 18, 2000). 407 P.
15. James R. Lewis Practical Speech User Interface Design (Human Factors and Ergonomics). // CRC Press (Dec. 15, 2010). 344 pages.
16. Fang Chen Designing Human Interface in Speech Technology // Springer; 1st Edition. edition (Oct. 29, 2010). 382 pages.
17. Величковский Б.М. Когнитивная наука: Основы психологии познания: В 2 т. Т. 1. - М.: Издательский центр "Академия", 2006. – 448с., 432с.
18. Дружинина В. Н., Ушакова Д. В. "Когнитивная психология. Учебник для вузов" М.: ПЕР СЭ. 2002. 480 с.
19. Роберт Солсо "Когнитивная психология" 6-е изд. – СПб.: Питер, 2006 – 589 с. (Серия "Мастера психологии").
20. Борис Величковский "Современная когнитивная психология" М.: Издво МГУ, 1982. 336 стр.
21. Билик Р.В., Мясоедова З.П., Петухова Н.В., Фархадов М.П. Под ред. проф. Жожикашвили В.А. Анализ речевого интерфейса при взаимодействии клиента с автоматизированной системой массового обслуживания. М.: МАКС Пресс. 2007. 112 с.
22. `http://www.etsi.org`
23. Билик Р.В., Мясоедова З.П., Петухова Н.В., Фархадов М.П. Речевой интерфейс как разновидность человеко-машинного взаимодействия / Материалы международной научно-практической конференции "Информационные технологии и информационная безопасность в науке, технике и образовании (ИНФОТЕХ-2009)". Севастополь: СевНТУ, 2009. С. 279-282.
24. Вертлиб В.А., Фархадов М.П., Петухова Н.В. "Электронное государство" как Автоматизированная система массового обслуживания населения. 2008: Макс Пресс, 2008. – 148 с.
25. Жилькин М.Б., Проталынский О.М., Френкель М.Б. Электроннное правительство: региональный аспект // Датчики и системы. 2010. є8. С.3–6

26. Васьковьский С.В. Построение современнных корпоративных телефонных сетей // Датчики и системы. 2009. є9. С. 54–56.
27. Васьковский С.В., Морозов В.П. Иерархический принцип построения инструментальных средств отладки // Датчики и системы. 2010. є6. С. 23–27.
28. Смирнов В.А., Гусев М.Н., Фархадов М.П. Функция лингвистического процессора в системе автоматического анализа неструктурированной речевой информации. // Автоматизация и современные технологии. 2013 є8. С. 22-28
29. Baker J. "The DRAGON system-An overview". // IEEE Transactions on Acoustics, Speech, and Signal Processing, 1975, pp. 24–29.
30. Manning C.D., Schütz H. Foundations of Statistical Natural Language Processing, MIT Press, 1999.
31. Nivre J. et al. Maltparser: A language-independent system for data-driven dependency parsing// Natural Language Engineering, vol. 13, no. 2, 2007. Pp. 95–135.
32. Charniak E.. Statistical Techniques for Natural Language Parsing //Artificial Intelligence.18(4):33–44, 1997.
33. Francis W.N., Kucera H. Brown Corpus Manual, Rhode Island, Brown University, 1964
34. Национальный корпус русского языка: 2003–2005. Сборник статей. Москва, 2005
35. DeRose S.J. Grammatical category disambiguation by statistical optimization // Computational Linguistics 14(1): 31–39, 1988.
36. Chomsky N. Syntactic Structures. The Hague: Mouton, 1957.
37. Uszkoreit H. New Chances for Deep Linguistic Processing // Proceedings of COLING 2002, pages xiv–xxvii, Taipei, 2002.
38. Pollard C., Ivan A. Sag. Head-driven phrase structure grammar. Chicago: University of Chicago Press, 1994.

CODE BOOKS FORMATION OF THE LIMITED CAPACITY AT LOW BIT RATE SPEECH CODING

A. Afanasjev[1], *V. Pimenov*[1]
[1] Academy of Federal Guard Service, Orel, Russia

Abstract

The given report considers problems of synthesis of speech signal processing systems based on the modernization of direct linear prediction method. It is shown that at the limited order of forming model the parameters of the synthesizing filter and appropriate excitation signals to them are becoming interdependent. Thus there is a possibility of the code books size reduction used at vector quantization.

ФОРМИРОВАНИЕ КОДОВЫХ КНИГ ОГРАНИЧЕННОЙ МОЩНОСТИ ПРИ НИЗКОСКОРОСТНОМ КОДИРОВАНИИ РЕЧИ

А. Афанасьев[1], *В. Пименов*[1]
[1] Академия Федеральной службы охраны Российской Федерации, Орел, Россия
fromnet@yandex.ru

Аннотация

В статье рассмотрена задача синтеза системы обработки речевого сигнала на основе модернизации метода прямого линейного предсказания. Показано, что при ограниченном порядке формирующей модели параметры синтезирующего фильтра и соответствующие им сигналы возбуждения становятся взаимозависимыми.

Ключевые слова: низкоскоростная передача речевого сигнала, кодирование речи, линейное предсказание, синтезирующий фильтр, векторное квантование

1. Введение

Для реализации перспективных алгоритмов обработки речи с целью дальнейшей его передаче по каналу связи предлагается использовать системы не только с переменными параметрами, но и с переменной структурой [1]. В процессе функционирования таких систем мощности пространств

представлений параметров и взаимосвязи между структурными элементами, а также их количество могут изменяться. Это позволит повысить показатели функционирования систем обработки речевых сигналов (РС), так как система будет адаптивно изменять свою структуру, выбирая наилучшую из заданного конечного множества вариантов структур.

2. Основная часть

Анализ методов и алгоритмов обработки речевых сигналов показал актуальность работ по созданию нового поколения систем подобного класса, основанных на исследовании статистических и параметрических характеристик распределения параметров речи и изменения в соответствии с ними структуры и параметров системы.

Такая система обработки может состоять из следующих элементов [2]:

– выделители параметров классификации;

– классификаторы;

– устройства, обеспечивающие реализацию макропроцедур обработки речи (векторные квантователи, липредеры);

– устройства, предназначенные для дополнительной обработки речевого сигнала (шумоподавление, изменение спектральных свойств).

Процессом изменения структуры и параметров устройства обработки в реальном масштабе времени должно управлять устройство, в котором будет происходить анализ и выделение параметров РС. В соответствии с полученными характеристиками и сравнении их с эталонами должна выбираться одна из возможных структур обрабатывающего устройства, которая наиболее адекватно отобразит РС на сегменте анализа [3, 4].

Среди многообразия разработанных методов обработки речевого сигнала одним из наиболее эффективных является метод прямого предсказания речи [5]. Из статистического анализа РС известно, что распределение формант и их число для различных звуков речи отличается друг от друга [6]. Данное свойство РС может использоваться для построения системы компрессии речи с изменяемым числом параметров формирующей модели. В зависимости от полученных статистических характеристик обработанного фрагмента речи принимается решение о передаче параметров, характеризующих передаточную функцию голосового тракта формирующего речевой сигнал – линейных спектральных пар и соответствующего сигнала возбуждения, их нахождение реализуется с использованием процедуры анализа через синтез.

При оптимизации системы является количественный и качественный выбор числа параметров, необходимых для обработки речевого сигнала с заданным качеством. При обработке речевых сигналов функционируют два адаптивных контура управления: первый производит оптимизацию структуры обрабатывающего устройства, второй – оптимизацию передаваемых параметров для выбранной структуры системы.

В общем случае при обработке РС и использовании байесова подхода в идеальном виде требуется знание функциональной зависимости ожидаемой ошибки $e(n)$ от принятого решения U и значений $S(n)$ обрабатываемого речевого сигнала. Для вычисления данного значения необходимо статистическое описание наблюдений $S(n)$ а также параметров состояния синтезирующей системы. При этом полезный объем этих данных определяет значение ожидаемой ошибки при любом из ограниченного множества возможных решений**.** Решение задачи синтеза системы в такой постановке осуществляется при известных распределениях вероятности $p(S\,|(\vec{a},\vec{e})$ и $p(\vec{a},\vec{e})$, где S – значения отсчетов речевого сигнала на сегменте анализа, $(\vec{a},\vec{e})$– пара векторов параметров голосового тракта и сигнала возбуждения соответственно, выбранные с использованием процедуры анализа через синтез.

Однако на практике полное статистическое описание $S(n)$ и $(\vec{a},\vec{e})$ получить невозможно, что связано в данном случае с высокой размерностью решаемой задачи и сложностью статистического анализа речевых данных. При этом задачи обработки речевых данных сопровождаются большей или меньшей априорной неопределенностью, которая ограничивает полноту статистического описания. Аппроксимация распределений $S(n)$ и $(\vec{a},\vec{e})$ с использованием нормального распределения носит достаточно приближенный характер, ограничивающий область применения решений, и привносит ошибки в конечные результаты вычислений. Недостатком такого представления и используемого подхода является априорное условие распределения параметров $(\vec{a},\vec{e})$ по нормальному закону, чего на практике зачастую не выполняется. Тем не менее, такой подход является удобным средством для учета имеющихся качественных представлений о статистическом поведении наблюдаемых данных $S(n)$ и параметров $(\vec{a},\vec{e})$ в сочетании с незнанием детальных количественных характеристик, точно определяющих это описание. Именно такое сочетание наиболее характерно для большинства прикладных задач обработки речи.

Качественные представления, основанные на физической сущности рассматриваемой задачи, дают возможность задать структуру распределений вероятности для $S(n)$ и $(\vec{a},\vec{e})$, при этом, если $S(n)$ или $(\vec{a},\vec{e})$ представляют собой конечное множество, то в качестве неизвестных параметров рассматриваются сами вероятности этих значений. Таким образом, параметрическое описание априорной неопределенности является достаточно универсальным средством учета ограниченных априорных сведений. Это описание должно удовлетворять двум подчас противоречивым требованиям. Во-первых, оно должно качественно правильно и, по возможности, количественно точно отражать ограниченные априорные знания, так, чтобы распределения с плотностями $p(\vec{a},\vec{e})$ при соответствующих переменных действительно представляли возможные в данной задаче распределения. Во-вторых, число параметров не должно быть слишком велико. Увеличение размерности приводит к ухудшению качества решения основной за-

дачи как из-за сложности технической реализации алгоритмов обработки данных наблюдения $S(n)$, так и из-за утраты некоторой доли входной информации, которую неизбежно приходится затрачивать для определения значений или исключения неизвестных мешающих параметров. Поэтому синтезируемую в условиях априорной неопределенности систему дополним подсистемой проверки правильности априорных предположений, положенных в основу принятого в данной задаче параметрического описания. Ее задача – установить, верно ли качественно введенное описание (например, при фильтрации процесса $S(n)$ – соответствует ли действительности аппроксимация $S(n)$ полиномом заданной степени с вычисленными коэффициентами или эта модель недостаточно адекватна) и достаточно ли число введенных параметров относительно $(\vec{a}, \vec{e})$. Такой алгоритм дает возможность при необходимости усложнить параметрическое описание, увеличив число параметров или качественно изменив модель априорной неопределенности, введя статистическую зависимость параметров $\{\vec{a}\}$и $\{\vec{e}\}$.

Наличие данных зависимостей объясняется тем, что в стандартах низкоскоростного кодирования речи используется ограниченный порядок анализирующего и синтезирующего фильтров, что определяется возможностью их физической реализации при необходимой и достаточной точности описания передаточной функции голосового тракта человека. Так, сущность метода линейного предсказания заключается в том, что выборка речевого сигнала $S(n)$ может быть предсказана линейной комбинацией предшествующих отсчетов этого сигнала:

$$S'(n) = \sum_{i=1}^{M} a_i S_{n-i} + e(n) \tag{1}$$

где *S'(n)* – предсказанное значение речевого сигнала;
ai – весовой коэффициент или коэффициент линейного предсказания;
M – число коэффициентов или порядок линейного предсказания,
$e(n)$ – ошибка предсказания.

Возникающая при этом ошибка предсказания находится по линейно-разностному уравнению (2), которое описывает функционирование фильтра анализа модели линейного предсказания:

$$e(n) = S(n) - S'(n) = S(n) - \sum_{i=1}^{M} a_i S(n-i). \tag{2}$$

Задача анализа речевого сигнала методом линейного предсказания заключается в его фильтрации линейной системой с передаточной характеристикой вида:

$$A(z) = 1 - \sum_{i=1}^{M} a_i Z^{-i}. \tag{3}$$

Обратная ей передаточная функция представляет собой фильтр синтеза и определяется соотношением

$$H(Z) = \frac{1}{A(Z)} = \frac{1}{1 - \sum_{i=1}^{M} a_i Z^{-i}}. \tag{4}$$

Теоретическим основополагающим базисом метода линейного предсказания является авторегрессионная модель, успешно применяемая для решения различных задач цифрового спектрального анализа и предполагающая в общем "идеальном"случае бесконечный порядок формирующей системы при возбуждении ее сигналом в виде дискретного белого гауссовского шума. Ее идентификация связана с решением системы алгебраических матричных уравнений Юла-Уокера [7]. В классической постановке задачи параметрического цифрового спектрального анализа возбуждение формирующего фильтра осуществляется сигналом $u(n)$, представляющим собой реализации белого шума с математическим ожиданием равным нулю и единичной дисперсией.

$$\begin{cases} M\{u(n)\} = 0 \\ D\{u(n)\} = \sigma^2\{u(n)\} = 1. \end{cases} \tag{5}$$

Точность идентификации математической модели исследуемого процесса напрямую связана с выбором величины ее порядка M. В качестве критерия настройки модели в предположении о гауссовском законе распределения исходного процесса используется взвешенная среднеквадратическая ошибка $e^2(n)$.

$$e^2(n) = d_2(\vec{S}, \vec{S}') = \frac{1}{N}(\vec{S} - \vec{S}')^T(\vec{S} - \vec{S}') = \frac{1}{N}\sum_{i=1}^{N}(s_i - s_i')^2. \tag{6}$$

где $\vec{S}$- вектор оригинального речевого сигнала, $\vec{S}'$- вектор синтезированного речевого сигнала, N- количество отсчетов на сегменте анализа.

Применительно к задаче предсказания речи повышение порядка передаточных функций фильтров анализа и синтеза приводит к "обелению"сигнала остатка предсказания.

В классической постановке задачи параметрического цифрового спектрального анализа на основе авторегрессионной модели линейное разностное уравнение формирующего фильтра выглядит следующим образом (**??**):

$$y\,(nT) = -\sum_{m=1}^{M-1} a_m \cdot y\,(nT - mT) + u\,(nT)\,, \tag{7}$$

где $y\,(nT)$ – выходной сигнал, T– интервал дискретизации, $\{a_m\}$- коэффициенты фильтра, M – порядок фильтра.

Его амплитудно-частотная характеристика определяется в виде:

$$A(\omega T) = \frac{1}{\sqrt{\left(1+\sum_{m=1}^{M-1} a_m \cdot \cos m\omega T\right)^2 + \left(\sum_{m=1}^{M-1} a_m \cdot \sin\ m\omega T\right)^2}}, \quad (8)$$

а спектральная плотность мощности:

$$G(\omega) = \frac{\sigma^2\{u(n)\} \cdot T}{\left(1+\sum_{m=1}^{M-1} a_m \cdot \cos m\omega T\right)^2 + \left(\sum_{m=1}^{M-1} a_m \cdot \sin\ m\omega T\right)^2}, \quad (9)$$

где ω - круговая частота дискретного преобразования Фурье.

Повышение порядка модели в выражениях (6), (7), (8) и (9) приводит к получению более точных оценок относительно анализируемого сигнала $\vec{S}$.

На практике при реализации линейного предсказания значение M всегда ограничено, что приводит к возникновению сигнала $e(n)$, являющегося сигналом возбуждения фильтра синтеза модели линейного предсказания. Таким образом, сигнал $e(n)$ уже не является реализациями белого шума с математическим ожиданием равным нулю и единичной дисперсией, а становится квазидетерминированным относительно множества $\{a_m\}$ и связан с ним соответствующими корреляционными зависимостями.

При формировании ограниченных множеств параметров голосового тракта $\{\alpha_i(n)\}$ и сигналов возбуждения на основе остатка предсказания в виде кодовых книг данные зависимости вырождаются в соответствующие классы подпространств соответствий между собой и определяют элементы декомпозиции речевого сигнала. Данное соответствие и взаимозависимости дают возможность значительного сокращения мощности кодовых книг хранящих эталоны параметров преобразований используемых при обработке речевого сигнала.

Представленный подход статистического описания $S(n)$ и $(\vec{a}, \vec{e})$ соответствуют тому, что в каждой конкретной практической задаче создается некоторая аналитическая модель для описания статистических свойств $S(n)$ и $(\vec{a}, \vec{e})$ с той степенью полноты и подробности, которая соответствует имеющимся знаниям о закономерностях их поведения, физических свойствах, взаимосвязи между собой. Подобная модель является наиболее сжатым описанием имеющегося опыта, содержащего, как результаты изучения данных закономерностей, так и, эмпирические данные относительно $S(n)$ и $(\vec{a}, \vec{e})$.

Однако на практике получить статистическое описание параметров $(\vec{a}, \vec{e})$ отдельного диктора можно лишь для реализации задачи аутентификации, в то время как для более широкого класса задач обработки речи (кодирование) такое представление достаточно проблематично, что связано с отсутствием априорного знания всех возможных голосов дикторов подвергаемых обработке. В итоге создаются кодовые книги, содержащие вектора

(образцы) параметров описывающих передаточную функцию голосового тракта и соответствующего им сигнала возбуждения, что позволяет перейти от учета статистики взаимосвязи между отдельными значениями данных векторов, вычисляемыми при обработке, к статистике взаимосвязи векторов между соответствующими кодовыми книгами.

На практике бывает и так, что кроме эмпирических данных всякая иная априорная информация относительно $S(n)$ и $(\vec{a}, \vec{e})$ отсутствует. Однако, если эти данные получены в обстановке, статистически однородной или хотя бы статистически связанной с той, в которой принимается решение по данным наблюдения $S(n)$, то они являются в определенной степени статистическим эквивалентом аналитических моделей для распределений вероятности $S(n)$ и $(\vec{a}, \vec{e})$, необходимых для нахождения оптимального правила принятия решения. Степень подобной эквивалентности, зависит от объема имеющихся эмпирических данных, которые, в свою очередь, могут быть использованы по-разному: непосредственно для нахождения недостающих распределений вероятности; для оценки функциональной зависимости апостериорного риска от $S(n)$ и решения U; для уточнения структуры и параметров решающего правила, т. е. алгоритма обработки данных наблюдения $S(n)$.

3. Заключение

Таким образом, в представленной работе показаны возможные пути сохранения приемлемых качественных характеристик синтезированного РС при уменьшения скорости передачи благодаря использованию выявленных зависимостей между элементами декомпозиции РС. Актуальность исследований и значимость полученных результатов подтверждаются существующими объективными требованиями, предъявляемыми к разрабатываемым системам обработки РС, в частном случае при низкоскоростном кодировании РС, сформированными с учетом тенденций развития систем инфокоммуникаций.

ЛИТЕРАТУРА

1. Емельянов С.В. Системы автоматического управления с переменной структурой. – М.: Наука. Физматлит, 1967. – 336с.
2. Справочник по теории автоматического управления. Под редакцией А.А.Красовского. – М.:Наука, 1987. – 712с.
3. Казаков И.Е. Статистическая динамика систем с переменной структурой. – М.: Наука, 1977. – 416с.
4. Бухалев В.А. Распознавание, оценивание и управление в системах со случайной скачкообразной структурой. – М. Наука, 1996. – 288с.
5. Прохоров Ю. Н. Статистические модели и рекуррентное предсказание речевых сигналов / Ю. Н. Прохоров. – М. : Радио и связь, 1984. – 240 с.

6. Шелухин О. И. Цифровая обработка и передача речи / Под ред. Шелухина О. И. – М.: Радио и связь, 2000. – 456 с. : ил.
7. Марпл – мл. С. Л. Цифровой спектральный анализ и его приложения. – М.: Мир, 1990. – С. 216-224

METHODIC OF MULTI-PARAMETER OPTIMIZATION OF DATA TRANSMISSION PERFORMANCE VIA RADIO FREQUENCY OF 868 MHZ

D.A. Aminev[1], *R.F. Azizov*[2]

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

[2]National Research University Higher School of Economics, Moscow, Russia

[1]aminev.d.a@yandex.ru, [2]radomir.azizov@gmail.com

Abstract

The possibility of better performance in data transmission via radio frequency of 868 MHz is investigated. Multi-parameter optimization problem is formulated and the methodic, which consists of several stages, is offered. A series of experiments according to the proposed methodic made.

1. Introduction

The rapid growth of wireless transmission due to the development of information technologies, reduction of cost, overall dimensions and power consumption of electronic devices in general and small-sized receiver-transmitters as well. The most common wireless transmission for radio frequencies are 433 MHz, 868 MHz and 2.4 GHz [1]. When deploying a wireless network designer needs to agree on a number of key parameters. Difficulties in assessing the performance of the wireless network are defined:

- a wide range of technologies used in the implementation of wireless transmission;
- a large number of parameters that determine the performance of the channel;
- the presence of regulated and unregulated settings that determine the performance of the channel wireless transmission;
- lack of legal channels wireless transmission dependencies performance of adjustable parameters.

This work is based on the parameters and their quantitative estimates obtained in the development of receiver-transmitter-based TI CC1101 transceiver with a frequency of 868 MHz Texas Instruments [2]. The aim is to improve the quality of design solutions, using data transmission technology based on the receiving-small low-power transmitter operating at a frequency of 868 MHz, by selecting the quantitative values of the adjustable parameters of channels to ensure their optimal performance.

2. Methodic of multi-parameter optimization performance radio frequency of 868 MHz

2.1. The formulation of the problem. We consider the ad hoc network's decentralized structure [3, 4], which is characterized by a set of parameters $\{p_i\}\,(i = 1, ..., n)$, each of which is known range $\Delta_i = \hat{x}_i - \check{x}_i$. The parameter p_1 corresponds to the channel capacity, and other parameters $\{p_j\}(j = 2, ..., n)$ corresponds to a set of adjustable parameters. For the channel in question is known as a set of experimentally determined dependencies $f_j(x_j)$ of values x_1 of performance p_1 from of varied parameter values x_j, provided that the remaining control parameters are fixed. Is required to determine a set of values $\{x_j^*\}$ adjustable parameters $\{p_j\}$, where the maximum value x_1^* of performance p_1.

2.2. Steps methodic. Methodic of solving the above problem has 6 stages.

- Determination of the parameters affecting the performance of the channel.
- Identify the composition of adjustable parameters.
- Development of a mobile stand for experimental verification of the channel allows you to specify the required values of adjustable parameters and to determine performance parameter values.
- Determination in step $s = d\,(s = \overline{0, (n - J)})$ set of basic values of $f\,\{x_j^0\}$ adjustable parameters $\{p_j\}$, relevant recommendations (from the point of view of the greatest guarantee of packet delivery) 868 MHz radio settings. Values $\{x_j^0\}$ are considered in the method of the coordinates of the point 0 of the multidimensional space.
- Using a stand, to experimentally determine x_i performance parameters at 0. In step $s > 0$ experimentally determined dependence $f_j(x_j)$ values x_i of performance p_1 from the values x_j of the variable parameter p_j provided that the other adjustable parameters are fixed and correspond to the coordinates of the point A of the previous step. Analysis of the relationship to determine the value x_j^s of the parameter p_j, maximizing the value of the performance of step s. By changing the j-th coordinate obtain the coordinates of s.
- Consistently performing experimental measurements of characteristics and analysis procedures to obtain on $(s - I)$-th step coordinates of the point $*$, corresponding to a set of values x_j^* radio settings that provide the best performance.

3. The implementation of multi-parameter optimization for radio performance

Step 1. Studies have established the following parameters defining the operation of a wireless network:

1) the performance of the channel;
2) code integrity control codes or recovery;

3) the size of the frame;
4) range, which regulates the expectation of delivery confirmation frame;
5) modulation radio.

Step 2. Parameters p2, p3, p4, p5, p6 - are variable parameters. The list of parameters is shown in Table 1.

Symbol	Name	Name
p2	frame (frame size)	2..60 bytes
p3	modulation	GFSK, 4-FSK, MSK
p4	rs (Reed-Solomon codes)	0 (off), 2,4,6,8,10 byte codes
p5	wh (whitering)	1 – on / 0 – off
p6	ack (confirmation)	1 – on / 0 – off

Table 1: Parameter list

Step 3. The basis of the stand are receiving and transmitting devices based TI SS1101 transceiver and microcontroller ST STM32F0, a set of specially developed software to estimate the rate of information exchange. Furthermore, the program shows received power level, dBm. The means employed are: PC with operating system GNU / Linux; Two of the wireless transceiver module (Fig. 1).

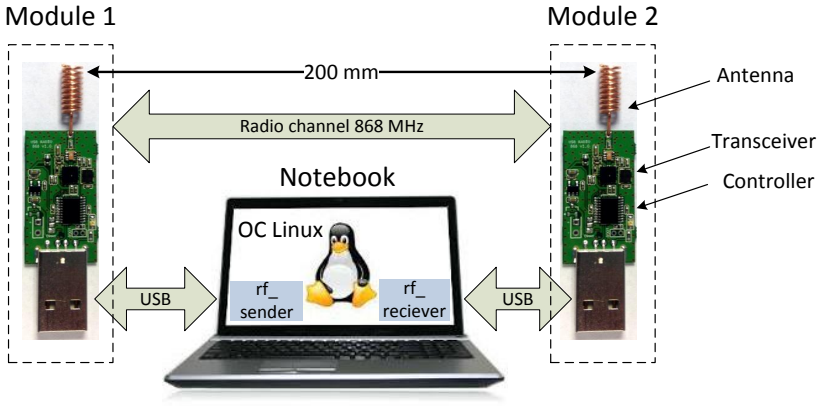


Fig. 1: The experimental scheme

The modules, developed by "Open development" Ltd. [5], are a small-sized wireless devices with a USB interface and a spiral antenna [6]. The modules

are located at a distance of 2 m from each other. To prevent saturation of receivers, transmission power is set at 0 dBm. For the experimental programs have been developed and rf_sender rf_reciever. Program rf_sender it is a traffic generator in accordance with the parameters of Table 1. Program rf_reciever - traffic analyzer produces records come and dropped packets, calculates the speed and shows the signal strength when receiving data.

Step 4: Find the optimal value of the bandwidth of 868 MHz radio channel starts with the values of adjustable parameters determined for the most guaranteed packet delivery.

Step 5. Examination of parametres

Parameter	p2 at p4=4, p5=0, p6=0			p4 at p2=50, p5=0, p6=0			p5 at p2=50, p3=5, p6=0		p6 at p2=50, p4=4, p5=0	
value	10	30	50	8	4	0	1	0	1	0
p1, Bps	480	1200	1750	120	870	1750	870	1750	800	1750

Table 2: Results of experiment

p2. The shorter the package, the more likely it passes, but lower overall rate. The actual length of the packet channel is the sum of service data useful and consisting of a header and integrity monitoring. Since sending packet contains the overhead of packet delivery and processing of the transmitter and, given the presence of ancillary data in the packet, it is advisable to use packets of at least 40-50 bytes. The packet structure is disclosed in Table 3.

Size, Byte	1	1	1	1	N	4
Name	PK	Seq	Dst	Src	Data	Ctrl

PK(Packet control) - header bytes								
Size, Bit	7	6	5	4	3	2	1	0
Name	Addr type		Pkt type		Ack req	Next header	Crypto	Reserved

Table 3: Structure of the information package

Here: PK - byte header; Seq - packet sequence number; Dst - destination address, a sequence of 1, 2, or 8 bytes, in which the address filtering; Src - address of the sender, the sequence of 1, 2 or 8 bytes; Data - data; Ctrl - code integrity checking and recovery; Addr type - addressing mode; Pkt type - the type of package; Ack Req - confirmation; Next header - followed by an additional header; Crypto - encrypted packet; Reserved - reserve.

p3. Various modes of modulation (GFSK, 4-FSK, MSK) had no impact on the network bandwidth in this experiment. Thus, when $p2 = 50$, $p4 = 4$, $p5 = 0$, $p6 = 0$ and all $p2$ combinations value $p1$ was 1750 bytes /sec.

p4. The longer the Reed-Solomon code, the more noise immunity, but increases the time to process the packet. If a significant deterioration in communication acceptable option would be an increase in code reduction to 4 bytes, which will be reduced to 2 bytes in the packet.

p5. Whitening - a logic procedure XOR data with a pseudorandom sequence. Mode bleaching increases the noise immunity of transmission that has a positive effect in low-level signal, but with a good signal reduces performance.

p6. When the standby time of the receipt confirmation if the sender does not receive a receipt for delivery at a specified time, there is a re-posting of data. Step 6. A graphical representation of the results of the experiment are shown in Fig. 2.

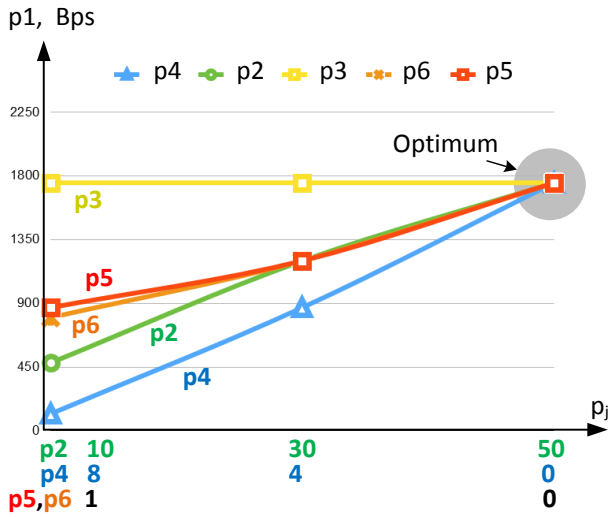


Fig. 2: The values of performance parameters change

As can be seen from the graphs, the area of optimum parameters of performance criteria is in the vicinity of frame size 50, disconnected mode handshake, bleaching and Reed-Solomon codes.

4. Conclusions

After performing the steps of a technique of optimization parameters of the radio channel 868 MHz and iterate combinations of parameters that affect the performance revealed that its maximum value is about 1800 bytes/sec, which is four times more combinations with a relatively high level of reliability of packet delivery of 450 bytes/sec.

With good signal quality for the transmitter to use the maximum possible size of the transmitted packet. Using a Reed-Solomon code of 4 bytes in length, will restore to two corrupted bytes in the packet and will not have a noticeable effect on transmission performance. The “whitening” and the actual receipt of parcels in low signal level. Changing modulation did not give a significant gain in throughput.

Data transfer via radio 868 MHz quite promising, it is important to select the wireless communication lines to achieve maximum performance.

REFERENCES

1. Tanenbaum Je., Ujezeroll D. Komp'juternye seti. 5-e izd. – SPb.: Piter, 2012. – 960 p.
2. Datasheet CC1101 Low-Power Sub-1 GHz RF Transceiver (Rev. I) 05 Nov 2013
3. Azizov R.F., Aminev D.A., Ivanov I.A.. Uvajsov S.U. Organizacija svjazi na osnove prioritetov dlja uluchshenija JeMS pri informacionnom obmene v decentralizovannoj seti. // Tehnologii jelektromagnitnoj sovmestimosti. 2013. №4(47). pp. 5–8.
4. Aminev D.A., Azizov R.F., Uvajsov S.U., Jurkov N.K. Opredelenie optimal'nyh harakteristik algoritma konkurentnogo dostupa k srede dlja minimizacii vremeni peredachi v decentralizovannyh besprovodnyh setjah // Prikaspijskij zhurnal: upravlenie i vysokie tehnologii. - Astrahan': -2015 №1. — pp. 101–107.
5. `open-dev.ru` — site of the Corp. "Open development"
6. Aminev D., Azizov R., Uvaysov S. Recommendations for the choice of antenna transceivers of decentralized self-organizing networks // V kn.: Innovacii na osnove informacionnyh i kommunikacionnyh tehnologij: materialy mezhdunarodnoj nauchno-tehnicheskoj konferencii / Otv. red.: I. A. Ivanov; pod obshh. red.: S. U. Uvajsov. M.: MIJeM NIU VShJe, 2013. pp. 480–481.

APPROXIMATE DESCRIPTION OF DYNAMICS OF A CLOSED QUEUEING NETWORK INCLUDING MULTI-SERVERS[1]

S. Anulova

V.A.Trapeznikov Institute of Control Sciences of RAS, Moscow, Russia
anulovas@ipu.ras.ru

Abstract

We investigate a closed network consisting of two multi-servers with n customers. Service requirements of customers at a server have a common cumulative distribution function. The state of the network is described by the following state parameter for each server: empirical measure of the age of customers being serviced multiplied by n^{-1}. For $n \to \infty$ we approximate the state function of time by means of measure-valued techniques. The approximation of a single infinite/multi-server dynamics is currently studied by famous scientists H. Kaspi, K. Ramanan, W. Whitt et al.

A motivation for studying such systems is that they arise as models of computer data systems and call centers.

Keywords: Multi-server queues, GI/G/n queue, fluid limits, mean-field limits, strong law of large numbers, measure-valued processes, call centers.

1. Introduction

1.1. Review of investigated contact centers models. In the last ten years an extensive research in mathematical models for telephone call centers has been carried out, cf. [2-16]. The object has been expanded to more general customer contact centers (with contact also made by other means, such as fax and e-mail). One of important relating questions is the dynamics of multi-server queues with a large number of servers and customer abandonment. In order to describe the object efficiently the state of the model must include: 1) for every customer in the queue the time that he has spent in it and 2) for every customer in the multi-server the time that he has spent after entering the service area, that is being received by one of the available servers. Thus the model is a hybrid process, with a discrete component of the arrival point process and continuous components of time spent in waiting and serving. The corresponding state process at every moment is a discrete random measure — with unit point masses. The number of point masses equals to the number of customers in the system at the present moment. The locations of point masses are equal to the described times spent by these customers until the present moment. The focus of research was on multi-server queues with a large number of servers, because it is typical of contact centers. For such queues were found fluid limits with the number of servers tending

[1]This work was supported by RFBR grant No. 14-01-00319 “Asymptotic analysis of queueing systems and nets”.

to infinity. Notice that such a limit is a deterministic fumction of time with values in a certain measure space, or in a space containing such a component. These developed deterministic fluid models provided simple first-order performance descriptions for multi-server queues under heavy loads, also allowing abandonment. The models were gradually generalized to the $G/GI/s + GI$ model and even more complicated ones. This notation implies the arrival process as a general stationary one, service times of all customers independent and identically distributed with a general distribution, the number of servers equal to s, and independent identically distributed abandon times while waiting in queue[2]. Further the arrival process was generalized to an arbitrary jump process independent of the sequence of service requirements and having a fluid limit [2].

1.2. A new model for contact centers. We suggest here a more suitable model for contact centers. The number of customers is fixed. Customers may be situated in two states: normal and failure. There is a multi-server which repairs customers in the failure state. The repair time/the time duration of a normal state is a random variable, independent and identically distributed for all customers. Now "the arrival process" in the multi-server does not correspond to that of the previous $G/GI/s + GI$ model. For a large number of customers and a suitable number of servers calculate the number of current failures, so much as an approximation. This is a continuation of our work [1], where a single multi-server in a network was functioning.

2. The mathematical model

Consider a closed network consisting of n customers and two multi-servers.

Multi-server 1 consists of n servers, the time they service a customer has distribution F_1. Multi-server 2 consists of $s^{(n)}n$ servers with $s^{(n)} \in (0, 1)$, the time they service a customer has distribution F_2. Service times are independent for all servers and all customers. We assume that F_i has a bounded and continuous density, $i = 1, 2$. Denote

$$\text{for } t \geq 0, u \geq 0 \qquad G_i(t) = 1 - F_i(t), \qquad H_i(t; u) = \frac{G_i(t + u)}{G_i(u)}, \ i = 1, 2 \ .$$

Denote by $\nu_i^{(n)}(t; du)$ the empirical measure of the age of customers in the multi-server i at moment t multiplied by n^{-1}, $i = 1, 2$ ("age" means the length of the time segment from entering the server up to moment t). Set $q^{(n)}(t) = 1 - (\nu_1^{(n)} + \nu_2^{(n)})(t; [0, \infty))$, $t \geq 0$. $q^{(n)}$ is the normed number of customers in the failure state waiting for service when the multi-server is completely occupied. Normed means multiplied by n^{-1}.

Definition 1. *A continuous measure-valued function $\nu_i(t; du)$, $t \geq 0$, is the "fluid" limit of $\nu_i^{(n)}(t; du)$, $t \geq 0$, if for every $t \geq 0$ $\sup_{s\in[0,t]} \rho(\nu_i^{(n)}(s; \cdot), \nu_i(s; \cdot)) \to 0$, $i = 1, 2$ (for metric ρ see section 19.2* [18]*).*

Our result is a generalization of Theorem 3.7 [2].

Assumptions

[2]See the complete description of shorthand system notations in [17, subsection III 1b Classification of Simple Queues]

- measure $\nu_i^{(n)}(0)$ has a weak limit $\nu_{0;i}$, $i = 1, 2$
- $\lim_{n\to\infty} s^{(n)} = s$ exists and belongs to $(0, 1)$.

We find the solution of the fluid limit problem in a special case — when the limit queue is locally (in time) uniformly positive, see the next section.

3. Equations system for the fluid limit and its solution

We shall describe the fluid limit dynamics by the following equations system.

State space and evolution in it: $\nu_i(t; du)$ is a non-negative finite measure function on $[0, \infty) \times [0, \infty)$, $i = 1, 2$, and $q(t)$ is a function on $[0, \infty)$ with values in $(0, 1)$. They represent the normed customers age distributions in multi-servers and in the queue.

Initial conditions: non-negative finite measures $\nu_{0;i}$ on $[0, \infty)$, $i = 1, 2$, and $q_0 > 0$ satisfying $\nu_{0;1}([0, \infty)) + \nu_{0;2}([0, \infty)) + q_0 = 1$ and $\nu_{0;2}([0, \infty)) = s$, that is the server 2 is fully occupied.

Equations:

- the multi-server i is losing its mass with the velocity:

$$v_i(t) = \int_0^\infty \frac{F_i'}{1 - F_i}(u)\nu_i(t; du), \ i = 1, 2,$$

  that is, the normed amount of abandoning customers (the mass increase takes place simultaneously with arriving customers);
- the mass of the queue develops with the velocity $v_1 - v_2$:

$$q(t) = q_0 + \int_0^t (v_1 - v_2)(s)ds;$$

- the normed customers age distributions in multi-servers develop in this way:

$$\nu_1(t; A) = \int_{\{u:u+t\in A\}} H_1(t; u)\nu_{0;1}(du) + \int_{\{s:t-s\in A\}} v_2(s)(1 - F_1)(t - s)ds,$$

$$\nu_2(t; A) = \int_{\{u:u+t\in A\}} H_2(t; u)\nu_{0;2}(du) + \int_{\{s:t-s\in A\}} v_2(s)(1 - F_2)(t - s)ds,$$

  for every Borel subset A of $[0, \infty)$.

Theorem 1. *If on some segment* $[0, T]$ *this system has a solution* (ν_1, ν_2, q) *with uniformly positive* q *and* $\nu_2(t; [0, \infty)) = \nu_{0;2}([0, \infty)) = s$, $t \in [0, T]$, *then the solution of the system on* $[0, T]$ *is unique, and this solution is the fluid limit of the sequence of processes* $(\nu_1^{(n)}, \nu_2^{(n)}, q^{(n)})$, $n = 1, 2, \ldots$, *restricted to* $[0, T]$.

The conditions of the full occupation of the server 2 and the uniform positivity of the queue size reduce our network model to a single multi-server model with the arrival process generated by the server 2. This process turns out to be a renewal process. We make use of renewal processes concepts presented in [17, subsection V.1].

REFERENCES

1. Anulova S. V. Age-distribution description and “fluid” approximation for a network with an infinite server // International Conference ”PROBABILITY THEORY and its APPLICATIONS” (Moscow, June 26-30, 2012), Abstracts, editors A.N. Shiryayev, A.V. Lebedev. M.: LENAND. 2012. P. 219-220. ISBN 978-5-9710-0492-9.
2. Kaspi, H. & Ramanan, K. Law of large numbers limits for many-server queues // Ann. Appl. Probab. 2011. V. 21. P. 33-114.
3. Whitt, W. Engineering solution of a basic call-center model // Management Sci. 2005. V. 51. P. 221–235.
4. Brown, L., N. Gans, A. Mandelbaum, A. Sakov, H. Shen, S. Zeltyn, L. Zhao. Statistical analysis of a telephone call center: A queueing-science perspective // J. Amer. Statist. Assoc. 2005. V. 100. P. 36–50.
5. Whitt, W. Fluid models for multiserver queues with abandonments // Oper. Res. 2006. V. 54. P. 37-54.
6. Pang, G.; Talreja, R. & Whitt, W. Martingale proofs of many-server heavy-traffic limits for Markovian queues // Probab. Surv. 2007. V. 4. P. 193-267.
7. Reed, J. The $G/GI/N$ queue in the Halfin-Whitt regime // Ann. Appl. Probab. 2009. V. 19. P. 2211-2269.
8. Xiong, W. & Altiok, T. An approximation for multi-server queues with deterministic reneging times // Ann. Oper. Res. 2009. V. 172. P. 143-151.
9. Koçağa, Y.L. & Ward, A.R. Admission control for a multi-server queue with abandonment // Queueing Syst. 2010. V. 65. P. 275-323.
10. Dai, J. & He, S. Many-server queues with customer abandonment: A survey of diffusion and fluid approximations // Journal of Systems Science and Systems Engineering, SP Systems Engineering Society of China. 2012. V. 21. P. 1-36.
11. Gamarnik, D. & Stolyar, A. L. Multiclass multiserver queueing system in the Halfin-Whitt heavy traffic regime: asymptotics of the stationary distribution // Queueing Syst. 2012. V. 71. P. 25-51.
12. Ward, A. R. Asymptotic analysis of queueing systems with reneging: A survey of results for FIFO, single class models // Surveys in Operations Research and Management Science. 2012. V. 17. P. 1 - 14.
13. Gamarnik, D. & Goldberg, D. A. On the rate of convergence to stationarity of the $M/M/n$ queue in the Halfin-Whitt regime // Ann. Appl. Probab. 2013. V. 23. P. 1879-1912.
14. Zhang, J. Fluid models of many-server queues with abandonment // Queueing Syst. 2013. V. 73. P. 147-193.
15. Walsh Zuñiga, A. Fluid limits of many-server queues with abandonments, general service and continuous patience time distributions // Stochastic Processes Appl. 2014. V. 124. P. 1436-1468.
16. Kang, W. & Pang, G. Equivalence of Fluid Models for $Gt/GI/N + GI$ Queues // ArXiv e-prints. 2015.
17. Asmussen, S. Applied probability and queues. 2nd revised and extended ed. Springer, New York, 2003.
18. P.-L. Hennequin and A. Tortrat. Théorie des probabilités et quelques applications. Masson et Cie, Paris, 1965.

THE COGNITIVE APPROACH TO PROCESSING LARGE AMOUNTS OF SPECIALIZED, UNSTRUCTURED, TEXT INFORMATION

Bakanov A.S.
Institute of Psychology of Russian Academy of Sciences, Moscow, Russia

Abstract

This paper proposes a cognitive approach to processing large amounts of specialized, unstructured text information. A description of the experiments and the results are given. In the described approach are used psychological techniques and content analysis of text by using a specially developed vocabulary. Described original method to visualize the structure and presentation of textual information.

Keywords: structuring of textual information, specialized textual information, information systems, decision-making.

КОГНИТИВНЫЙ ПОДХОД К ОБРАБОТКЕ БОЛЬШИХ ОБЪЕМОВ СПЕЦИАЛИЗИРОВАННОЙ, НЕСТРУКТУРИРОВАННОЙ ТЕКСТОВОЙ ИНФОРМАЦИИ[1]

А.С. Баканов
Институт психологии РАН, Москва, Россия
arsb3151@gmail.com

Аннотация

В статье предложен когнитивный подход к обработке специализированной, неструктурированной текстовой информации, приведено описание проведенных экспериментов, описаны полученные результаты. В предлагаемом подходе используются как психологические методики, так и контент-анализ текста с использованием специально разработанного словаря. Описаны оригинальные способы визуализации структуры и представления текстовой информации.

Ключевые слова: структурирование текстовой информации, специализированная текстовая информация, информационные системы, принятие решений.

[1]Государственное задание ФАНО РФ № 0159-2015-0007

1. Введение

Статья посвящена вопросам, структурирования и обработки специализированной текстовой информации человеком, в процессе человеко-компьютерного взаимодействия. В данной статье описан подход к обработке и структурированию текстовой информации. Данный подход разработан в результате исследования особенностей профессионального труда специалистов управленческой деятельности, чьи должностные обязанности связаны с решением производственных задач посредством системы электронного документооборота. В процессе профессиональной деятельности при обработке входящих документов, осуществляется сортировка, т.е. документ направляется в профильное подразделение организации. Задачи о сортировке являются классическими задачами принятия решений. Принятие решений является основой управленческой деятельности. Вообще говоря, принятие решений входит в любую профессиональную деятельность и может относиться и ко всей деятельности в целом, и к отдельным действиям или даже его компонентам. Процесс принятия решения включает выявление проблемной ситуации, сбор и анализ информации необходимой для разрешения проблемной ситуации, мысленное выдвижение гипотез (вариантов решения), оценку гипотез, выбор того варианта решения, который наилучшим образом обеспечивает достижение результата. Применительно к нашему случаю, т.е. к задаче о сортировке документа, проблемной ситуацией является наличие документа, который необходимо направить в профильное подразделение для исполнения. На этапе сбора и анализа информации, специалист управленческой деятельности читает документ, анализирует и структурирует полученную информацию. Сравнивает имеющиеся у него представления о данной проблемной области с информацией, полученной в процессе чтения документа, и затем выдвигает гипотезы, т.е. возможные варианты решения задачи о сортировке. После чего осуществляется выбор альтернативы, осуществляется решение задачи о сортировке.

2. Обработка специализированной текстовой информации

Для исследования процесса обработки человеком специализированной текстовой информации, были проведены исследования [3], [4]. Первоначально предъявлялся стимульный материал (в виде вопросов, ключевых слов, графических образов), а затем испытуемому предъявлялась текстовая информация, после прочтения которой, испытуемый принимал решение о сортировке (т. е. определял департамент или подразделение организации, куда документ должен быть направлен для исполнения), отвечал на вопросы по тексту и проходил психологическое тестирование [10]. Испытуемым предъявлялись специализированные тексты определенной тематики, как прошедшие предварительную обработку – т.е. с выделенными опорными словами (рис. 1), так и тексты без предварительной обработки. В ходе исследований, использовалась установка контроля движения взора испытуемого (`www.smivision.com`). С ее помощью были получены эксперимен-

тальные данные: траектория взора, диаметр зрачка испытуемого, скорость перемещения взора, как в процессе чтения текста, так и в процессе принятия решения.

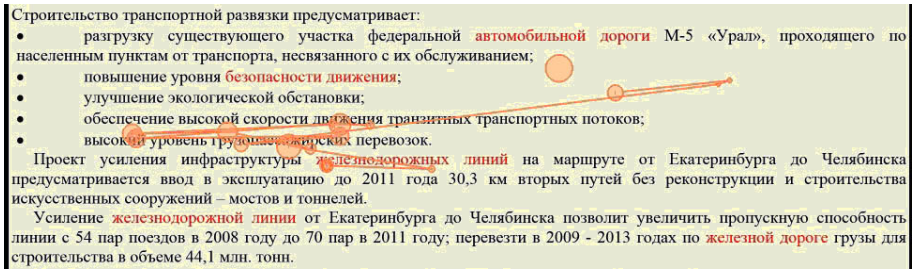


Рис. 1

На рисунке 1 представлена траектория движения взора испытуемого по строкам текста и длительность фиксации взгляда на определенных словах или словосочетаниях. Длительность фиксации взгляда на словах или словосочетаниях пропорциональна диаметру окружности см. рис. 1. Взаимодействие человека с информационными системами в процессе профессиональной деятельности, изучается различными научными направлениями: информационные технологии, инженерная психология и т.д. Это обусловлено актуальностью, а также широким и повсеместным использованием информационных систем [5], [11], [12]. Вопросам, связанным с проектированием информационных систем, а также проблемам взаимодействия человека с информационными системами посвящено значительное количество работ [2]. Среди исследований, касающихся проектирования информационных систем необходимо выделить работу Вишневского В.М. [6], также необходимо отметить работу коллектива авторов: Т. В. Атанасовой, Т.Н. Савченко, Г.М. Головиной и др., посвященную вопросам информационного взаимодействия человека с информационными системами, в которой описаны психологические механизмы взаимодействия человека с интеллектуальной информационной средой обитания [1].

В настоящее время в управленческой деятельности широко используются информационные и компьютерные системы. На современном уровне развития информационных и компьютерных систем, а также телекоммуникационных технологий особую актуальность представляют вопросы эффективного взаимодействия человека с этими системами в целях решения управленческих задач. В процессе взаимодействия с информационной системой, к человеку по различным информационным каналам (с использованием различных телекоммуникационных устройств) могут поступать значительные объемы информации для переработки и последующего принятия решения. Количество каналов информации, по которым информация поступает к человеку, продолжает стремительно увеличиваться, как и

количество поступающей информации. В процессе взаимодействия, человеку приходится обрабатывать массивы текстовой и графической информации, учитывать значительное количество различных факторов, а также решать задачи многокритериального выбора. Для человеческой системы переработки информации многокритериальные задачи представляют собой особо сложный класс задач [8]. Наличие многих критериев приводит к нагрузке на человеческую систему переработки информации, заставляя человека использовать различные, зачастую оригинальные эвристики для того, чтобы решить поставленную задачу [7].

Для исследования процессов структурирования и обработки специализированной текстовой информации человеком, в процессе человеко-компьютерного взаимодействия было проведено экспериментальное исследование. В ходе исследования моделировались процессы чтения специализированной текстовой информации, обработки и анализа информации и процессы принятия решений. В рамках эксперимента, испытуемому необходимо было прочитать документ — специализированный текст, предъявленный на мониторе компьютера, и затем ответить на предложенные вопросы — принять решение о сортировке (decision making), т. е. определить департамент или подразделение организации, к которому по тематике относится прочитанный текст.

Необходимо отметить, что в состав информационной системы был включен программный модуль, реализующий функции системы поддержки принятия решений. Данный модуль осуществлял предварительную обработку текста документа, выделял опорные слова цветом (рис. 1) и визуализировал структуру документа. В процессе исследования, решения принимались испытуемыми как с помощью системы поддержки принятия решений, так и без помощи системы поддержки принятия решений. Последовательность предъявления текстов (с обработкой и без обработки) на мониторе компьютера, менялась, чтобы нивелировать привыкание испытуемого.

В ходе проведения экспериментов удалось выявить характерные особенности траектории взора испытуемых. Большинство испытуемых читали предлагаемый текст дважды (рис. 1). Причем во второй раз взор испытуемого перемещался от одного опорного слова (фрагмента текста) к другому опорному слову (фрагменту текста) в независимости были ли они выделены в процессе обработки текста или нет. В процессе движения от одного опорного слова к другому, взор испытуемого «перескакивал» через строки и абзацы, иногда возвращаясь к отдельным фрагментам текста (рис. 1). Можно предположить, что таким образом испытуемый фиксировал наиболее важные (для него) слова или фрагменты текста. Такие последовательности слов или фрагментов текста можно представить графически, см. рис. 2.

На рисунке 2 последовательности слов или фрагментов текста представлены прямоугольниками с номерами от 1 до 5. Стрелками на рисунке 2 показана траектория взора испытуемого.

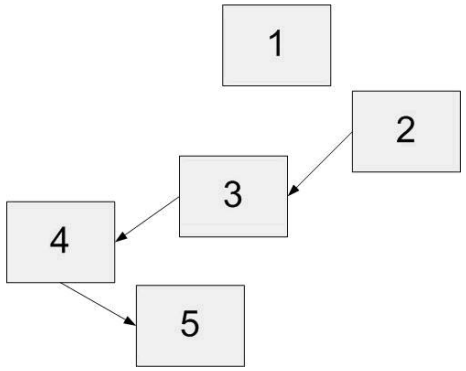


Рис. 2

3. Структурирование и визуализация

Анализируя последовательности опорных слов, полученных как в процессе предварительной обработки специализированного текста, так и выделенных в процессе чтения испытуемым и проводя последующее сравнение, определяем соответствие с имеющимися альтернативами. Визуализация полученной структуры может быть представлена в виде рис. 3. На рисунке 3 количество уровней иерархии в структуре и количество дуг сходившихся к одному узлу характеризуют структуру ментальной репрезентации испытуемого, сформировавшуюся после прочтения документа на момент принятия решения о сортировке. Дуги, сходящиеся к одному узлу на рис. 3, показывают процесс принятия решения испытуемым, т.е. выбор единственной альтернативы в процессе сортировки.

На рисунке 3 показана структура с тремя уровнями иерархии и процесс сортировки, т.е. выбор одной альтернативы ксњ из пяти возможных «a», «b», «c», «d», «e».

4. Краткие выводы

Анализ траектории движения взора испытуемого по строкам текста и длительность фиксации взгляда на определенных (опорных) словах или словосочетаниях и последующее сопоставление с когнитивным стилем испытуемого позволяет характеризовать когнитивные процессы, протекающие при считывании и анализе информации испытуемым. Действительно анализируя траекторию взора, можно с определенной долей уверенности судить о когнитивных процессах испытуемого по считыванию и анализу текстовой информации [3], [10]. Анализируя количество опорных слов или словосочетаний, а также время фиксации взгляда на опорных словах, группируя опорные слова в соответствии с имеющимися альтернативами см. рис. 1, 2 и 3, можно представить репрезентацию структуры документа,

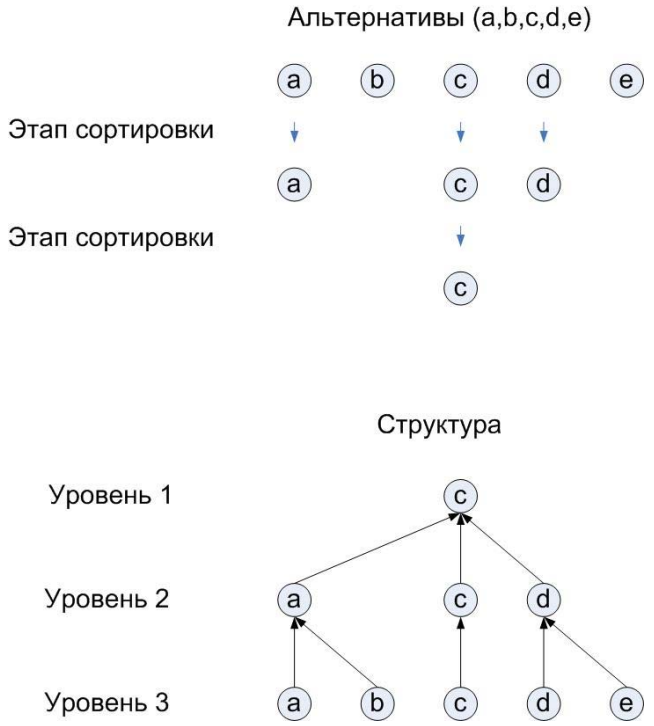


Рис. 3

сформировавшуюся у испытуемого в процессе чтения специализированной текстовой информации.

ЛИТЕРАТУРА

1. Атанасова Т., Савченко Т.Н., Головина Г.М., Баканов А.С. Интеллектуальная информационная среда обитания и субъективное восприятие качества жизни // Методы исследования психологических структур и их динамики. Труды ИП РАН. М., 2010.
2. Баканов А.С. Особенности психологического подхода к моделированию человеко-компьютерного взаимодействия // Вестник ГУУ. 2009. №6. С. 15–18.
3. Баканов А.С. Зеленова М.Е., Алдашева А.А. Когнитивные стили и эффективность работы с документацией // Сборник научных трудов SWorld. – Выпуск 2. Том 15. 2014 С. 74–78
4. Баканов А. С. Психологический подход к обработке текстовой информации // Труды Двадцать первой Международной Конференции “Крым 2014” “Библиотеки и информационные ресурсы в современном

мире науки, культуры, образования и бизнеса" ISBN 978-5-85638-177-0 Сборник трудов конференции в электронной форме.
5. Баканова Н. Б. Использование программно-технических комплексов для повышения эффективности контроля в системах документооборота //«Электросвязь». 2007. №6. С. 51–53.
6. Вишневский В.М. Теоретические основы проектирования компьютерных сетей // - М. Техносфера 2003 с. 512 ISBN: 5-94836-011-3
7. Ларичев О.И., Петровский А.Б. Системы поддержки принятия решений. Современное состояние и перспективы развития. // Итоги науки и техники. Серия Техническая кибернетика. М. ВИНИТИ, 1987. т.21, с.131–164.
8. Петровский А.Б. Многокритериальное принятие решений по противоречивыми данным: подход теории мультимножеств. // Информационные технологии и вычислительные системы, 2004, №2, 56-66.
9. Савченко Т.Н., Головина Г.М. Роль математической психологии в гуманитарном знании // Психология. Журнал Высшей школы экономики. 2014. Т. 11. №3. С. 8–22.
10. Холодная М. А. Когнитивные стили: О природе индивидуального ума. Учебное пособие – М.: ПЕР СЭ, 2002. -304 с.
11. Ташев Т., Баканова Н., Ташева Р. Исследование верхней границы пропускной способности коммутационного узла при входящем трафике типа «горячей точки». International Journal "Information Technologies & Knowledge", Vol.7, №2, 2013, pp. 182–189.
12. Шрайберг Я. Л. Доступ к библиотечно-информационным ресурсам сферы образования и науки: первые результаты Федерального проекта Министерства образования и науки РФ/ Я. Л. Шрайберг //Формирование и использование электронных ресурсов библиотек России: материалы ежегод. совещ. руководителей федер. и центр. регион. б-к России (Москва, 23-24 окт. 2012 г.). – М. : Пашков дом, 2013. – С. 88-93. ISBN 978-5-7510-0579-6

FEATURES OF PRAGMATIC DATA ANALYSIS OF TEXTUAL INFORMATION WHEN PROCESSING INFORMATION FLOWS

Bakanova N.B.
Keldysh Institute of Applied Mathematics, Moscow, Russia

Abstract

The article describes the method of pragmatic analysis of texts, which the organization received in the input stream of business documents. The method is aimed at the implementation of information support of decision-making in the allocation of work in the organization. The idea of the method is based on the method of content analysis and the creation directory of jobs in the organization and thematic orientation of departaments.

Keywords: organizational management, electronic docflow, information support of management, content analysis.

ОСОБЕННОСТИ ПРАГМАТИЧЕСКОГО АНАЛИЗА ТЕКСТОВОЙ ИНФОРМАЦИИ ПРИ ОБРАБОТКЕ ИНФОРМАЦИОННЫХ ПОТОКОВ

Н.Б. Баканова

Институт прикладной математики им. М.В. Келдыша, Москва, Россия
nina@keldysh.ru

Аннотация

В статье дается описание метода прагматического анализа текстов документов, поступающих в организацию во входном потоке деловых документов. Метод направлен на реализацию информационной поддержки принятия решений при распределении работ в организации. Идея метода базируется на методе контент-анализа и создании справочника распределения должностных полномочий в организации с указанием тематической направленности подразделений.

Ключевые слова: организационное управление, электронный документооборот, информационная поддержка управленческой деятельности, контент-анализ.

1. Введение

Активное проведение информатизации в организационных структурах существенно повлияло на управленческие процессы в методологическом

и технологическом аспекте. Потребовались новые многофункциональные системы, направленные на автоматизацию разнообразных задач организаций, расширения функциональных возможностей информационных систем в направлении поддержки управленческой деятельности, мониторинга выполнения заданий, взаимодействия с подведомственными и вышестоящими организациями. Развитие информационных технологий предоставило новые возможности в области: передачи данных, хранения и обработки информации, создания и проектирования информационных систем. С другой стороны, открывшиеся возможности усложнили процессы обмена информацией, увеличили информационные потоки, определили новые требования к реализации электронного обмена документами. В организациях появились новые требования к интерактивным режимам работы с документами, к средствам мобильного доступа к информационным ресурсам, потребовался анализ специализированных информационных потоков, включая межведомственный документооборот (МЭДО), межведомственное электронное взаимодействие (СМЭВ) [1, 2].

Перечисленные задачи информатизации в организационном управлении повысили научный интерес к проблемам создания крупных информационных систем, к исследованию управленческой деятельности, к систематизации процессов, происходящих в организации, к разработке новых методов поддержки принятия решений и анализа деятельности организаций.

В статье рассматриваются проблемы информационной поддержки одного из важнейших видов управленческой деятельности — подготовки проектов указаний по исполнению решений. Разработка средств информационной поддержки этого вида деятельности требует решения комплекса научных задач, включая поиск источников данных, реализация механизмов поддержки, анализ взаимодействия человека с информационной системой в процессе профессиональной деятельности, построение специализированных интерфейсов [3,4,8].

2. Постановка задачи

Деятельность любой управленческой организации отражается в деловых документах, прохождение и исполнение которых фиксирует документооборот организации в соответствии с нормативами документационного обеспечения управления (ДОУ) и ГОСТ Р 6.30 — 2003 (Унифицированная система организационно-распорядительной документации). Процесс обработки документов предусматривает следующие этапы работ:

- регистрация документа, выполняется при поступлении документа в систему (экспедиция, канцелярия);
- подготовка проекта указания по исполнению решений (помощники руководителя: советники, референты);
- утверждение проекта указания (руководитель);

- исполнение документа (сотрудники подразделений).

Указание по исполнению решения (резолюция на документ) сопровождает каждый документ, который передается по служебной иерархии в подразделения организации. Указание содержит текст и список исполнителей (3 — 4 участника). Первый исполнитель, отмеченный в списке, считается ответственным по всему комплексу работ.

Подготовка проекта указания по исполнению решений является важным элементом деятельности помощников, референтов и советников руководителей в управленческих организациях и выполняется на всех уровнях организационной иерархии. Формально процедура состоит в анализе поступившего документа для определения подразделений или конкретных должностных лиц в чьи полномочия входят вопросы, затронутые в документе. Работа включает определение проблематики документа, выбор подразделения — исполнителя, в чью компетенцию входит решение затронутых в документе вопросов, установление иерархии исполнителей (ответственный, соисполнитель). Выбор подразделения — исполнителя осуществляется на основании нормативов распределения должностных полномочий.

Наиболее трудоемкая часть подготовки проекта выполняется на высшем уровне обработки (начальный анализ документа). На этом уровне должны быть определены пути решения проблемы, указанной в документе, и основные участники работы (ответственные по направлениям). При большом входном потоке и значительном размере текстовой части документа указанная задача является нетривиальной. Для выявления проблем, которым посвящен документ, и подготовки проекта указания по исполнению решений требуется высокий профессионализм, хорошее знание проблемной области и сферы компетенции исполнителей данной организации.

В крупных организациях начальный анализ осложняется большим объемом входного потока документов, широтой аспектов управленческой деятельности. Можно привести некоторые данные, характеризующие поток документов крупных государственных организаций:

- объем входного потока документов — 50–100 тысяч в год;
- поступающие документы содержат до 100 страниц текста, включающего отраслевую специфику;
- сложная иерархическая структура нормативов распределения должностных полномочий в крупной организации;
- необходимость выдачи пояснений по подготовленным проектам.

Функционально указания по исполнению решений обеспечивают документированное взаимодействие и распределение работ между подразделениями организации. На рисунке 1 показаны уровни и этапы обработки документа.

На верхнем уровене обработки документ рассматривается подразделениями центрального аппарата (ЦА), далее обработка документа может быть передана нескольким подразделениям для обработки документа в

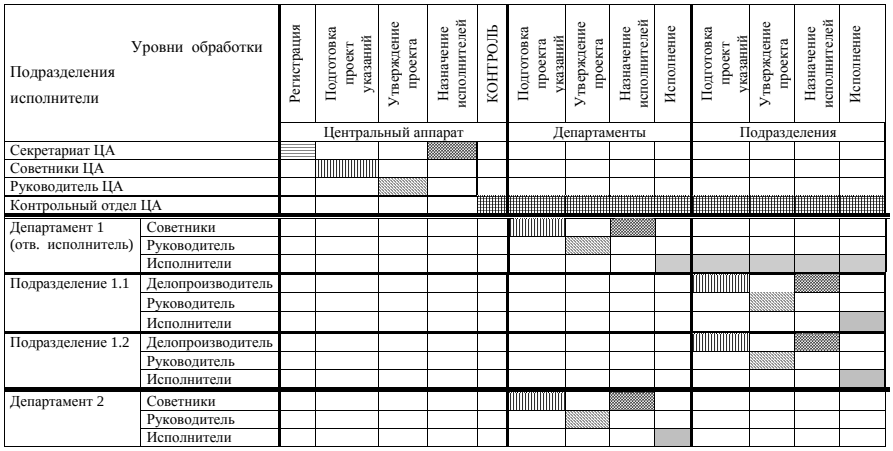


Рис. 1

рамках их компетенции. Обработка документа в подразделениях организации также выполняется по нескольким направлениям в зависимости от специализации исполнителей. На рисунке видно, что этап подготовка проекта указания присутствует на всех уровнях обработки документа. При этом на каждом уровне иерархии проводится анализе текст документа на соответствие задачам, которые входят в компетенцию исполнителей данного подразделения.

3. Принципы реализации

Для информационной поддержки этого трудоемкого процесса разработан метод прагматического анализа текста документа, поступающего на исполнение в организацию. Идея базируется на использовании метода контент-анализа, адаптированного к алгоритмической обработке текстовых массивов, и направленного поиска терминов специализированного справочника. Технологическая основа метода — подсчет встречаемости исследуемых компонентов в анализируемом информационном массиве, дополняемый выявлением статистических взаимосвязей и анализом структурных связей между ними. Для анализа статистических взаимосвязей выбран, существующий в каждой организации, справочник распределения должностных полномочий, расширенный парадигматическими отношениями, с учетом отраслевой тематики.

Разработанный метод предусматривает создание программного комплекса, обеспечивающего следующие функциональные возможности:

- создание и поддержка в актуальном состоянии специализированного справочника терминов, включающего связь терминов с подразделениями организации;

- анализ текста документа для поиска совпадений с терминами справочника и формирование схемы совпадений по направлениям компетенции подразделений организации;
- анализ схемы совпадений и определение частоты встречаемости терминов, относящихся к конкретным подразделениям организации;
- интерфейс программного комплекса, реализованный в соответствии с поставленной задачей и отвечающий ментальным репрезентациям сотрудников при подготовке проектов указаний, для эффективного использования разработки;
- передача подготовленного проекта в систему документооборота для продолжения обработки документа.

Для обеспечения требуемой функциональности программного комплекса необходимо решение следующих технологических задач:

- предварительная подготовка текстовых массивов поступающих документов для анализа на совпадение с множеством понятийно — тематических единиц, специализированного справочника;
- создание специализированного формата для объединения результатов анализа понятийно-тематических единиц;
- разработка алгоритма поиска данных на соответствие терминам специализированного тезауруса;
- представление результата анализа для реализации возможности получения объяснений по каждому назначенному исполнителю;
- разработка алгоритма подготовки проектов указаний по исполнению решений в соответствии с нормативами ДОУ;
- разработка средств взаимодействия программного комплекса с системой документооборота.

4. Основные решения по реализации метода

Для анализа текста выбран формат Rich Text Format (RTF), позволяющий работать с документами, как полученными в результате электронного обмена, так и поступившими в традиционном (бумажном) виде. В результате анализа текста должны быть сформированы весовые коэффициенты, показывающие количество совпадений по департаментам с учетом основных направлений и терминов расширения этих направлений. Разработанная структура хранения данных позволяет в режиме корректировки выделять цветом все термины, относящиеся к конкретному исполнителю, что обеспечивает последующее объяснение выполненного анализа [4].

Структура хранения данных представляет таблицу, которая создается при переводе документа в формат RTF. Структура таблицы включает следующие элементы:

- «идентификатор документа» — связывает карту с номером зарегистрированного в системе документооборота поступившего документа;

- «департамент» — используется для указания отношения найденного совпадения к направлениям деятельности конкретного департамента;
- «вид термина» — используется для указания группы термина (основные направления или расширенные).

Данные о местоположении найденных совпадений сохраняются в виде отношения:

$$M =< t, p, c, n, l, \ldots >,$$

где

t — основной код термина в справочнике;
r — код расширения термина в справочнике;
p — страница документа;
c — строка на странице;
n — начальная позиция на строке;
l — количество позиций для выделения;
$\ldots$ другие служебные признаки.

В силу специфики задачи анализу подвергается сравнительно небольшой текст отраслевой направленности (до 100 страниц), поэтому для поиска совпадений терминов с текстом документа использован метод «широкого поиска», отличающийся большим количеством результатов по сравнению с методом «узкого поиска». Наличие нерелевантной информации корректируется специалистами при подготовке проекта указаний по исполнению документа.

Разработанный интерфейс обеспечивает представление результата в виде списка исполнителей, для каждого из которых указано количество терминов, найденных в тексте документа. Индикация количества совпадений подчеркивается усилением фона цветового индикатора для данного исполнителя. Для объяснения и подтверждения полученных результатов, при выборе конкретного исполнителя выбранные термины выделяются в тексте документа цветом, и показывается результат подсчета встречаемости терминов. Разработка интерфейса проводилась с учетом исследования ментальных репрезентаций при подготовке проектов управленческих решений [5, 6].

5. Заключение

Метод прагматического анализа текстовой информации при обработке документов входного потока является новой разработкой, предназначенной для информационной поддержки принятия решений в процессе подготовки проектов управленческих решений. Структура метода демонстрирует возможности нового подхода к использованию известных методов в условиях развития информационных технологий.

Выполненная разработка показывает, что исследования процессов управленческой деятельности и информационного потенциала систем, поддерживающих эту деятельность, открывают новые возможности совершенствования управленческой деятельности [7].

6. Благодарности

Работа выполнена при поддержке гранта РФФИ 14-07-00216а «Разработка принципов реализации информационной поддержки принятия управленческих решений, на базе информационного потенциала систем организационного управления».

ЛИТЕРАТУРА

1. Баканова Н. Б., Вишневский В. М. Моделирование процесса движения документов в корпоративных системах документооборота // Автоматика и телемеханика. — 2008. — №9. — С. 183—189
2. Баканова Н. Б. Проблемы реализации электронного взаимодействия в распределенных системах документооборота // Электросвязь. — 2013. — №10. — С. 43—45.
3. Алдашева А.А., Баканов А.С. Зеленова М.Е. Когнитивные стили и эффективность работы с документацией // Сборник научных трудов SWorld. — Выпуск 2. Том 15. 2014 С. 74-78.
4. Баканова Н. Б. Информационная поддержка подготовки проектов управленческих решений по организационной деятельности // Information Technologies & Knowledge: International Journal / ed. Kr. Markov ; Institute of Information Theories and Applications FOI ITHEA. — Sofia: Institute of Information Theories and Applications FOI ITHEA. — 2014. — Vol. 8. — №2. — Pp. 119—123.
5. Баканов А.С. Психологический подход к обработке текстовой информации // Труды Двадцать первой Международной Конференции «Крым 2014» «Библиотеки и информационные ресурсы в современном мире науки, культуры, образования и бизнеса» ISBN 978-5-85638-177-0 Сборник трудов конференции в электронной форме.
6. Баканов А.С. Особенности психологического подхода к моделированию человеко-компьютерного взаимодействия // Вестник ГУУ. 2009. №6. С. 15—1.
7. Петровский А.Б. Теория принятия решений. — М.: Издательский центр «Академия», 2009.
8. Ташев Т., Баканова Н., Ташева Р. Исследование верхней границы пропускной способности коммутационного узла при входящем трафике типа «горячей точки». International Journal «Information Technologies & Knowledge», Vol.7, №2, 2013, pp. 182-189.

REASONING OF THE TRANSITION TO POLYMODAL INFOCOMMUNICATIONAL SYSTEMS

O. Basov
Academy of FGS of Russia
oobasov@mail.ru

Abstract

Multimodality of traditional interpersonal communications points to the purposefulness of using the polymodal dialogue in the process of communicative interaction of infocommunications subscribers. The creation of the polymodal communication systems became possible due to the development of cognitive science and current results in the design area of multimodal interfaces interaction. Application of the existing and expected outcome of signal processing tasks of different modalities in the synthesis of polymodal systems will provide all parties to communicate and their further intellectualization will allow us to approach the infocommunicational interaction of subscribers to the traditional interpersonal communication.

Keywords: Interpersonal communication, communication channel, infocommunication system, multimodal interface, polymodal system

1. Introduction

Interpersonal communication (communication) – a system of transfer and exchange of information between the representatives of the society – is a complex, multifaceted and multi-part phenomenon. People can exchange information at different levels of abstraction, and the communication is not limited to oral or written communications. An important role in the communicative process is played by the personality characteristics of interlocutors, their mood, and their physiological and psycho-emotional conditions that must be taken into account in business communication [1]. According to anthropologists and ethologists the information, conveyed in words, is only about 7% of the total amount of information obtained by a person, whereas non-verbal signals have to 93% (facial expressions, postures, gestures, touches, smells up to 55%, and the share of voice paralinguistic component accounts for up to 38%). At the average a person speaks for only 10-12 minutes a day, with a verbal component accounts for only 35% of meaning and nonverbal much more – 65% [2]. These data require the necessity of interpersonal communication structure analysis, classification of the channels and means of communication of information and their correlation with existing and prospective infocommunication systems.

2. Projection of interpersonal communication on existing infocommunication systems

Development of tools for effective communication between people through technology systems is now one of the priority directions of the development of artificial intelligence and computer science in general. This is because even now the technical tools are not fully used because of the lack of a full-fledged, habitual interface for a subscriber to interact with the hardware-software means of communication. Today, the majority of telecommunication systems provide a very limited way of interaction: voice input using directional microphone, recording a low quality image using video cameras, printing by using the keyboard, handwriting using touch screens, management of virtual objects with the mouse cursor, the display of visual information in the form of text and images on the monitor screen and mono or stereo audio playback. At that in each application their communication channels are used, realizing only communicative (exchange of information) side of communication. For example, telephone communication speech and acoustic channels are involved, but the passed information is not shared [3]. During a video conference verbal (including text), acoustic and visual communication channels are used, but there are no tactile and olfactory channels, at that the data sent by textual communication channel, is not synchronized with verbal information. Such methods of communication force subscribers to adapt to the means of communication and learn the virtual method of communication. In the process of its extensive development on the way of implementing multiservice (speech, video and data transfer) telecommunication systems have evolved into *infocommunication*, providing in some degree an interactive side of communication (act exchange). However, in recent years there has been a trend of depersonalization - subscriber often requires one type of service under the name *network connection*, implying the possibility of getting affordable or most convenient way of interaction, determined in accordance with disabilities and individual preferences, on the basis of the context of the communicative act.

3. Multimodal character of interpersonal communication

For the decision of global problem of communicative interaction between people through the technical systems is necessary to use additional means of information transfer, implementing multi-modal interaction. Here modality should be understood as belonging of the reflected stimulus to certain sensory system - part of the human nervous system, responsible for the perception of certain signals from the environment or internal environment [4, 5]. In the framework of solving the concrete problem of human-computer interaction *multimodal interfaces* [6, 7] are developed, typical for interpersonal communication. They allow providing an effective and natural to human interaction with various automatic means of control and communication. In systems that are implemented on the basis of multimodal interfaces, information from the verbal and non-verbal communication channels is continuously processed, cre-

ating a real or virtual environment, that allows to satisfy the user's needs, and quickly adapt to the context. Some combinations of modalities for transferring information are well suited to certain situations and application tasks, but worse or even entirely inapplicable to others. Possibility of choice of modality by user is an important feature of multimodal interfaces. Currently [6]] multimodal interfaces are already used abroad in some application areas: the cartographic systems, virtual reality systems, medical systems, robotics, web applications, etc. In Russia scientific researches in this area have started recently, and their successful implementation is complicated by the fact that it is necessary to combine the efforts of various research groups engaged in separate processing of speech, video, handwriting, etc. in various scientific research institutes. Since 2003 the group of linguistic informatics SPIIRAN conducts fundamental and applied works on multimodal interfaces.

4. Interpersonal communication in the framework of polymodalcommunication systems

Taking into account the established fact of interpersonal communication and existing interest in multimodal interfaces we should expect that the subscribers of infocommunication systems will want to use the polymodal dialogue in the process of communicative interaction (fig. 1). The possibility of creating the polymodal communication systems due to the fact that cognitive science that studies the mechanisms of human perception and interaction provided the fundamental information for modeling the subscriber's behavior, as well as information about how the multimodal architectures should be organized [8].

The transition from multimodal to polymodal systems is associated with a complex character of interaction of separate modalities (table) and, while developing multimodal interfaces the problems associated with the synchronization, joint processing and the integration of the multimodal input of information are solved, in polymodal systems there are additionally problems associated with the division and transfer of such information and its representation of output modalities.

Input modalities' signals (task group I) are analyzed and coded in accordance with the defined channels of communication, thereby realizing communicative and partly interactive sides of communication like traditional information and communication systems [1]. However, a comparison of data processing technologies with the tasks associated with the recognition of the information transmitted by text (task II.1), acoustic (task II.2) and visual (tasks III.3) channels, indicates complete compliance processed on the same communication channels of the modalities. So text channel's signals (modality 1 and 2) are used in encoding (task I.1), and by the recognition (task II.1) [9, 10]. Speech (modality 4) and nonvoice (modality 6) signals and pauses (modality 5) are served as input data in the existing methods and algorithms of coding (objective I.2) [3] and speech recognition (objective II.2) [11, 12]. Lips articulation (modality 3) acts as an object of analysis in the problems of lip reading

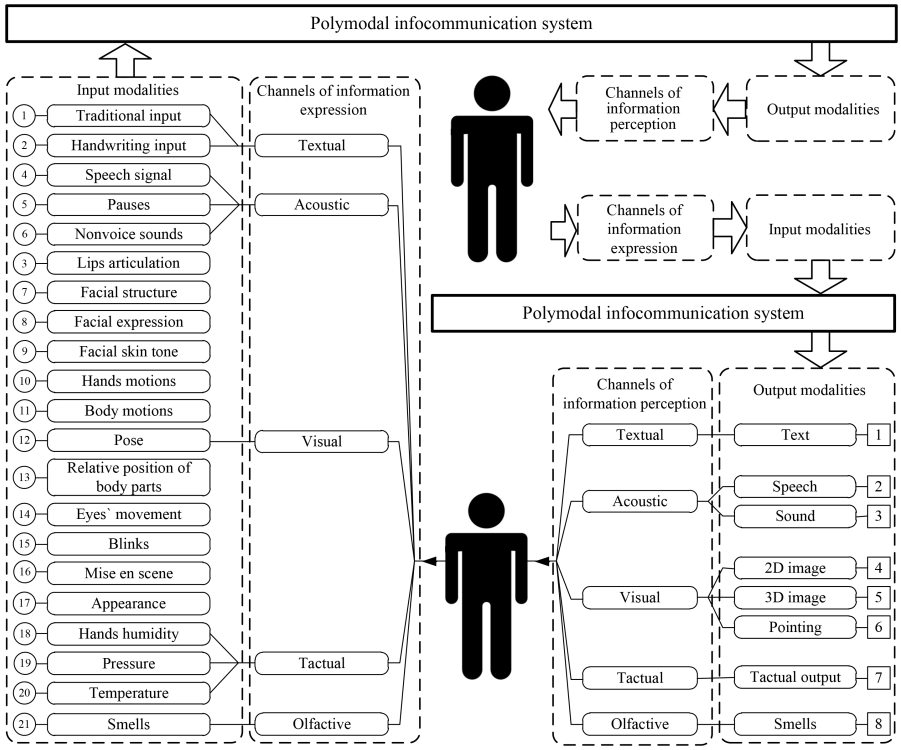


Fig. 1: Communicative interaction in polymodal communication systems.

(task II.3), on the other hand, the results of studies [13, 14] point to the advisability of its integration at the level of the parametric representation with duplicate information on speech communication channel. A similar conclusion is fair and for the speech recognition methods – joint processing of multimodal information (modality 3–6) allows to increase their effectiveness [15]. Other visual modalities (7-17) are separately investigated in problems of faces' recognition, facial expressions, gestures and determine the position of the human body and its sight (tasks II.3) and their solutions are widely used in systems of computer vision and recognition of sign languages, navigation and cartography, medicine [16, 17]. At the same time, the information received through the visual channel of communication, should be coded and transmitted by infocommunication (task I.3). In the traditional infocommunication systems it is transmitted in the form of mobile low-level (as the element of some mathematical space, for example, in the form of an array of brightness) images [18], while the analysis of individual visual modalities (3, 7-17) in polymodal systems allows the use the structural methods of presentation of visual information on the basis of contour or similar descriptions [19]. Integration at the

Tasks		Input modalities	Processing technologies	Output modalities
I	I.1	1–2	Message coding	1
	I.2	4–6	Speech coding	2, 5
	I.3	3, 7–17	Image coding	4, 5
	I.4	18–20	Tactual signals coding	7
	I.5	21	Chemical analysis and synthesis	8
II	II.1	1–2	Determining message semantics	1
	II.2	4–6	Speech recognition	1, 4, 5
	II.3	3	Lip reading	1, 5
		3, 7–9, 13	Face recognition	1, 2, 4, 5
		3, 7, 8	Facial expression recognition	
		12–15	Determining message semantics	4–6
		11–13	Determining of pose	2, 4, 5, 6
		10–13	Determining of gesture	1, 2, 5, 6
III	III.1	1–15, 17–21	Subscriber identification	1–4
	III.2		Determining of physiological condition	1, 2, 4, 7
			Determining psychoemotional condition	
IV	IV.1	1–6, 8–15, 18–20	Assessment of truth send of information	1–4, 7
	IV.2	1-21	Subscriber‘s intentions forecastinh	1–3, 7

Table 1: Processing technologies of certain modalities within the framework of polymodal communication systems.

level of making decision semantically different information transmitted through different communication channels, creates preconditions for formulating higher level tasks, for example, identification of subscribers (task III.1), including the conditions of departure from their physiological and/or emotional condition from normal (tasks III.2). Application of existing and expected decisions in the synthesis of polymodal communication systems will provide the perceptive side of communication (each other communication partners' perception), and their further intellectualization (solution of problems of group IV) will bring the infocommunication subscribers' interaction to the traditional interpersonal communication (fig. 3).

5. Conclusion

In the subject field theories there are two universally recognized approaches to the presentation of information in the infocommunication systems that are based on the division of the transmitted information on services, as well as poly-

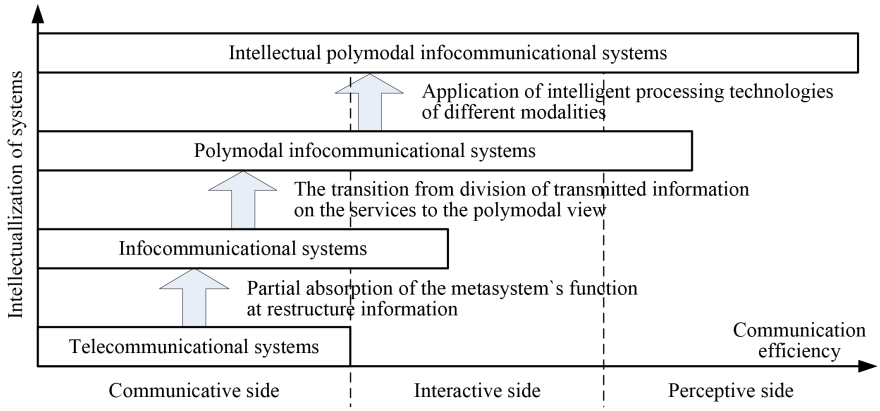


Fig. 2: Rise of communication efficiency.

modal presentation of information. The first of them is "rooted" in existing infocommunications, however, as the analysis of interpersonal communication shows, does not provide the required communication and efficiency, by and large, has no further development. The cause of this situation is persisting in the scientific and technical community attitude towards infocommunication system as a distributed system that implements the functions of receiving, processing, and transmission of data recovery. The consequence of this system is the increase of capacity (and, as a consequence, the cost of the whole system) due to consistent and independent development of applied modalities (manual input, speech signal, the image) when processing and data transfer. The second approach is widely used in purely informational technologies and has enough good results from their use, creates the suppositions to building infocommunication systems with "perceptual" functions. The need to improve the effectiveness of communicative interaction is caused by: 1) transformation of formal role in business communication, whereby together with the exchange of information should be taken into account the features of the subscriber's identity, his mood, physiological and psycho-emotional state; 2) increasing psychological stress of taking managerial decisions associated with the reduction of quotas trust communicating to each other; 3) increased speed of change of the situation and the growth of the transmitted information, requiring subscribers to increase the impact of their actions. The abandoning of the traditional principles of the transmitted information separation on services in favor of the polymodal information presentation will require the development of a new, strict, but at the same time, constructive theory of building polymodal (including intellectual) communication systems, which would allow to judge from unified methodological positions, to assess the current state of affairs in the subject area, to explore the proposed new solutions for building (synthesis) of such systems, as well as

to justify proposals for their optimization taking into account the specifics of the functioning.

REFERENCES

1. Basov O.O. Prerequisites of the Polymodal Infocommunicational Systems // Sbornik materialov Mezhdunarodnojnauchnomodal infocommunicational systems construction – The Compilation of the International research and practice conference materials. "NedeljanaukiSPbGPU". St-Petersburg state poly-technical university, December 3-8 2012. // Section "Solution of the difficult problems in sphere of modern information and computer technologies". 5–6 pp.
2. Butovskaja M.L. Bodylanguage: Nature and Culture (Evolutionary and cross-cultural basics of the human non-verbal communication. M.: Nauchnyjmir, 2004. 440 p.
3. Sheluhin O.I., Luk'jancev N.F.Digital processing and transmission of speech/ edited by O. I. Sheluhina. M.: Radio i svjaz', 2000. 456 p.
4. Ostrovskij M.A., Shevelev I.A. Human Physiology in two volumes. Student's Book. vol. II) / edited by. V.M. Pokrovsky, G.F. Korot'ko. 201–259p.
5. Handverker H.: Human Physiology in 3 volumes. Vol.1.translated from English. / edited by R. Shmidtai G. Tevsa (translation edited by. P.G. Kostjuk). M.: Mir, 1996. p. 178–196.
6. Oviatt S.L. Multimodal Interactive Maps: Designing for Human Performance // Human Computer Interaction. Special Issue on Multimodal Interfaces. Vol. 12, 1997. P. 93–129.
7. Ronzhin A.L., Karpov A.A. Multimodal interface: basic principles, cognitive aspects. // Tr. SPIIRAN. SPb.: Nauka, 2006. Issue. 3, vol. 1. S. 300–319.
8. Ekman P., Friesen W.V. The repertoire of non-verbal behavior: categories, origins, usage and coding // Semiotica. 1, 1969. P.49–98.
9. Juan A., et al. Integrated Handwriting Recognition and Interpretation via Finite-State Models // Technical Report ITI-ITE-01/1. Institut Tecnologic d'Informatica. Valencia (Spain), July 2001.
10. Mankoff J., Abowd G.D. Error Correction Techniques for Handwriting, Speech, and other ambiguous or error prone systems // GVU TechReport GIT-GVU-99-18. June, 1999.
11. Rabiner L.R., Juang B.H. Fundamentals of speech recognition. New Jersey: Prentice Hall PTR Englewood Cliffs, 1993. 507 p.
12. Kipyatkova I.S., Karpov A.A.Analitical review of the Russian speech recognition systems with wide vocabulary // Tr. SPIIRAN. SPb.: Nauka, 2010. Issue. 1. P. 7–20.
13. Chan T. HMM-based audio-visual speech recognition integrating geometric and appearance-based visual features // IEEE 2nd Workshop on Multimedia Signal Processing. – Oct. 3–5, 2001. P. 9–14.

14. Matthews I., et al. Lipreading Using Shape, Shading and Scale. School of Information Systems, University of East Anglia, Norwich, 1999.
15. Sascha F. Audiovisual speech: analysis, synthesis, perception and recognition // 16th International Congress of Phonetic Science. Saarbrucken, August 2007. P. 275–278.
16. Aggarwal J.K., Cai Q. Human motion analysis: review // Comput. Vis. Image Understanding. Vol. 73, 1999. P. 428–440.
17. Fasel B., et al. Automatic Facial Expression Analysis: survey // Pattern Recognition. 36, 2003. P.259-275.
18. Pinz A. Interpretation and fusion recognition versus reconstruction. Vision Milestones, OGAI lecture series, 1995. P. 9-21.
19. Rares A., Reinders M.J.T., Hendriks E.A. Image Interpretation Systems // Technical Report (MCCWS 2.1.1.3.C), MCCWS project, Information and Communication Theory Group. TU Delft, 1999. 32 p.

SYNTHESIS OF LINEAR MODULATORS AND DEMODULATORS FOR PHYSICAL LAYER TECHNOLOGY

K. Batenkov
Academy of Federal Guard Service, Orel, Russia

Abstract

Task solution of linear modulator and demodulator for linear filter channel with additive noise in accordance with minimal average squared error is taken. Obtained solution technical effect estimation in case of two-position amplitude modulation is given. It shown that energy gain (about 3 dB) relatively widely known modal modulation is represented.

СИНТЕЗ ЛИНЕЙНЫХ МОДУЛЯТОРОВ И ДЕМОДУЛЯТОРОВ ДЛЯ ТЕХНОЛОГИЙ ФИЗИЧЕСКОГО УРОВНЯ

К. Батенков
Академия Федеральной службы охраны Российской Федерации,
Орел, Россия
pustur@yandex.ru

Аннотация

Получено решение задачи синтеза линейного модулятора и демодулятора для линейного фильтрового канала связи с аддитивным шумом по критерию минимальной среднеквадратической ошибки. Проведена оценка технического эффекта полученных решений для случая передачи одномерных двухпозиционных амплитудно- модулированных сигналов, показавшая наличие энергетического выигрыша (порядка 3 дБ) относительно широко используемой модальной модуляции.

Ключевые слова: физический уровень, модулятор, демодулятор, критерий минимальной среднеквадратической ошибки, линейный фильтровой канал связи с аддитивным шумом

1. Введение

Известно, что непосредственно сами процедуры как модуляции так и модуляции в теоретических исследованиях сводятся к их заданию в форме

обобщенных рядов Фурье, причем в целях упрощения дальнейших математических выкладок базисы выбираются в виде ортогональных функций. При этом сама форма данных функций не конкретизируется, поскольку в дальнейшем для получения правил решения находится предел при стремлении числа подобных функций к бесконечности, а в конечном правиле непосредственно базисные функции не фигурируют. В практических же приложениях формы базисных функций оказываются существенными, так как они определяют формы сигналов передаваемых в непрерывный канал связи, а также функций корреляторов на приеме. Так, известно довольно обширное число работ, посвященных синтезу несущих колебаний (и корреляторов) для определенных типов непрерывных каналов связи. Для канала связи с аддитивным белым гауссовским шумом оптимальным оказывается гармонический базис, для канала ограниченного и по частоте и по времени – функции Слепяна, для канала с окрашенным шумом и неравномерной частотной характеристикой – базисные функции разложения Карунена–Лоэва, реализующие так называемую модальную модуляцию [1]. Однако следует подчеркнуть, что в качестве критерия синтеза всех упомянутых функций служит независимость формируемых при этом подканалов, что не всегда приводит к оптимальности по другим критериям, таким как минимум среднеквадратической ошибки или вероятности ошибки [2].

С позиции же вычислительной сложности осуществляемых процедур несомненным преимуществом обладает именно гармонический базис, являющийся непрерывным аналогом дискретного базиса в широко используемых процедурах быстрого прямого и обратного преобразования Фурье. В то же время сам по себе гармонический базис является оптимальным для задач синтеза модулятора и демодулятора только в случае идеализированной модели канала – модели с аддитивным белым гауссовским шумом и прямоугольной амплитудно-частотной характеристикой канала. В результате стремление к снижению вычислительной сложности цифровой обработки сигналов на основе быстрых алгоритмов Фурье зачастую приводит к существенному снижению количественных и качественных показателей передаваемой по каналу связи информации. Именно поэтому реализация операций модуляции и демодуляции на основе их описания в виде ряда Вольтерра, являющегося обобщением ряда Фурье, с последующим представлением результирующих ядер Вольтерра в дискретной форме, пригодной для реализации на современных процессорах цифровой обработки сигналов, с одной стороны позволяет реализовать, пусть и с некоторым увеличением требований к производительности процессоров, алгоритмы модуляции и демодуляции на современной элементной базе, а с другой обеспечивает их согласованность со стохастическими характеристиками непрерывного канала связи, поскольку ядра Вольтерра получаются решением задачи синтеза, а не выбираются произвольно, как в случае гармонического базиса.

Таким образом, основной задачей данной статьи является разработка линейных процедур модуляции и демодуляции для линейного фильтрового

непрерывного канала связи с аддитивным шумом, являющихся оптимальными по критерию минимума среднеквадратической ошибки и реализуемых с использованием процессоров цифровой обработки сигналов. Актуальность же данной задачи следует из постоянного стремления современных систем связи к повышению качества и количества передаваемой информации в условиях ограниченного ресурса непрерывных каналов связи, а также недостаточной проработанностью теории модуляции и демодуляции применительно к оптимальному согласованию стохастических свойств произвольных линейных фильтровых каналов связи с аддитивным шумом с непосредственно процедурами модуляции и демодуляции.

2. Постановка задачи

Общая постановка задачи формулируется следующим образом. В модели дискретного канала связи, имеющей форму:

$$x \xrightarrow{\Phi} x(t) \xrightarrow{\mathrm{H}} x'(t') \xrightarrow{\Phi'} x', \tag{1}$$

где $\mathbf{x}$ – сигнал (вектор координат точек сигнального созвездия) на входе модулятора (на входе дискретного канала связи); $x(t)$ – сигнал на входе непрерывного канала связи (на выходе модулятора); $x'(t')$ – сигнал на выходе непрерывного канала связи (на входе демодулятора); $\mathbf{x}$' – сигнал (вектор) на выходе демодулятора (на выходе дискретного канала связи; Φ – оператор модуляции; H – оператор непрерывного канала связи; Φ' – оператор демодуляции;

необходимо определить оптимальные операторы модуляции Φ и демодуляции Φ' по критерию минимума среднеквадратической ошибки между сигналами на входе модулятора $\mathbf{x}$ и выходе демодулятора $\mathbf{x}$' на основе известных свойств линейного непрерывного канала связи H, а также стохастических характеристик сигнала на входе модулятора $\mathbf{x}$. Следует подчеркнуть, что модулятор и демодулятор осуществляют преобразование непрерывного канала связи в дискретный и, по сути, задают операцию дискретного отображения непрерывного канала.

Таким образом, задача синтеза детерминированного дискретного отображения непрерывного канала связи заключается в минимизации среднеквадратической ошибки между сигналами на входе и на выходе образуемого дискретного канала:

$$\bar{\sigma}^2 \to \min_{\mathbf{\Phi}_i, \mathbf{\Phi}'_j}, i = \overline{1, N_a}, j = \overline{1, N_b}. \tag{2}$$

при ограничении на энергию e_x передаваемых сигналов [3]:

$$\sum_{i=1}^{N_a} \sum_{j=1}^{N_a} \left(\mathbf{\Phi}_i \left\{ i+1 \middle| j+1 \right\} \mathbf{\Phi}_j\right) \left\{ 1, ..., i+j \middle| 1, ..., i+j \right\} M_{x,i+j} \le e_x, \tag{3}$$

где $\mathbf{\Phi}_i = \{\varphi_{k_1,\ldots,k_i,j}\}_{k_1,\ldots,k_i=\overline{1,N},\ j=\overline{1,\infty}}$ – $(i+1)$-мерная матрица переменного порядка коэффициентов разложения базисных функций модуляции $\varphi_{k_1,\ldots,k_i,j}$; $\mathbf{\Phi'}_i = \left\{\varphi'_{i,k,k_1,\ldots,k_i}\right\}_{k=\overline{1,N'},\ k_1,\ldots,k_i=\overline{1,\infty}}$ – $(i+1)$-мерная матрица переменного порядка коэффициентов разложения базисных функций демодуляции $\varphi'_{i,k,k_1,\ldots,k_i}$; N – размерность сигналов на входе модулятора (входе дискретного канала связи); N' – размерность сигналов на выходе демодулятора (выходе дискретного канала связи); N_a – степень нелинейности модулятора; N_b – степень нелинейности демодулятора; $\delta(x)$ – многомерная дельта-функция; $\mathrm{tr}\mathbf{A}$ – след матрицы $\mathbf{A}$; $M_{x',i,x,j}$ – матрица совместных моментов $(i+j)$-го порядка сигналов x' и x.

Таким образом, задача синтеза дискретного отображения непрерывного канала связи для детерминированных операторов модуляции и демодуляции по критерию минимума среднеквадратической ошибки (2), (3) классифицируется как задача нелинейного программирования с нелинейным ограничением в виде неравенства.

3. Решение задачи в общем виде

В настоящей работе в качестве модели непрерывного канала рассматривается модель линейного фильтрового канала связи с аддитивным шумом, для которой сигнал на выходе $x'(t')$ (входе демодулятора) определяется как сумма свертки сигнала на входе $x(t)$ (выходе модулятора) и импульсной характеристики канала $h(t - t')$ и некоторого шума $n(t')$ [4]:

$$x'(t') = \int_t h(t - t')\, x(t)\, dt + n(t'). \tag{4}$$

При этом для линейных операторов модуляции и демодуляции, представимых в форме классических матриц размерности два, решение рассматриваемой задачи существует в явном виде, что позволяет достаточно просто указать нижнюю границу показателя качества для общего нелинейного случая.

Так, в данной ситуации $N_a = N_b = 1$, то выражение для оптимальной матрицы демодуляции (линейного оператора демодуляции) получается путем соответствующего дифференцирования:

$$\mathbf{\Phi}'_1 = \left(M_{n,2} + H^T \mathbf{\Phi}_1 M_{x,2} \mathbf{\Phi}_1^T H\right)^{-1} H^T \mathbf{\Phi}_1 M_{x,2}, \tag{5}$$

где H – матрица коэффициентов разложения импульсной характеристики; $M_{n,2}$ –матрица вторых начальных моментов аддитивного шума в канале связи.

Составление лагранжиана: с последующим его дифференцированием по матрице модуляции $\mathbf{\Phi}_1$ согласно правил дифференцирования скалярной функции матричного аргумента [5, 6] позволяет получить в явном виде линейные дискретные отображения непрерывных линейных фильтровых каналов связи с аддитивным шумом, оптимальные по критерию минимума

среднеквадратической ошибки, в форме матриц демодуляции (5) и модуляции (линейного оператора модуляции):

$$\mathbf{\Phi}_1 = \left(H^{-1}\right)^T M_{n,2}^{1/2} \bar{U}_n^T \bar{U}_x V^{1/2} V, \tag{6}$$

где $\bar{U}_n$ – матрица размером $N_c \times N$, составленная из собственных векторов матрицы $M_{n,2}^{1/2} H^{-1} \left(H^{-1}\right)^T M_{n,2}^{1/2}$, соответствующих ее минимальным собственным числам; $\bar{U}_x$ – ортогональная матрица в каноническом разложении матрицы $V^{1/2} V M_{x,2} V^T \left(V^{1/2}\right)^T$; V – ортогональная матрица из канонического разложения матрицы $\frac{1}{\sqrt{-a}} U N^{1/2} U^{-1} M_{x,2}^{-1/2} - M_{x,2}^{-1}$; U – неособенная матрица, преобразующая матрицу $-a \left(M_{x,2}^{-1} + \mathbf{\Phi}_1^T H M_{n,2}^{-1} H^T \mathbf{\Phi}_1\right)^2 M_{x,2}$ к канонической жордановой форме $N = -aU^{-1} \left(M_{x,2}^{-1} + \mathbf{\Phi}_1^T H M_{n,2}^{-1} H^T \mathbf{\Phi}_1\right)^2 \times$ $\times M_{x,2} U$; a – множитель Лагранжа, вычисляемый на основе предположения о значении среднеквадратической ошибки: $a = -\left[\bar{\sigma}^2 / \mathrm{tr}\left(M^{1/2} N^{-1/2}\right)\right]^2$.

При этом в результате вычисления матриц модуляции (6) и демодуляции (5) рассчитывается величина энергии сигнала на выходе модулятора в форме:

$$E_x = \mathrm{tr}\left(\bar{U}_n^d \bar{U}_x^d\right), \tag{7}$$

где $\bar{U}_n^d$ – усеченная до размера $N \times N$ диагональная матрица, полученная из диагональной матрицы в каноническом разложении матрицы $M_{n,2}^{1/2} H^{-1} \left(H^{-1}\right)^T M_{n,2}^{1/2}$; $\bar{U}_x^d$ – диагональная матрица в каноническом разложении матрицы $V^{1/2} V M_{x,2} V^T \left(V^{1/2}\right)^T$.

4. Оценка технического эффекта

С целью оценки технического эффекта от применения полученных линейных операторов модуляции и демодуляции проведено вычисление матриц модуляции и демодуляции для линейного фильтрового канала связи с аддитивным гауссовским шумом с использованием пакета MathCad. В качестве канала связи рассматривался канал с импульсной характеристикой идеального фильтра высоких частот в диапазоне от 1 кГц до 100 МГц и аддитивным гауссовским шумом с неравномерной спектральной плотностью мощности. При этом рассматривалось три типа шумовых сценариев. В первом дисперсия шума по каждому из измерений σ_n^2 варьировалась в диапазоне от 0,9 мкВ^2 до 1 мкВ^2, во втором – от 0,5 мкВ^2 до 1 мкВ^2, а в третьем – от 0,25 мкВ^2 до 1 мкВ^2 (единицы измерения приведены из расчета на один подканал). Длительность тактового (символьного [7]) интервала составляет 100 мкс при количестве отсчетов (коэффициентов разложения базисных функций модуляции и демодуляции) равном 128.

В качестве аналога использовались несущие, описанные в [1] и представляющие собой собственные колебания автокорреляционной функции

канала (базисные функции разложения Карунена–Лоэва), позволяющие представить исходный аналоговый канал в виде совокупности независимых подканалов. Подобный тип модуляции в литературе именуется модальной, а его базисом является ортонормальный набор собственных функций канала связи, имеющих наибольшие собственные числа [1]. При демодуляции использовался оптимальный линейный демодулятор, полученный в настоящей работе.

Вычисленные зависимости среднеквадратической ошибки $\bar{\sigma}^2$ от отношения сигнал-шум $\gamma = \frac{E_x}{\sum_n \sigma_n^2}$ в канале связи для случая передачи одномерных двухпозиционных амплитудно-модулированных сигналов (сигнальные точки являются противоположными и равновероятными [7]) представлены на рисунке 1.

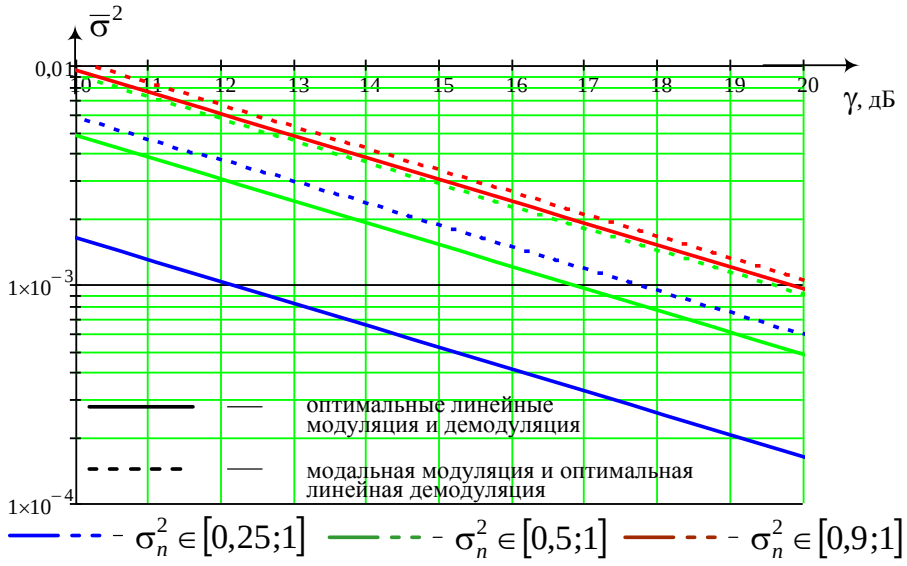


Рис. 1: Зависимость среднеквадратической ошибки $\bar{\sigma}^2$ от отношения сигнал-шум γ при оптимальной линейной и модальной модуляции в условиях оптимальной линейной демодуляции для случая передачи одномерных двухпозиционных амплитудно-модулированных сигналов

Следует отметить, что разность между принятым сигналом на выходе демодулятора и поданным на вход модулятора, по сути, является реализацией интегральной помехи (всех аддитивных шумов) на выходе демодулятора. Следовательно среднеквадратическое отклонение эквивалентно энергии помехи, содержащейся в сигнале на выходе демодулятора. В результате для заданной энергии передаваемых сигналов целесообразно рассматривать не просто среднеквадратическое отклонение как энергию неустранен-

ной помехи, а отношение энергии полезного сигнала к помехе (отношение сигнал–помеха $\gamma' = E_x/\bar{\sigma}^2$). Кроме того, при подобной трактовке показателя технического эффекта появляется возможность прогнозировать помехоустойчивость последующих схем обработки (при переходе от непрерывного выхода к дискретному), поскольку для них наиболее часто исследуемыми являются зависимости достоверности именно от отношения сигнал–помеха, а не от просто энергии вредной составляющей.

В работе были получены зависимости отношения сигнал–помеха на выходе демодулятора γ' от отношения сигнал-шум γ, представленные на рисунке 2.

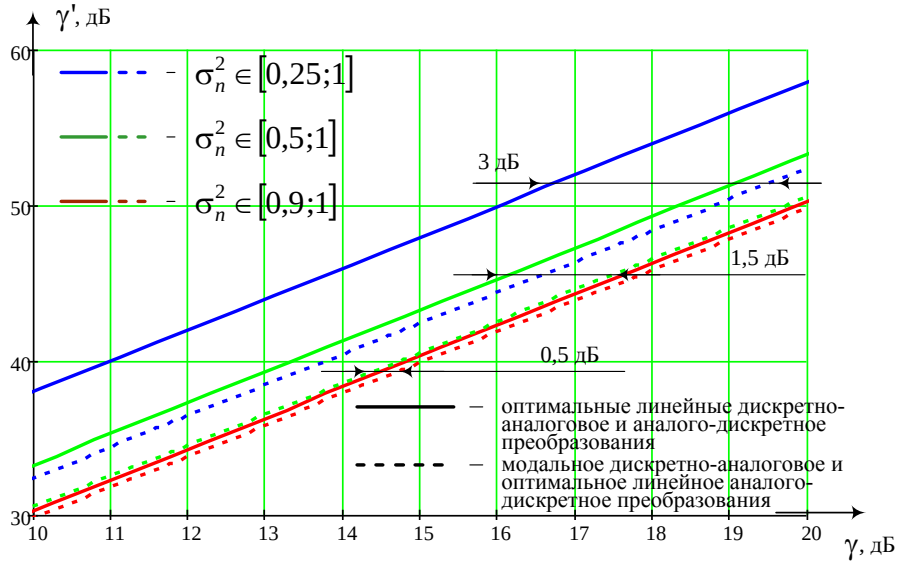


Рис. 2: Зависимость отношения сигнал–помеха на выходе демодулятора γ' от отношения сигнал-шум γ при оптимальной линейной и модальной модуляции в условиях оптимальной линейной демодуляции для случая передачи одномерных двухпозиционных амплитудно-модулированных сигналов

5. Заключение

Таким обрзом, с ростом неравномерности дисперсии аддитивного шума снижается величина среднеквадратической ошибки, что вызвано, прежде всего, появлением измерений с низкими относительно других значениями дисперсий шума, использование которых и позволяет повысить технический эффект. При этом следует заметить, что модальная модуляция приводит к существенно большей среднеквадратической ошибки при одном и том же отношении сигнал-шум по сравнению с оптимальной линейной модуля-

цией, предложенной в настоящей работе. Это связано в первую очередь с тем, что при модальной модуляции несущие преобразуют исходный аналоговый канал связи в совокупность независимых подканалов и, по сути, пытаются полностью устранить соканальную помеху, жертвуя при этом отношением коэффициента передачи каждой из базисных функций по каналу связи к дисперсии шума. В случае же оптимальной линейной модуляции полное уничтожение соканальной помехи не производится, а получаемые решения, позволяют найти компромисс между максимально допустимой величиной отношения коэффициента передачи к дисперсии шума и минимально допустимым значением соканальной помехи. Следует заметить, что в случае исследуемых одномерных сигналов как таковой соканальной помехи нет и оптимальным решением является несущая с максимальным отношением коэффициента передачи к дисперсии шума.

ЛИТЕРАТУРА

1. Advanced Digital Communications. Classic EE379 Series Courses / John M. Cioffi ... [et al.]. – Department of Electrical Engineering, Stanford University. – URL: http://www.stanford.edu/group/cioffi/book/chap4.pdf. Дата обращения: 02.10.2013.
2. Батенков К. А. Математические модели модулятора и демодулятора с заданным порядком нелинейности // Цифровая обработка сигналов. – 2013. – є 1. – С. 14–21.
3. Батенков К. А. Обобщенный пространственно-матричный вид энергетических ограничений систем связи // Известия Тульского государственного университета. Технические науки. – 2013. – є 3. – С. 238–245.
4. Батенков К. А. Математическое моделирование непрерывных многопараметрических каналов связи в операторной форме // Телекоммуникации. – 2013. – є 10. – С. 2–4.
5. Магнус Я. Р. Матричное дифференциальное исчисление с приложениями к статистике и эконометрике / Я. Р. Магнус, Х. Нейдеккер ; пер. с англ. под ред. С. А. Айвазяна. – М. : Физматлит, 2002. – 496 с.
6. Амосов А. А. Скалярно-матричное дифференцирование и его приложение к конструктивным задачам теории связи / А. А. Амосов, В. В. Колпаков // Проблемы передачи информации. 1972. Т. VIII. Вып. 1. С. 3–15.
7. Прокис Дж. Цифровая связь / Дж. Прокис ; пер. с англ. под ред. Д. Д. Кловского. – М. : Радио и связь, 2000. – 800 с.

REDUNDANT DISTRIBUTION OF REQUESTS ACROSS THE NETWORK BY TRANSFERRING THEM OVER MULTIPLE PATHS

V.A. Bogatyrev[1], *S.A. Parshutina*[2]
ITMO University, Saint-Petersburg, Russia
[1]vladimir.bogatyrev@gmail.com, [2]svetlana.parshutina@gmail.com

Abstract

The research focuses on the opportunities offered by multipath routing for providing reliability and fault-tolerance of distributed computing systems, when requests are distributed dynamically across a given network. Models for non-redundant and redundant search for an available server, which is ready to serve arriving requests, are proposed. A server may be unavailable due to its faults, temporary shutdown, or overload with requests arriving from the network. An unavailable server rejects incoming requests, which then underlie the flow of repeated requests, or those sent to other servers, time and again, until they are finally accepted. Transmitting requests repeatedly and polling several servers simultaneously in case of redundant search cause the increase in the network load. It is found that searching for an available server over multiple paths at once results in the decreased requests' time in the system, provided that server availability and the intensity of request flow are not high.

Keywords: reliability, cluster, request, redundancy, routing, queuing systems

1. Introduction

Reliability and fault-tolerance of distributed computing systems are achieved owing to the use of redundant computing and communication resources, when requests are distributed adaptively among computing nodes. Distribution of requests, which are transmitted across a given network in this study, can be combined with routing, i.e. the choice of the route that meets specified requirements for the current dataflow. In case of single-path routing, requests are sent over a single path, *the best* one from a certain point of view. In case of multipath routing, there exists a set of appropriate paths from the sender – the source of requests – to their receiver and each request is transmitted over several paths simultaneously.

Contemporary models and methods for providing reliability of distributed computing systems are expected to rest upon the notion of Quality of Service (QoS) [1]. Those methods have already been widely discussed, for example in [2-8], but the opportunities offered in this regard by dynamic distribution of requests across the network, on the basis of multipath routing, remain understudied for today.

This research paper discusses how to provide reliability of distributed computing systems using redundant requests, distributed across the network over multiple paths. Two ways of transmitting requests by polling servers, which are possible receivers of the requests, in a serial and in a simultaneous manner are proposed and underlie the corresponding models. The models take into account the fact that the network load may increase because of simultaneous transmission of requests over several routes and the existence of the flow of repeated (unserved) requests – the result of polling unavailable servers.

2. Problem Statement and System Design

The focus of the study is considering two alternative ways of distributing requests across the network based on multipath routing. The simplest case of disjoint routes is examined; each path is a single-channel non-preemptive M/M/1 queuing system with the infinite queue [9, 10]. The probability of loss of requests and the probability p of server availability are included in the models in question.

A server is available and, as a result, ready to serve requests if it is in working order, powered up, and not overloaded with other requests coming from the network. A request will be served or, if the server is busy, wait on its queue before entering service and the sender of the request will be notified of success. Otherwise, the request will be rejected by the server and sent to another one over some different path and the sender will receive a notice of refusal. Servers are supposed to be accessed repeatedly in a random or a Round-Robin fashion until the success is achieved.

The general scheme of distribution of requests, taking into account the flow of the rejected ones and hence sent to other servers repeatedly, is shown in Fig.1. In this paper, transmitting a request over one of all appropriate paths at a time is called serial search for available servers while sending it over multiple paths simultaneously is called parallel (redundant) search. Time in the system, or the period of time a request resides in the system, starting from its initial distribution until its sender receives the notification of acceptance, is chosen as the criterion of efficiency of the systems under consideration. All calculations are carried out using Mathcad 14.

2.1. Non-redundant distribution of requests across the network. Non-redundant distribution of requests implies transmitting a request over one appropriate route at once. The intensity of request flow Λ_0 is determined by the intensity of the initial request flow Λ, comprised of requests first entering the system, and the intensity of the flow of repeated requests, rejected by unavailable servers and returned to the system. The intensity Λ_0 is calculated as

$$\Lambda_0(\Lambda) = \Lambda p \sum_{i=1}^{\infty} i(1-p)^{i-1} = \frac{\Lambda}{p}, \tag{1}$$

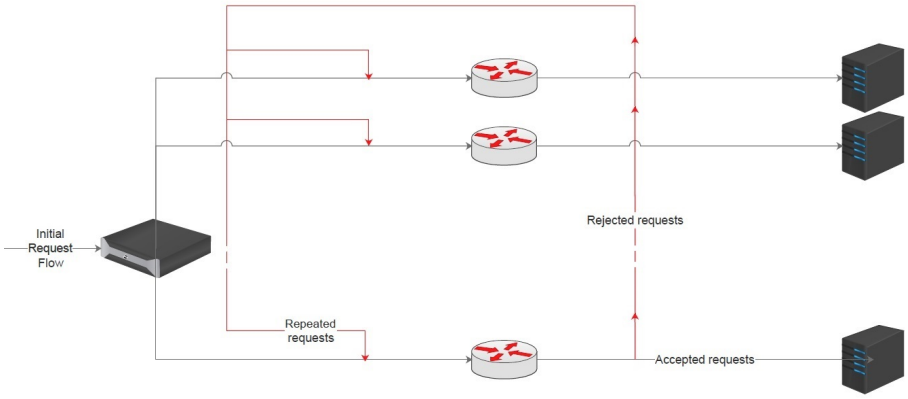


Fig. 1: The general scheme of searching for an available server

where p is the probability of server availability, $i(1-p)^{i-1}$ is the expected value of the number of repeated efforts to distribute a request across the network.

Time delay arising from each of those efforts, when using (1), is represented by

$$T(\Lambda) = \frac{v}{1 - \frac{(\Lambda_0(\Lambda)+\alpha\Lambda)v}{n}} = \frac{vnp}{p(n+\alpha v\Lambda) - v\Lambda}. \tag{2}$$

Here v is the average time of request transmission across the network, without taking into account waiting on queue, t is the time of response from the given server (available or not), n is the number of possible routes from the source to one of the servers, α is the share of requests coming from other sources.

Time in the system, or average residence time – $T(\Lambda)$ is replaced with (2) – is

$$T_0(\Lambda) = p\sum_{i=1}^{\infty} i(1-p)^{i-1}(T(\Lambda)+t) = \frac{vt\Lambda + p[\alpha vt\Lambda - n(v+t)]}{p[v\Lambda + p(\alpha v\Lambda - n)]}. \tag{3}$$

2.2. Redundant distribution of requests across the network. Requests are distributed across the network in a redundant manner when each of them is sent to k servers over k routes simultaneously. The probability r saying that at least one server is available is

$$r = 1 - (1-p)^k. \tag{4}$$

The intensity of request flow Λ_1, when (4) substitutes for r, is

$$\Lambda_1(\Lambda) = \Lambda r\sum_{i=1}^{\infty} i(1-r)^{i-1} = \frac{\Lambda}{1-(1-p)^k}. \tag{5}$$

Time delay arising from each effort to deliver a request, granting that Λ_1 can be replaced with (5) is

$$T(\Lambda) = \frac{v}{1 - \frac{(k\Lambda_1(\Lambda)+\alpha\Lambda)v}{n}} = \frac{vnr}{r(n + \alpha v\Lambda) - kv\Lambda}. \quad (6)$$

Time in the system is calculated based on (6) as follows:

$$T_1(\Lambda) = r\sum_{i=1}^{\infty} i(1-r)^{i-1}(T(\Lambda) + t) = \frac{kvt\Lambda + r[\alpha vt\Lambda - n(v+t)]}{r[kv\Lambda + r(\alpha v\Lambda - n)]}. \quad (7)$$

3. Calculations and Evaluation

The proposed ways of distributing requests across the network are to be evaluated by comparing them for efficiency, described in terms of time in the system. For that purpose, some values should be assigned to the parameters of the models under consideration, like the following ones: *v = 1 s*, *t = 0.1 s*, *n = 10* routes as well as servers, *α = 0*. The redundancy order k varies from 2 to 10 routes while server availability p ranges between 0 and 1.

Fig.2 illustrates an example calculation of time in the system T, which depends on the intensity of request flow Λ and is computed based on non-redundant distribution of requests, redundant distribution when the redundancy order k equals to 2, and redundant distribution when k equals to 3, in case server availability *p = 0.95* and *p = 0.85*. It is clear that the decrease of server availability from 0.95 to 0.85 leads to the increase of time in the system T in any case; nevertheless, the change of T is more significant in case of non-redundant distribution of requests. Thus, the non-redundant system is more preferable than the redundant ones when *p = 0.95* and $\Lambda >= 0.5$ 1/s but is less attractive as compared with them when *p = 0.85* and $\Lambda < 0.5$ 1/s. It should also be noted that the redundant system with the redundancy order *k = 3* produces worse results than with *k = 2* in case *p = 0.95* but proves to be better than the latter in case *p = 0.85* and $\Lambda < 0.25$ 1/s.

To conclude, the simultaneous use of several routes is pointless in case of high server availability but, with its value decreasing, the efficiency of redundant search starts to increase. However, this will continue until a certain point (the threshold value), which is represented by the intersection of the functions of non-redundant and redundant search in Fig.2. On exceeding the threshold value, the intensity of the flow of primary requests combined with the flow of repeated requests, all sent over multiple paths, causes the dramatic rise of time in the system. Moreover, the lower server availability is, the greater the number of routes that are needed to find an available server in the shortest time is.

4. Conclusion

In this paper, we propose models for fault-tolerant distributed computing systems with adaptive distribution of requests across the network. The models

rest upon multipath routing and describe the ways of polling servers: one at a time in case of serial (non-redundant) search and several servers at once in case of parallel (redundant) search. Servers are supposed to be available, i.e. ready for serving requests, or unavailable due to their possible faults and temporary shutdown. Time in the system, or average residence time of requests in the system, takes into account the flow of repeated requests – previously rejected by unavailable servers – and is chosen as the criterion of efficiency of the systems in question.

It is found that non-redundant search for an available server, when each request is sent over only one appropriate path at a time while the others are regarded to be reserve, turns out to be the best option if server availability is high. The less server availability is, the more preferable redundant search is, when each request is being transmitted over multiple paths simultaneously. However, it is true only until the intensity of request flow is less than a certain threshold value. Increasing the number of routes used to search for an available server (the redundancy order) gives little advantage in case server availability is high but proves significant if the availability is decreasing.

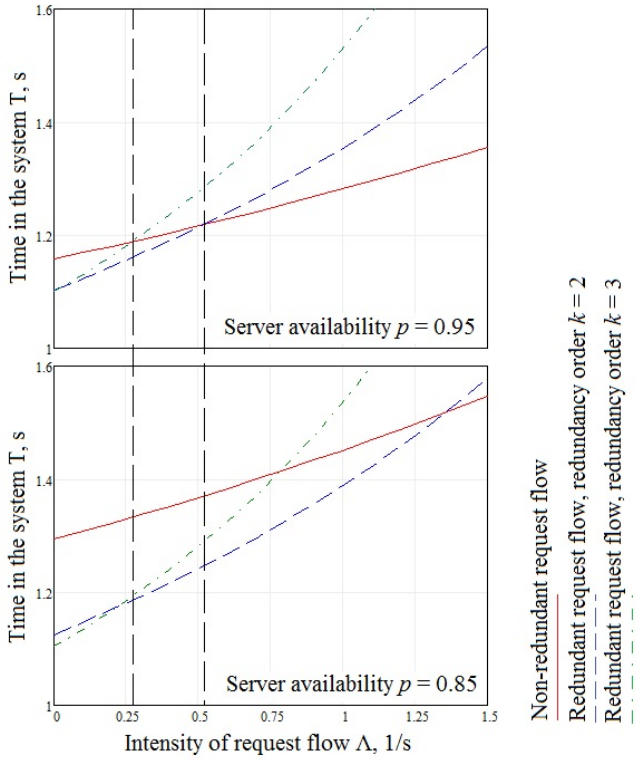


Fig. 2: Time in the system, or residence time of requests in the system

REFERENCES

1. McDysan D. QoS and Traffic Management in IP and ATM Networks. McGraw-Hill, 2000.
2. Andreev S., Saffer Z., Turlikov A. Delay Analysis of Wireless Broadband Networks with Non Real-Time Traffic // Lecture Notes in Computer Science. 2011. V. 6886. P. 206-217.
3. Bogatyrev V.A. Exchange of Duplicated Computing Complexes in Fault tolerant Systems // Automatic Control and Computer Sciences. – 2011. – V. 46. – N 5. – pp. 268–276.
4. Bogatyrev V.A , Bogatyrev S.V., Golubev I. Y. Optimization and the Process of Task Distribution between Computer System Clusters // Automatic Control and Computer Sciences. - 2012. - N 3. - pp. 103-111. DOI: 10.3103/S0146411612030029.
5. Bogatyrev V.A Fault Tolerance of Clusters Configurations with Direct Connection of Storage Devices // Automatic Control and Computer Sciences, 2011, Vol. 45, No. 6, pp. 330–337. DOI: 10.3103/S0146411611060046.
6. Bogatyrev V.A An interval signal method of dynamic interrupt handling with load balancing, Automatic Control and Computer Sciences , 34(6), 2000, pp. 51-57.
7. Bogatyrev V.A. Protocols for dynamic distribution of requests through a bus with variablelogic ring for reception authority transfer // Automatic Control and Computer Sciences. 33(1), 1999, pp. 57-63.
8. 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. Vol. 49. No. 1. Pp. 46-56. DOI 10.3103/S0146411615010022.
9. Vishnevsky V.M. Teoreticheskie osnovy proektirovaniya komp'uternykh setey [Theoretical Fundamentals for Design of Computer Networks]. Moscow, Tekhnosfera, 2003 (in Russian).
10. Aliev T.I. Osnovy modelirovaniya diskretnykh system [Fundamentals of simulation of discrete systems]. St. Petersburg, SPbSU ITMO Publ., 2009, 363 p. (in Russian).

RFID WITH OPTICAL INTERFACE

Borodula V., Maklakov V.
V.A.Trapeznikov Institute of Control Sciences of RAS,
Moscow, Russia

Abstract

Considered possibility building RFID system with optical communication interface between reader and tag. The solution allows to expand the range of applications of RFID systems, particularly for jewelry and small objects.

Keywords: RFID, optical communication channel , reader, tag

RFID С ОПТИЧЕСКИМ ИНТЕРФЕЙСОМ

В. Бородуля[1] *, В. Маклаков*[2]
[1,2] Институт проблем управления им. В.А. Трапезникова РАН, Москва
[1] borodula@mail.ru, [2] vvmaklakov@mail.ru

Аннотация

Рассматривается возможность создания RFID-системы с оптическим каналом связи между ридером и меткой в рамках стандартных протоколов . Рассмотрен вариант построения LF RFID с меткой пассивного типа. Предложенное решение позволяет расширить спектр применения RFID систем, в частности проводить идентификацию мелких объектов, объектов удаленных на значительные расстояния, объектов находящихся в прозрачных жидких средах.

Ключевые слова: RFID, оптический канал, ридер, метка

Введение

RFID технологии представляет собой способ удаленной идентификации объектов при котором на объекте устанавливается электронная метка содержащая уникальный код (идентификатор) который может быть дистанционно считан специальным устройством — ридером. Ридер работает на фиксированных частотах и постоянно излучает электромагнитную энергию в окружающее пространство. Большинство применяемых меток являются пассивными, т.е. не имеющих собственного источника питания и получающих энергию от излучения ридера. Метка передает свой уникальный код путем модуляции несущей частоты ридера.

1. Традиционная структура RFID

Структура типовой RFID - системы представлена на Рис. 1. Основу конструкции составляет резонансный контур и микросхема-транспондер, содержащая идентификационный код. Транспондеры существуют в нескольких конфигурациях, отличаются протоколом кодирования и скоростью передачи данных. Основными характеристиками системы являются:

- Несущая частота: 125 кГц, 13,56 МГц, 433 МГц, обычно с амплитудной модуляцией
- Метод кодирования: манчестерский код, двухфазная или фазовая манипуляция
- Скорость передачи: 1, 2, 4 кбит/с или более
- «Вшитый» идентификатор, подсчет контрольных сумм

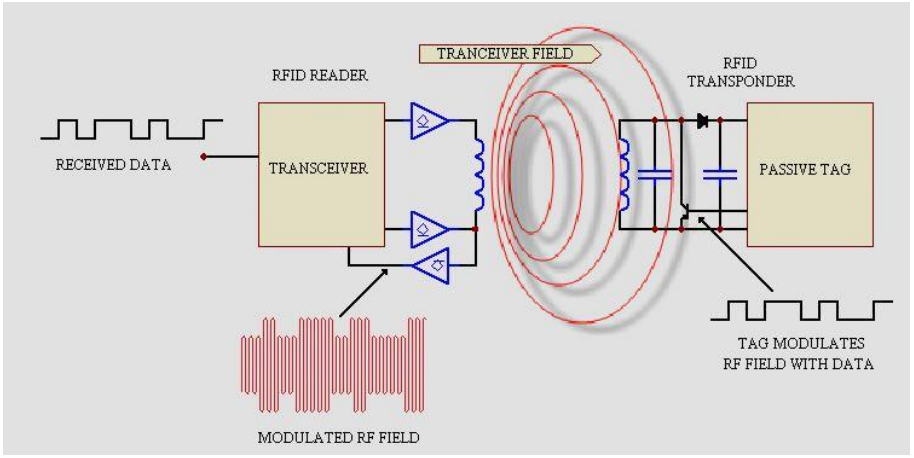


Рис. 1: Структура типовой RFID - системы

Ридер на фиксированных частотах постоянно излучает электромагнитную энергию в окружающее пространство. Попадая в поле ридера, входной контур метки резонансно возбуждается и через диод заряжает входной конденсатор. При достижении определенного уровня напряжения транспондер включается и выдает идентификационный код на базу транзистора, который, согласно кодовой последовательности, осуществляет модуляцию добротности входного контура. Входной контур индуктивно связан с приемной катушкой ридера и передает на него код.

2. Проблемы

Наряду с достоинствами RFID системам присущи и недостатки

- Ограничение по дальности применения (десятки метров)

- Невозможность применения технологии к миниатюрным объектам (ювелирные изделия, медицина и т.д.) в силу значительных размеров используемых антенн
- Подверженность помехам в виде электромагнитных полей.

3. Решение

Для расширения возможностей применения RFID технологий в работе предлагается новый подход и схемное решения - использование оптического интерфейса взаимодействия между ридером и меткой, при сохранении стандартных протоколов.

Блок-схема RFID с оптическим интерфейсом представлена на Рис. 2.

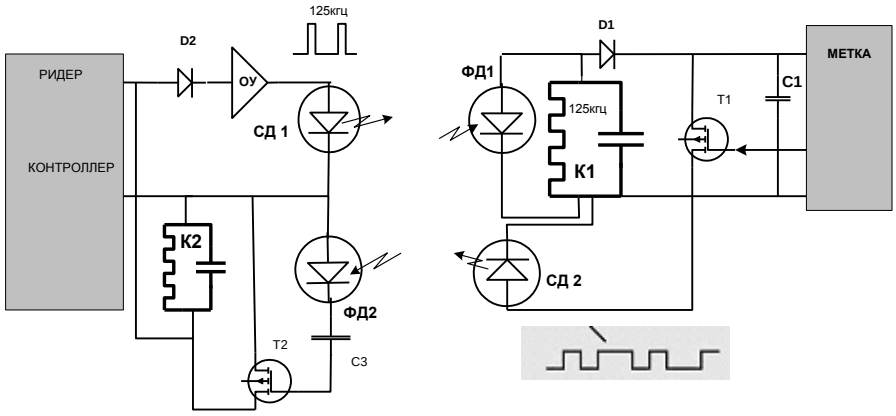


Рис. 2: Блок-схема RFID с оптическим интерфейсом

Сигнал несущей ридера, через диод D2 и усилитель-ограничитель ОУ, в виде импульсной последовательности частотой 125 кгц подается на светодиод СД1. Излучение диода регистрируется фотодиодом ФД1 включенным параллельно входному контуру К1 . В результате «оптической» раскачки контура, резонансное напряжение, через диод D1 заряжает входной конденсатор С1 так, что при достижении определенного порога происходит включение транспондера. В результате идентификационный код подается на затвор транзистора Т1 и управляет работой светодиода СД2. Излученный СД2 свет попадает на фотодиодом ФД2 и через транзистор Т2, по уже вышеописанной схеме , модулирует добротность контура К2, подключенного к входной цепи ридера, чем обеспечиваются стандартные условия декодирования сигнала метки.

Работа системы возможна в условиях прямой видимости между ридером и меткой.

В качестве пассивных меток использовались RFID производства Mikron (Россия), в качестве фото\свето диодов — ARPL-FL1J7 (длина волны 940 нм).

ЛИТЕРАТУРА

1. К. Финкенцеллер. Справочник по RFID. Теоретические основы и практическое применение индуктивных радиоустройств, траспондеров и бесконтактных чип-карт // Изд-во «Додэка XXI», 2008. 496 с.
2. «Как выбрать правильную RFID систему: пошаговое руководство», RFID Journal, 2011, 37 с.
3. В. Л. Дшхунян, В. Ф. Шаньгин. Электронная идентификация. Бесконтактные электронные идентификаторы и смарт-карты // АСТ, НТ Пресс, 2004, 696 с.

SPACE SIMULATION WIRELESS BROADBAND NETWORK IN ELECTROMAGNETIC INTERFERENCE, AND OBSTRUCTIONS

Victor M. Churkov
Moscow State Institute of Electronics and Mathematics Higher School of Economics (MIEM HSE), Moscow, Russia
wchur@mail.ru

Abstract

The paper presents a mathematical model, the technology, tools and simulation results broadband heterogeneous network under different broadband noise spectrum, the intensity of natural and fabricated obstacles. The features of the joint use of existing and emerging technologies of broadband networks in various ranges of the radio spectrum.

Keywords: Mathematical model, wireless networks and modeling tools, broadband networks, electromagnetic interference, system interference, natural and artificial obstacles, the radio frequency spectrum, the coverage area.

1. The spatial mathematical model of the service area of a wireless broadband network.

Famous figures CINR and RSSI measure the quality of the radio signal in the client equipment. CINR (Carrier to Interference + Noise Ratio) - the ratio of carrier signal level to noise level is used in telecommunications for assessing the quality of the signal and connection.

If the signal quality is above a certain standard for the norm. RSSI (Received Signal Strength Indication) - the power level of the received signal can be used to approximate the signal quality estimate.

Indicator CINR measured periodically in the channel and radio signal receiver is transmitted source for analysis and decision-making on how to connect and uniquely defines the client's network speed.

Thus, CINR is an integral indicator of the quality of the radio signal, taking into account:

natural and artificial obstacles; internal and external noise, electromagnetic interference; dynamics of change in all the factors of time and space.

It is obvious that the use of the index CINR fundamentally necessary as you would a broadband network and a qualitative design and development of its infrastructure, taking into account the relative position coordinate effectively and to minimize the number of elements.

External factors - the desired signal level, noise, and electromagnetic disturbance largely determine the bandwidth of the radio channel, which can

significantly reduce the data transmission rate and, consequently, the quality of the traffic [1].

Spatial model of the wireless network comprises a matrix parameter CINR, consider the quality of the radio as a function of given above influencing factors defining network traffic at any point.

Formation of the matrix takes place based on the results of an experimental study of the parameters of the radio signal CINR in the network coverage area.

Mathematical model [2] of the spatial organization of the coverage of wireless broadband network comprises a plurality of functions, variables and constants which:U - universal set of elements of the system wireless network (coverage):

S – A lot of system coverage subsets (regions);
C – The set of centers of areas;
R – Radius of the area;
(x, y, z) – Spatial coordinates;
$u(x, y, z)$, $c(x, y, z)$, $s(x, y, z)$ – Elements of the set;
m, n – Cardinality of the set;
$\rho(x, y, z)$ – An indicator of the effectiveness of the radio signal coverage.

Let $U, C, S, R, (x, y, z), \rho(x, y, z)$ — variable mathematical model:

$U, R, (x, y, z), \rho(x, y, z)$ – Asked variables;
C, S – Expanded variables.

Let the sets U, C, S associated functional dependency.

We introduce additional features $f, \phi, \psi, |\nu_u|$ specifying constraints and performance indicators for coverage, operating:

f – Computation $C \subset U$ and subject to the limitations of the selected performance criteria (for example $|S_j| \to \max \to |U|$);
ϕ – Calculation of the elements subsets U_i;
ψ – Calculation of the subset $U_j \setminus U_{j+1}$;
$|\overrightarrow{\nu_u}|$ – Calculation of magnitude of the vector in the coordinates U;

Express $f, \phi, \psi, |\nu_u|$ through the original variables:

$$\begin{cases} f\Big[\phi\big(u, |\overrightarrow{\nu_u(x, y, z)}|, R, \rho(x, y, z), \psi(U)\big)\Big]; \\ \phi\Big(u, |\overrightarrow{\nu_u(x, y, z)}|, R, \rho(x, y, z), \psi(U)\Big); \\ |\overrightarrow{\nu_u(x, y, z)}|; \\ \psi(U); \end{cases}$$

Then the desired sets C, S can be written as a function of variables $U, R, (x, y, z), \rho(x, y, z), f, \phi, \nu_u, \psi$; following system of equations.

$$\begin{cases} C = \Big\{u : f\big[\phi(u, |\overrightarrow{\nu_u(x, y, z)}|, R, \rho(x, y, z), \psi(U))\big]\Big\}; \\ S = \Big\{u : \phi(c, |\overrightarrow{\nu_u(x, y, z)}|, R, \rho(x, y, z), \psi(U))\Big\}. \end{cases} \quad (1)$$

We introduce the function ξ convolution common to the mathematical description of the sets C, S. Function $\xi = (|\overrightarrow{\nu_u(x,y,z)}|, R, \rho(x,y,z), \psi(U))$ calculates the occurrence of an element u_i in the set $\psi(U)$ with parameters R for $\rho(x,y,z)$ all elements $\psi(U)$. Then the system (1) can be minimized.

$$\begin{cases} C = \Big\{u : f\big[\phi(u,\xi)\big]\Big\}; \\ S = \Big\{u : \phi(c,\xi)\Big\}. \end{cases} \tag{2}$$

Thus, the system (2) is a generalized mathematical model of the spatial organization of wireless coverage. The model provides the calculation of coverage areas and centers for wireless network infrastructure based on defined performance criteria and constraints.

2. Simulation of electromagnetic parameters

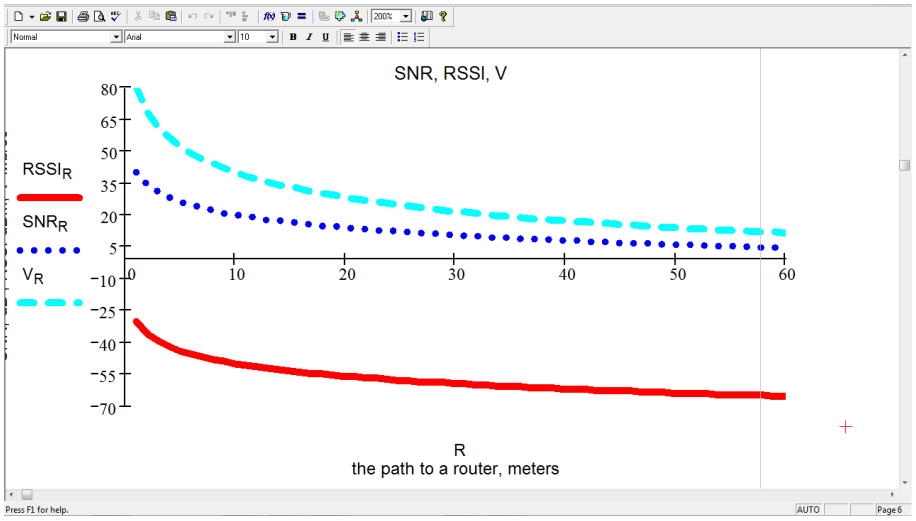


Fig. 1: Distribution of SNR and network speed under nominal and operating noise.

Fig. 1 - Fig. 3 - to select the network technology and suitable models of equipment;

Safety tasks;

anti-interference;

determining speed limits, covering radius cells;

frequency of training WDS mode;

Ensure the sustainability of the network traffic to fluctuations and other external and internal loads; ensure scalability;

Modeling parameters of the latest and advanced networking technologies to address issues of development, forecasting and analysis of fundamental solutions.

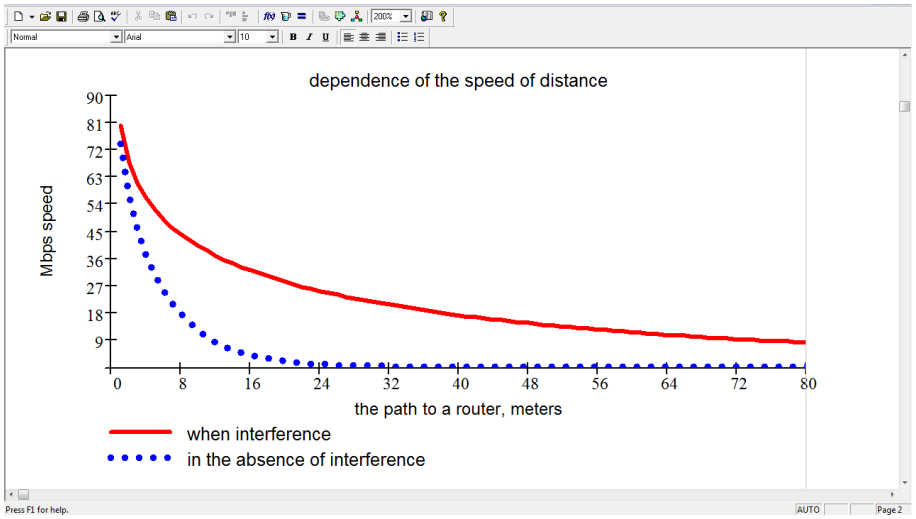


Fig. 2: Comparison of the speed of the network at different levels of interference at different distances from the access point.

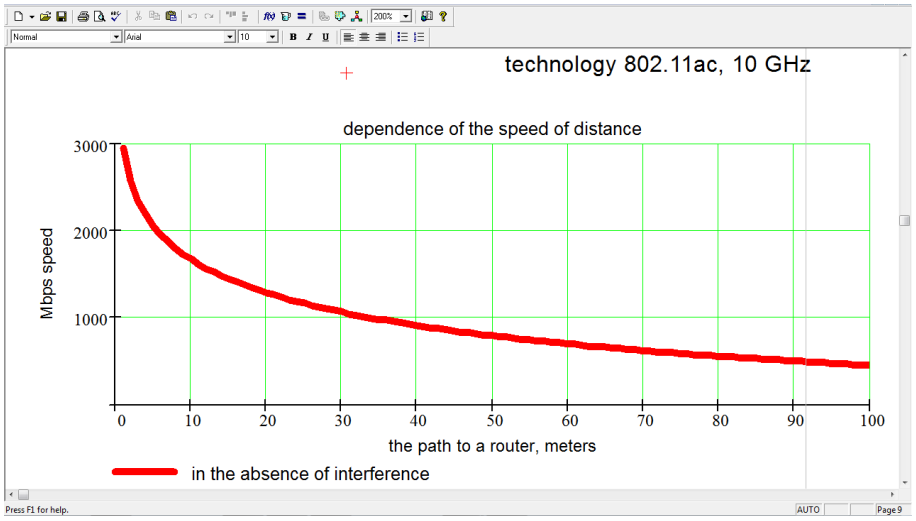


Fig. 3: The speed of the network when the standard 802.11n 5 GHz, standard 802.11ac 10 GHz.

3. Spatial modeling networks

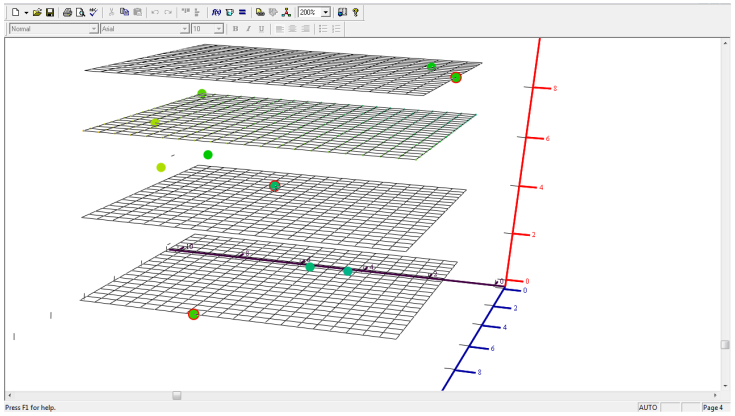


Fig. 4: Result of 3D modeling space coverage heterogeneous network.

Spatial modeling networks

Fig. 4 - for the formation of a mathematical model of the spatial coverage;

Visual analysis and processing machine to determine the coordinates of access points for networks with complex spatial configuration;

Determining the structural network diagrams required frequency resources, the number of cells;

Equipment selection - bridges, repeaters, repeaters, routers, routers, and other tools for network infrastructure.

4. Modelling of the network infrastructure

Modelling of the network infrastructure

Fig. 5 - for the construction of multi-level network optimization;

localization, verification under-interference, obstacles and predictable sporadic impacts;

the formation of design decisions on the system and application levels;

providing a carrier mode traffic with QoS.

5. Conclusion

The presented models and tools ensure the full, high-quality design and testing of broadband wireless networks and systems for existing and future applications.

At the initial stage of the deployment of complex wireless networks, modernization and operation of these funds increase reliability; the stability of the technical parameters, functional stability.

Reduces the complexity, cost, recovery time during operation and maintenance.

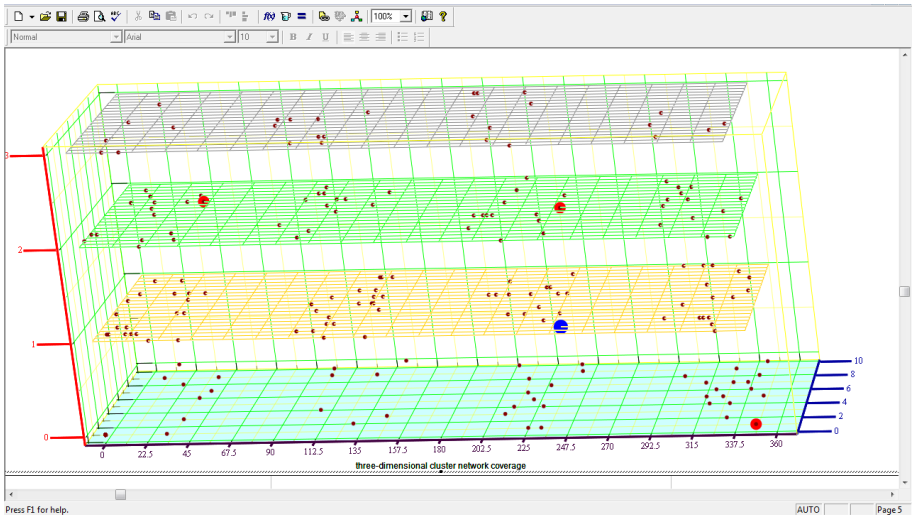


Fig. 5: Resulted space simulation coverage while minimizing infrastructure.

REFERENCES

1. Vishnevsky V.M. Broadband wireless data transmission networks. Technosphera, Moscow, 2005, 592 p.
2. Churkov V.M., Kosilov N.A., Zhdanov V.M. Optimization of Spatial Models Wireless Sensor Networks // Proceedings of Seventh Intern. Conf. "Information and telecommunication technologies in intelligent system" Schweiz, July 03-09, 2010, pp 9–16.

PERFORMANCE ASSURANCE IN SOFTWARE-DEFINED NETWORKS

V. Efimushkin, T. Ledovskikh, D. Korabelnikov, D. Iazykov
«Intellect Telecom» JSC, Moscow, Russia

Abstract

Migration to software-defined networks (SDN) is an effective solution of problems, emerging while routing dynamically changing traffic of simultaneously operating heterogeneous applications. Despite the fact that both SDN and Network Function Virtualization (NFV) have been recognized as key world-wide telecommunication trends, despite large-scale research and development efforts in SDN/NFV undertaken by major telecommunication vendors and network operators, approaches to performance assurance in SDN/NFV networks are still underdeveloped.

ОБЕСПЕЧЕНИЕ КАЧЕСТВА ФУНКЦИОНИРОВАНИЯ ПРОГРАММНО-КОНФИГУРИРУЕМЫХ СЕТЕЙ

В.А. Ефимушкин1[1], *Т.В. Ледовских*[2], *Д.М. Корабельников*[3], *Д.Н. Языков*[4]

[1,2,3,4] АО «Интеллект Телеком», Москва, Россия
[1]efimushkin@i-tc.ru, [2]ledovskikh@i-tc.ru, [3]korabelnikov@i-tc.ru, [4]yazikov@i-tc.ru

Аннотация

Эффективным решением задач, возникающих в ходе маршрутизации динамически изменяющегося трафика одновременно функционирующих разнохарактерных приложений, является переход на технологии программно-конфигурируемых сетей (Software Defined Network, SDN). Сегодня, в условиях признания технологий SDN и виртуализации сетевых функций (Network Function Virtualization, NFV), в качестве основного тренда развития телекоммуникаций в мире и проведения разработок в области SDN/NFV практически всеми крупнейшими операторами и производителями оборудования связи, вопросы обеспечения качества функционирования в сетях SDN/NFV операторского класса остаются недостаточно проработанными.

Ключевые слова: программно-конфигурируемая сеть, контроллер, качество функционирования, качество услуги, QoS, SDN.

1. Введение

Непроработанность архитектуры SDN, включая ее наиболее развитый вариант – OpenFlow, в части обеспечения качества услуг объясняется как новизной технологии [1, 2], так и структурной сложностью моделей, необходимых для исследования и разработки практических решений в области качества в сетях SDN [3]. В процессе развития стандарта OpenFlow в протокол добавлялись некоторые механизмы QoS, однако даже последняя версия предлагает ограниченные возможности и не содержит концептуальных положений по вопросам качества необходимой детализации [4].

Продолжающиеся активные работы в области качества [5, 6, 7, 8, 9, 10, 11], в том числе в международных Форумах и организациях по стандартизации, предположительно в течение ближайшего времени дадут достаточные для применения практические подходы и механизмы обеспечения качества в сетях SDN. Некоторые из них рассматриваются в статье.

2. Механизмы обеспечения QoS протокола OpenFlow

В протокол OpenFlow версии 1.0 были внедрены некоторые возможности QoS: пакеты могут направляться в очереди на выходных портах с помощью опционального действия постановки в очередь, которое в версии 1.2 было преобразовано в опциональное действие «Установить очередь». Контроллеры могут запрашивать статистику по очередям и параметры конфигурации очередей с помощью протокола OpenFlow, информацию о гарантированных минимальных скоростях, ассоциированных с очередями, но сам протокол не поддерживает возможности создания очередей или изменения поведения существующих очередей, а необходимость конфигурации очередей лежит за рамками OpenFlow.

В версиях не ниже 1.2, коммутатор может сообщать контроллерам также и максимальную скорость. Максимальной и минимальной скорости конечно не достаточно для построения умеренно сложного сценария QoS, т.к. поведение очередей описывается и другими характеристиками, которые невозможно запросить. Например, контроллер OpenFlow не может отличить очереди малой длины с нулевым минимальным пороговым значением при дисциплине произвольного раннего обнаружения (Random Early Detection, RED) от больших очередей, вызванных переполнением буферов при использовании дисциплины «пришедший первым обслуживается первым» (First In First Out, FIFO) [4].

Чтобы обойти ограничения протокола OpenFlow по обеспечению QoS, некоторые производители оборудования разрабатывают собственные решения. Так, компания HP разработала расширения протокола, с помощью которых контроллер может удалено создать в коммутаторе OpenFlow компании HP объект ограничения скорости, сбрасывающий пакеты, превышающие по их числу предустановленный порог, и добавлять действия к таблицам потоков, направляющих соответствующие пакеты через объекты ограничения скорости. Компания предложила это расширение, и версия

1.3 OpenFlow поддерживает подобную функциональность по ограничению скорости с использованием таблиц измерителей.

В дальнейшем, не расширяя возможностей по конфигурированию коммутатора в протоколе OpenFlow, консорциум ONF создал дополнительный протокол OF-Config. С точки зрения поддержки QoS, протокол может использоваться только для удаленной установки минимальной гарантированной и максимальной скоростей для очередей, т.е. OF-Config должен использоваться для конфигурации очередей, а OpenFlow – для конфигурации таблиц измерителей.

В соответствии со спецификацией версии 1.5.0 (раздел 7.3.5.8) протокол OpenFlow обеспечивает ограниченную поддержку QoS за счет возможности перезаписи битовой части DiffServCodePoint (DSCP) поля ToS (Type of Service) в заголовке IP (базовая поддержка QoS в OpenFlow) и сопоставления (согласования) битов IP ToS/DSCP, посредством использования простого механизма организации очередей и использования измерителей [4].

2.1. Механизм организации очередей. В коммутаторе OpenFlow заложен пока простой механизм обеспечения QoS - организация очередей. Одна или несколько очередей могут быть закреплены за портом и использоваться для отображения (сопоставления) записей потоков на них; записи, сопоставленные очередью, обрабатываются в соответствии с конфигурацией очереди.

Сам протокол OpenFlow не отвечает за конфигурацию очереди, это реализуется либо средствами командной строки, либо соответствующим внешним протоколом конфигурирования, например, протоколом OF-Config. Контроллер может запросить у коммутатора информацию о сконфигурированных очередях на порте, используя запрос OFPMP_QUEUE_DESC.

Сообщение запроса OFPMP_QUEUE_DESC предоставляет описания очередей для одного или более портов и одной или более очередей. Тело запроса содержит поле «port_no» (номер порта), определяющее порт OpenFlow, для которого запрашивается статистика, или OFPP_ANY – для всех портов. Поле «queue_id» (идентификатор очереди) определяет одну из приоритетных очередей или OFPQ_ALL – все очереди, сконфигурированные на определенном порте.

Коммутатором OpenFlow предусмотрена поддержка опционального действия «Установить очередь». Данное действие устанавливает идентификатор очереди в пакете, который будет указывать на заранее конфигурируемую очередь на порте для буферизации потока без учета полей ToS (Type of Service) и VLAN PCP (Priority Code Point). При поступлении пакета на порт определяется очередь, закрепленная за этим портом и используемая для диспетчеризации и передачи пакетов. Иногда коммутатор может поддерживать только очереди, связанные со специфическими битами ToS/PCP. В этом случае невозможно сопоставить произвольную запись потока с определенной очередью, поэтому действие «Установить очередь» не поддерживается, а пользователь может использовать очереди и сопостав-

лять записи потока для них с помощью установки соответствующих полей ToS, VLAN PCP.

2.2. Использование измерителей. Измеритель – элемент коммутатора, который может измерять и контролировать скорость передачи пакетов. В соответствии с заданным диапазоном значений, при превышении числа пакетов или байтов в единицу времени измеритель сбрасывает пакеты, если указано такое действие, при этом действие с заданным значением диапазона называется ограничителем скорости. Записи в таблице измерителей описывают измерители всех потоков.

Измерители потоков позволяют протоколу OpenFlow реализовывать различные простые операции по обеспечению QoS, как, например, ограничение скорости передачи, и могут быть объединены с механизмами организации очередей на портах для реализации комплексных подходов к обеспечению QoS, наподобие DiffServ.

Измерители закреплены за записями потоков в противоположность очередям, которые закреплены за портами. В наборе инструкций любой записи потока может быть указан измеритель. Соответственно, измеритель измеряет и контролирует совокупную скорость потоков, в записях которых он указан. В одной таблице потоков могут использоваться несколько измерителей, но они не должны фигурировать в записях одного потока.

Запись измерителя включает следующие компоненты:

- идентификатор (meter identifier): 32-битное целое число без знака однозначно определяющее измеритель;
- диапазоны значений (meter bands): неупорядоченный перечень (список) диапазонов значений измерителей, где каждый диапазон определяет скорость и соответствующий способ обработки пакета;
- счетчики (counters), которые обновляются при обработке пакетов измерителем.

Каждый измеритель может иметь один или несколько диапазонов значений; каждый диапазон определяет скорость, на которой он применяется, и способ обработки пакетов. Пакеты обрабатываются в соответствии с диапазоном, исходя из измеренной измерителем скорости. Измеритель применяет диапазон с наиболее высокой скоростью, меньшей текущей измеренной скорости. Если текущая скорость меньше, чем любая определенная диапазонами измерителя скорость, то действия, диапазонов не применяются.

Диапазон значений измерителя включает следующие основные компоненты:

- тип (band type): определяет, каким образом обрабатываются пакеты;
- скорость (rate): используется измерителем для выбора диапазона, определяет самую низкую скорость, на которой может применяться диапазон;
- счетчики (counters): обновляются при обработке пакетов диапазоном;

- специальные параметры (type specific arguments): некоторые типы диапазонов имеют опциональные параметры.

В текущей спецификации протокола OpenFlow версии 1.5.0 обязательные типы диапазонов не определены. Контроллер может запросить коммутатор, какой из опциональных (необязательных) типов диапазонов значений измерителя он поддерживает. Спецификацией определены следующие опциональные типы:

- drop: сбросить пакет (для определения диапазона ограничителя скорости);
- dscp remark: увеличивает приоритет сброса пакета в поле DSCP заголовка IP (для определения простой политики DiffServ).

3. Архитектурные решения по обеспечению QoS

Ограниченные возможности OpenFlow по обеспечению QoS ставят актуальные задачи в части проведения исследований и их апробации в различных проектах: предлагаются решения, реализующие как различные архитектурные расширения к уже существующим экспериментальным моделям контроллеров OpenFlow [5], так и новые архитектуры, например, [6, 7], и механизмы обеспечения QoS [8, 9, 10, 11].

3.1. Архитектура инфраструктуры управления Ofelia с поддержкой QoS. В [5] описывается архитектурное расширение к главной Европейской экспериментальной сети OpenFlow – Ofelia [12]. Его целью являлась реализация на базе Ofelia управляемой интегрированной поддержки QoS для дальнейшего проведения экспериментов по QoS и превращения Ofelia в главный двигатель инноваций в части QoS в OpenFlow. Для расширения возможностей сети Ofelia в целях осуществления настройки QoS, была предложена платформа по управлению QoS с полной интеграцией в существующую инфраструктуру управления, рис. 1.

В архитектуру сети добавлен модуль управления очередями (Queue Manager Plugin) с возможностями их унифицированного конфигурирования, с пользовательским интерфейсом, единообразным для разнородного оборудования, позволяющим настраивать параметры очередей коммутаторов и управлять ими.

Модуль управления очередями транслирует запрос пользователя в последовательности команд конфигурации оборудования соответствующего производителя и затем исполняет их через различные интерфейсы конфигурации/управления. В качестве таких интерфейсов могут использоваться стандартные интерфейсы аппаратных коммутаторов (например, SNMP), собственная разработка для программных коммутаторов на базе интерфейса Netconf с применением расширений по QoS, а также протокол OF-Config, реализованный в аппаратных и программных (виртуальных) коммутаторах (рис. 2). Кроме того, модуль управления очередями может взаимодействовать и с другими устройствами в сети Ofelia, например, расширения на

базе Netconf могут быть легко интегрированы в беспроводные маршрутизаторы OpenWrt. Предлагается также внедрить в модули инфраструктуры управления Ofelia «FlowVisor» и «Менеджер подключений» дополнительные возможности по защите очередей и выбору пользователями заранее сконфигурированных очередей.

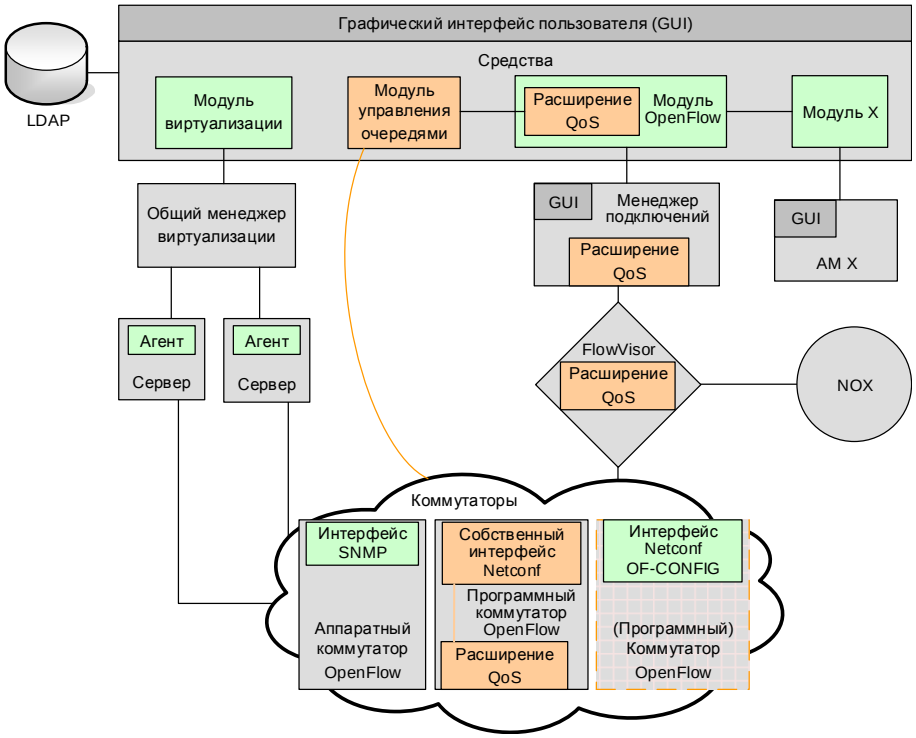


Рис. 1: Общая архитектура инфраструктуры управления Ofelia с поддержкой QoS

3.2. Архитектура OpenQoS. В [6] предлагается расширение стандартной архитектуры контроллера OpenFlow – OpenQoS для реализации нового механизма обеспечения сквозного QoS при предоставлении услуг мультимедиа посредством дополнительных интерфейсов и функций QoS, рис. 2. В предложенной архитектуре через интерфейс *контроллер – сетевые устройства* по защищенному каналу контроллером осуществляется отправка таблиц потоков, запросы и получение информации о состоянии сети от коммутаторов, мониторинг сети. При увеличении числа узлов OpenFlow, могут потребоваться несколько контроллеров, и интерфейс *контроллер – контроллер* используется в этом случае для совместного управления сетью. Контроллер предоставляет открытый, безопасный интерфейс

контроллер – услуга поставщикам услуг для установки определений потоков и правил их маршрутизации, для оповещения контроллера о начале передачи потоков данных новым приложением.

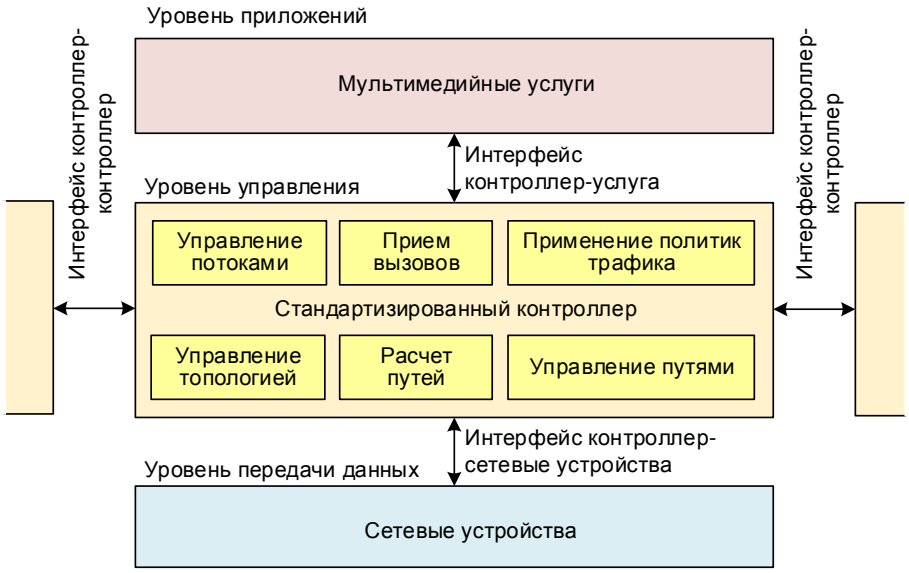


Рис. 2: Архитектура контроллера OpenQoS

Функции контроллера в архитектуре OpenQoS включают *управление топологией* в целях мониторинга и поддержания связности сети на основе данных от коммутаторов; *управление маршрутами*, их перерасчет при несоответствии уровня готовности узлов сети и качества передачи пакетов заданному; *управление потоками*, включая их агрегацию и получение определений потоков от поставщиков услуг; *расчет маршрутов* для потоков различного типа на базе информации о топологии сети и управлении маршрутами (параллельно могут применяться несколько алгоритмов маршрутизации) с учетом резервирования и удовлетворения требований по качеству функционирования; *управление доступом вызовов* с блокировкой вызова с необеспеченным QoS (например, при отсутствии возможных маршрутов) и информированием контроллера о необходимых действиях; политики трафика с определением согласования потоков с запрашиваемыми параметрами QoS и применением правил политики, когда они не согласованы, например, для приоритетного трафика или выборочного сброса пакетов.

В OpenQoS предлагается новая схема приоритезации для обеспечения требуемого сквозного качества и динамическая QoS-маршрутизация для QoS-потоков. Входящий трафик группируется в потоки мультимедиа (QoS-

потоки) и потоки данных (остальные); QoS-потоки динамически помещаются на маршруты с гарантированным QoS, а потоки данных передаются по обычным маршрутам (кратчайшего пути).

Данный подход отличается от существующих схем QoS (например, IntServ, DiffServ), т.к. в нем не используются ни резервирование ресурсов, ни организация очередей (т.е. формирование скорости). Отказ от использования этих методов позволяет, прежде всего, минимизировать такие отрицательные эффекты от настройки не QoS-потоков как потеря пакетов и задержка.

Традиционная сетевая архитектура не позволяет осуществлять маршрутизацию на базе потоков. При поступлении пакета в маршрутизатор, сравниваются адреса источника и получателя с записями в таблице маршрутизации, и он передается согласно предопределенным правилам (например, протокол маршрутизации), сконфигурированным оператором сети. С другой стороны, OpenFlow обеспечивает гибкость в определении различных типов потоков, к которым могут устанавливаться набор действий и правил. Например, поток одного типа может передаваться, используя маршрутизацию по кратчайшему пути, другие потоки могут передаваться по сконфигурированным вручную маршрутам. Таким образом, пакеты каждого потока могут по-разному обрабатываться на сетевом уровне. В OpenFlow потоки можно определить любым способом; они могут содержать одинаковые или различные типы пакетов. Например, пакеты с портом 80 протокола TCP (зарезервирован для HTTP) могут быть определением для потока, или пакеты, имеющие заголовок RTP, могут указывать на поток, который передает голос, видео или одновременно и то, и другое. Таким образом, можно установить поток как комбинацию полей разных уровней заголовка пакета: L1 (Ingress Port, Metadata), L2 (Ethernet Source, Ethernet Destination, Ethernet Type, VLAN ID, VLAN Priority), L2.5 (MPLS Label, MPLS Traffic Class), L3 (IP Source, IP Destination, IP Protocol, IP TOS field), L4 (Transport Source Port, Transport Destination Port), однако необходимо учитывать ограниченные возможности сетевого оборудования (маршрутизаторы или коммутаторы) по их обработке.

Во избежание сложных поисков в таблицах потоков, определения потоков должны быть четко установлены и по возможности агрегированы. В OpenFlow сетевые устройства хранят потоки и соответствующие им правила в таблицах потоков с конвейерной обработкой пакетов с целью сокращения времени обработки.

В архитектуре OpenQoS используется принцип передачи на базе потоков OpenFlow с разделением трафика на данные и мультимедиа. Потоки мультимедиа можно определять, используя поля в заголовках пакетов: поле заголовка класса трафика в MPLS, поле ToS в IPv4, поле класса трафика в IPv6, адрес IP источника, номер порта источника и/или пункта назначения. Поскольку сложность анализа пакетов ниже по сравнению с обработкой верхних уровней (L4), желательно определять потоки согласно заголовкам уровней L2, L3. В [6] предлагается определять потоки мульти-

медиа, используя поля в MPLS (L2.5) и обеспечивая возможности быстрой коммутации. В некоторых случаях для лучшего определения типов пакетов могут потребоваться поля заголовков уровня L4, что OpenFlow позволяет сделать. Кроме того, определение потоков может основываться не только на протоколе IP: любая схема адресации с информацией сервисного уровня может использоваться для определения типов потоков мультимедиа.

Для расчета QoS-маршрутов собирается информация о состоянии сети, такая как задержки, ширина полосы пропускания, потери пакетов для каждого канала. Сегодня большинство реализаций коммутаторов OpenFlow не поддерживают сбор статистики по задержке, джиттеру. Эффективность любого алгоритма маршрутизации напрямую зависит от точности информации о состоянии сети. В больших сетях, сбор такого рода информации зависит от размера сети и ее архитектуры. OpenFlow облегчает эту задачу посредством использования централизованного контроллера – вместо того, чтобы в традиционной сети маршрутизатору распространять информацию о состоянии сети между всеми маршрутизаторами, коммутаторы OpenFlow в сети SDN посылают информацию о состоянии контроллеру, который рассчитывает наилучшие маршруты.

4. Заключение

В заключение отметим, что предлагаемые архитектуры и возможные новые механизмы обеспечения QoS предполагают, что для передачи информации между составляющими сети связи могут использоваться «южный» и/или «северный» интерфейсы OpenFlow и соответствующие протоколы, а функции обеспечения QoS (например, управление политиками, управление очередями, управление и расчет маршрутов, мониторинг QoS и SLA, и др.) могут выполняться контроллером SDN или выноситься в отдельные приложения (на уровень приложений) [1, 2]. Исследуются и предлагаются расширения и улучшения существующих механизмов конфигурации очередей в коммутаторах OpenFlow, а также дополнительные возможности для реализации в протоколе OpenFlow [5, 6, 7, 8, 9, 10, 11].

На текущем этапе развития стандартов OpenFlow в части механизмов обеспечения качества функционирования сетей SDN и предоставления услуг в этих сетях рекомендуется использовать существующие решения производителей оборудования на базе гибридных коммутаторов, реализующих как возможности протокола OpenFlow, так и существующие механизмы QoS коммутаторов и маршрутизаторов традиционных сетей IP (политики QoS, маркировка трафика, организация очередей, формирование трафика, случайный и принудительный сброс трафика), а также возможно дополнительные решения, предлагаемые производителями оборудования (дополнительная реализация механизмов QoS, например, CBWFQ, WRED, приоритетная организация очередей с распространением, иерархический QoS, и др.).

ЛИТЕРАТУРА

1. Ефимушкин В.А., Ледовских Т.В., Корабельников Д.М., Языков Д.Н. Международная стандартизация программно-конфигурируемых сетей // Электросвязь. – 2014. – № 8. – С.3-9.
2. Ефимушкин В.А., Ледовских Т.В., Корабельников Д.М., Языков Д.Н. Сравнительный анализ архитектур и протоколов программно-конфигурируемых сетей // Электросвязь. – 2014. – № 8. – С.9-14.
3. Ефимушкин В.А., Языков Д.Н. Анализ характеристик функционирования коммутатора программно-конфигурируемой сети // XII Всероссийское совещание по проблемам управления ВСПУ-2014. Москва, 16-19 июня 2014 г.: Труды. М.: ИПУ РАН. – 2014. – С.8536-8543.
4. ONF OpenFlow Switch Specification, version 1.5.1 (Protocol version 0x06), 03/2015.
5. Sonkoly B. et. al. On QoS Support to Ofelia and OpenFlow // In: Proc. of the European Workshop on Software Defined Networking. – 2012. – Pp.109-113.
6. Egilmez H.E., Dane S.T., Bagci K.T. et. al. OpenQoS: An OpenFlow Controller Design for Multimedia Delivery with End-to-End Quality of Service over Software-Defined Networks // In: Proc. of the APSIPA Annual Conf., Los Angeles. – 2012.
7. Jeong K., Kim J., Kim Y.-T. QoS-aware Network Operating System for Software Defined Networking with Generalized OpenFlows // In: Proc. of the IEEE/IFIP 4th Workshop on Management of the Future Internet (ManFI). – 2012. – Pp.1167-1174.
8. Ko N.-S., Heo H., Park J.-D. OpenQFlow: Scalable OpenFlow with Flow-Based QoS // IEICE Transactions on Communications. – 2013. – Vol.E96-B, No.2. – Pp.479-488.
9. Cao C., Wang J., Tang X. A SDN-Controlled ECMP QoS Solution for Data Networks // Advanced Materials Research. – 2013. – V.765-767. – Pp.1730-1733.
10. Ishimori A., Farias F., Cerqueira E. et. al. Control of Multiple Packet Schedulers for Improving QoS on OpenFlow/SDN Networking // In: Proc. of the Second European Workshop on Software Defined Networks (EWSDN). – 2013. – Pp.81-86.
11. Wallner R., Cannistra R. An SDN Approach: Quality of Service using Big Switch's Floodlight Open-source Controller // In: Proc. of the Asia-Pacific Advanced Network. – 2013. – V.35. – Pp.14-19.
12. OpenFlow in Europe – Ofelia // `www.fp7-ofelia.eu/`

ARCHITECTURE OF USER APPLICATIONS FOR A NETWORK WITH MOBILE NODES

Farkhadov M.P.[1], *Blinova O.V.*[2], *Abramenkov A.N.*[3], *Vorontsov Y.A.*[4]
[1,2,3]V.A. Trapeznikov Institute of Control Sciences of RAS, Moscow, Russia,
[4]Moscow Technical University of Communications and Informatics, Russia

Abstract

The report analyzes the activities of the emergency services in the aftermath of an extraordinary situation. The basic requirements and proposed a solution for the implementation of user applications for mobile network nodes of communication that maximally meet the conditions solved problems, existing restrictions. To investigate the network and assess its effectiveness requires a deep understanding of the processes occurring in the network, as well as a range of tasks faced by the user. Originality and diversity of tasks the user requires a flexible and effective solution that combines performance with many features. Used modern approaches to designing of the software architecture.

Keywords: Network with mobile nodes, service-oriented architecture, emergency services, distributed networks.

АРХИТЕКТУРА ПОЛЬЗОВАТЕЛЬСКОГО ПРИЛОЖЕНИЯ ДЛЯ СЕТИ С ПОДВИЖНЫМИ УЗЛАМИ

М.П. Фархадов[1], *О.В. Блинова*[2], *А.Н. Абраменков*[3], *Ю.А. Воронцов*[4]
[1,2,3] ФГБУН Институт проблем управления им. В.А. Трапезникова РАН,
[4]Московский технический университет связи и информатики
[1]mais@ipu.ru, [2]aabramenkov@asmon.ru, [3]blinova_olga_v@mail.ru

Аннотация

В докладе проведен анализ деятельности аварийно-спасательных служб по ликвидации последствий чрезвычайной ситуации. Выявлены основные требования и предложено решение для реализации пользовательского приложения для сети с мобильными узлами связи, максимально отвечающего условиям, решаемым задачам, существующим ограничениям. Для исследования сети и оценки ее эффективности необходимо глубокое понимание всех процессов, происходящих в сети, а так же круга задач, возникающих перед пользователем. Неординарность и многоплановость задач пользователя требует гибкого и эффективного решения, сочетающего быстродействие

с широкими возможностями. Использованы современные подходы к проектированию архитектуры программного обеспечения.

Ключевые слова: Сеть с подвижными узлами, сервис-ориентированная архитектура, аварийно-спасательные службы, распределенные сети.

1. Введение

В докладе [1] рассмотрена модель информационной сети (ИС) с подвижными узлами для аварийно спасательных служб, предложена методика расчета коэффициента доступности сети. Разрабатываемая сеть предназначена в первую очередь для отрядов аварийно-спасательных служб, работающих в горных районах, с перспективой дальнейшего расширения на другие отряды. Выбор обусловлен и существующей потребностью и удобством тестирования системы. При общем небольшом объеме информации и численности отрядов, условия работы задают множество ограничений. Необходимость разработки эффективной и производительной сети побуждает обращаться к самым современным подходам к проектированию сетей, а внедрение той или иной технологии проще по сравнению с крупными предприятиями. Построение модели функционирования спасательных отрядов необходимо для оценки различных возможных вариантов создания информационной сети, а возможность тестирования и внедрения для ограниченного числа пользователей позволит использовать эти отряды как испытательный полигон для дальнейшего расширения сети на все аварийно-спасательные службы.

Основные особенности проведения спасательных работ в горных районах:

- полное или частичное отсутствие мобильной связи на месте ЧП, а так же источников электропитания;
- большое количество видов возможных ЧП (от травм в туристических группах до техногенных катастроф);
- объединение различных технологий, разнообразие пользовательских устройств;
- мобильное разворачивание сети связи в месте ЧП;
- большая часть пользовательских устройств подвижна.

Исходя из вышеперечисленного, основные требования к сети:

- Наличие постоянно функционирующей части для сохранения опыта проведения операций, справочных сведений, а так же для удобства мониторинга ситуации в рассматриваемом районе.
- Наличие динамически разворачиваемой части для использования на местности, сбора и передачи оперативной информации и управленческих решений.
- Динамическая часть сети должна быть максимально приспособлена для сложных условий работы (в т.ч. погодных), отсутствия электрической сети и покрытия мобильной сетью.

- Возможность доступа к ИС с любого устройства, в т.ч. удаленно, для возможности привлечения сторонних специалистов. Возможность использования и мобильных сотовых и спутниковых сетей.
- Возможность подключения к приложению различных внешних сервисов, например погодных, картографических и т.д.
- Обеспечение безопасности передаваемой информации и гибкой системой управления доступом пользователей к ИС.

Для выбора оптимальной архитектуры были использованы современные подходы к организации информационных сетей, основывающихся на мобильных беспроводных технологиях. Особое внимание уделялось сервис-ориентированной архитектуре, облачным технологиям, сетецентрическому принципу организации ИС [2-6].

Использованы следующие ключевые черты, присущие облачным технологиям и сервис-ориентированной архитектуре:

- Широкий сетевой доступ (Broad network access). Предоставляемые вычислительные ресурсы доступны по сети через стандартные механизмы для различных платформ, тонких и толстых клиентов (мобильных телефонов, планшетов, ноутбуков, рабочих станций и т. п.)
- Мгновенная эластичность (Rapid elasticity). Ресурсы могут быть эластично выделены и освобождены, в некоторых случаях автоматически, для быстрого масштабирования соразмерно со спросом (реализован не в полной мере).[2]

Возможности проведения спасательных операций, возникающие в результате внедрения ИС, общие с сетецентрическими системами для ведения боевых действий:

- Гибкое перераспределение резервных сил.
- Увеличение скорости принятия решений (speed of command) — время, необходимое для прохождения полного цикла Бойда: Наблюдение — Ориентация — Решение — Действие (Observe — Orient — Decide — Act, OODA)
- Самосинхронизация (self-synchronisation) — способность корректировать поведение в соответствии с действиями других участников системы и общей целью.
- Возможность доступа к нелокальным ресурсам (reachback) — возможность использовать ресурсы, географически расположенные вне зоны боевых действий. Например, консультации удаленных аналитиков.
- Повышения уровня распределенной ситуационной осведомленности (level of shared situational awareness) — степени общности выводов, к которым приходят участники системы по мере поступления информации, и действий, которые они планируют в последствие
- Интероперабельность в рамках информационного пространства (interoperability in the information domain) — способность сил к подготовке,

осуществлению и эффективному совместному взаимодействию с целью выполнения поставленных задач и целей [3]

Построение схемы архитектуры пользовательского приложения

Характер и масштаб возникающих чрезвычайных ситуаций (ЧС) может быть самым разным, поэтому в первую очередь выделен ряд общих этапов по ликвидации последствий чрезвычайной ситуации, не зависящих от обстоятельств:

1. Мониторинг ситуации
2. Поступление информации о ЧП
3. Первичная реакция на получение информации о ЧП
   3.1. Формирование штаба ликвидации последствий ЧП
   3.2. Выдвижение сил и средств на место происшествия
4. Основной этап ликвидации ЧП
   4.1. Принятие решения о привлечении дополнительных сил и средств
   4.2. Действия на местности по ликвидации последствий ЧП, эвакуация пострадавших.
   4.3. Разведка ситуации на месте и передача подробной информации в штаб
   4.4. Обработка получаемой информации, принятие решений о текущих действиях на местности
   4.5. Управление поступающими силами и ресурсами
5. Завершение операции
   5.1. Завершение действий по ликвидации ЧП, демобилизация.
   5.2. Приведение сил и средств в готовность.

На рис. 1 представлена общая схема ликвидации последствий ЧП с использованием ИС с динамически разворачиваемой на местности подсистемой. На 1,2 и 5 этапе используется стационарная часть сети, при получении информации о ЧП разворачивается динамическая часть для передачи оперативной информации, на этапе 5 динамическая часть сети сворачивается. Для повышения скорости развертывания сети оборудование динамической части предлагается собирать в комплекты, включающие аккумуляторы.

На рис. 2 изображена схема пользовательского приложения для работы в ИС. Один сервер обеспечивает длительное хранение архивной информации, расположен в дата-центре или там, где возможно обеспечить необходимые условия и широкополосный доступ к сети Интернет. Второй сервер расположен в ближайшем к месту ЧП поисково-спасательном пункте (ПСП) и отвечает за сбор и обработку оперативной информации. Главная задача ИС — обеспечения связи всех участников операции с сервером 2, и передача управляющих решений от ПСП к непосредственным исполнителям. Универсальный web интерфейс позволяет подключать к системе волонтеров и сторонних специалистов с любого пользовательского устройства, а так же облегчает подключение к системе внешних сервисов.

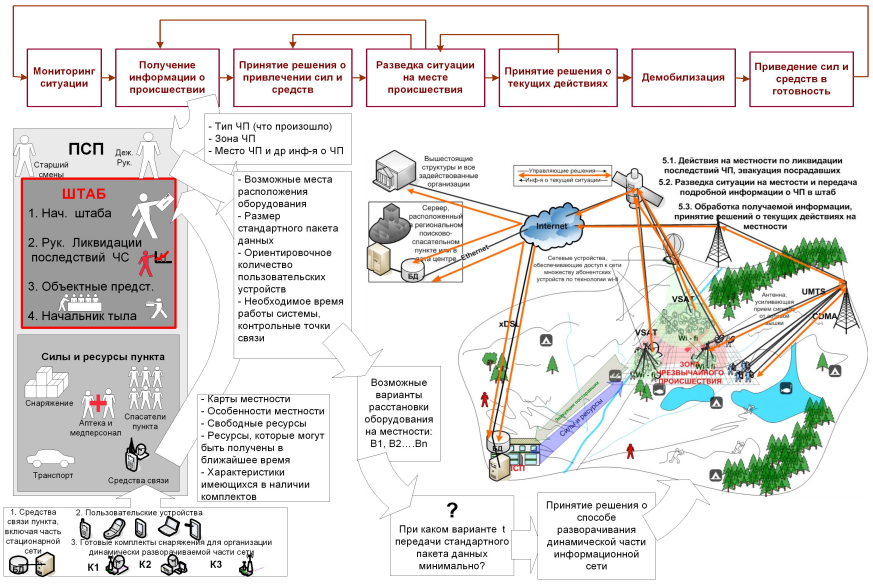


Рис. 1: Общая схема ликвидации последствий ЧП с использованием ИС

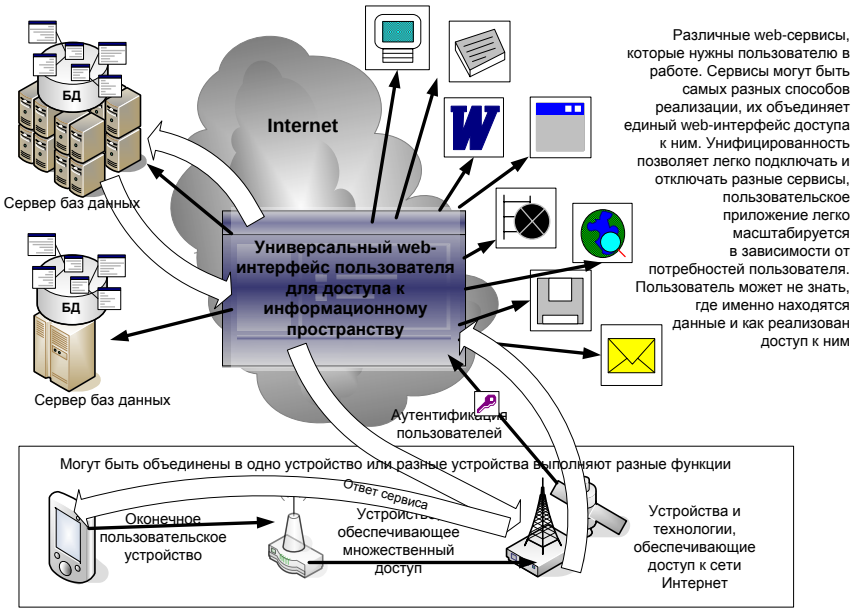


Рис. 2: Архитектура пользовательского приложения

2. Заключение

Проведен анализ работы поисково-спасательных отрядов при ликвидации последствий ЧП. Выявлены основные требования к разрабатываемой системе. Предложена концептуальная схема пользовательского приложения с учетом современных тенденций проектирования сетей. Построенные схемы можно использовать как основу для построения моделей функционирования сети в практических ситуациях и расчета показателей эффективности внедрения сети, а так же выбора оборудования и программного обеспечения, максимально отвечающего поставленным целям.

ЛИТЕРАТУРА

1. Фархадов М.П., Блинова О.В., Абраменков А.Н., Воронцов Ю.А. Модель сети с подвижными узлами связи для аварийно-спасательных служб. // 5-й научно-технический семинар кСовременные проблемы прикладной математики, информатики, автоматизации и управлениянь (Севастополь, 2015). (в печати)
2. Батура Т.В., Мурзин Ф.А., Семич Д.Ф. Облачные технологии: основные понятия, задачи и тенденции развития. // Программные продукты и системы и алгоритмы є 1, 2014 г. С. 64–72.
3. Фархадов М.П., Душкин Д.Н. Сетецентрические технологии: эволюция, текущее положение и области дальнейших исследований // Автоматизация и современные технологии. 2012. є 1. С. 21–29.
4. Ефремов А. Ю., Максимов Д. Ю. Сетецентрическая система управления — что вкладывается в это понятие? Институт проблем управления им. В.А. Трапезникова РАН, г. Москва, 2012. С.158–161.
5. В.С Балицкий, М.В. Кривенков, А.В. Каверный. Мобильные комплексы связи: проблемы, варианты, решения // Мобильные системы, 2007. С. 36–37.
6. Сорокин А.А. Разработка программного комплекса для исследования телекоммуникационных систем с динамической топологией сети // Вестник Астраханского государственного технического университета. Серия: Управление, вычислительная техника и информатика, є 2 / 2011. С. 137–142.

CURRENT APPROACHES TO THE INTERNET PHYSICAL CHANNELS TRAFFIC CLASSIFICATION

P. Filimonov, M. Ivanov
Academy of FGS of Russia, Orel, Russia

Abstract

The article discusses the main current approaches to the classification of IP-traffic on the Internet. It describes the basic classification methods, their advantages and disadvantages.

СОВРЕМЕННЫЕ ПОДХОДЫ К КЛАССИФИКАЦИИ ТРАФИКА ФИЗИЧЕСКИХ КАНАЛОВ СЕТИ ИНТЕРНЕТ

П. Филимонов[1], *М. Иванов*[2]
Академия ФСО России, Орел, Россия
[1]pafilimonov@gmail.com, [2]maximivanov@mail.ru

Аннотация

В статье рассматриваются основные современные подходы к классификации IP-трафика в сети Интернет. Описываются основные методы классификации, их преимущества и недостатки. Развернуто приводятся подходы к классификации на основе полезной нагрузки в зависимости от методов обработки и требований к памяти. Производится сравнение способов машинного обучения при классификации на основе статистических методов.

Ключевые слова: IP-трафик, классификация, полезная нагрузка, статистические методы, машинное обучение.

1. Введение

В настоящее время системы мониторинга каналов связи все чаще сталкиваются с недостатком ресурсов для постоянного наблюдения всех находящихся в доступности каналов. При этом требуется производить оценку заданных параметров каждого канала связи из доступных. Эта задача осложняется тем, что зачастую для некоторых каналов возможны лишь периодические наблюдения. Однако, задача может быть решена, если предположить, что доступные для наблюдения каналы связи можно объединить в группы со схожими характеристиками так, что на основании оценки заданных параметров одного или нескольких каналов из группы можно спрогнозировать их значения для остальных каналов в группе. Для объединения физических каналов сети Интернет в такие группы, обладающие

схожими характеристиками, предполагается использовать классификацию по типам IP-трафика.

В настоящее время существуют следующие методы классификации IP-трафика физических каналов сети Интернет (рисунок 1):

- анализ номеров портов в пакетах TCP и UDP;
- восстановление сигнатур протокола из его полезной нагрузки (классификация, основанная на полезной нагрузке);
- анализ статистических характеристик обмена сообщениями между узлами сети и статистических характеристик трафика (классификация, основанная на использовании статистических методов) [1].

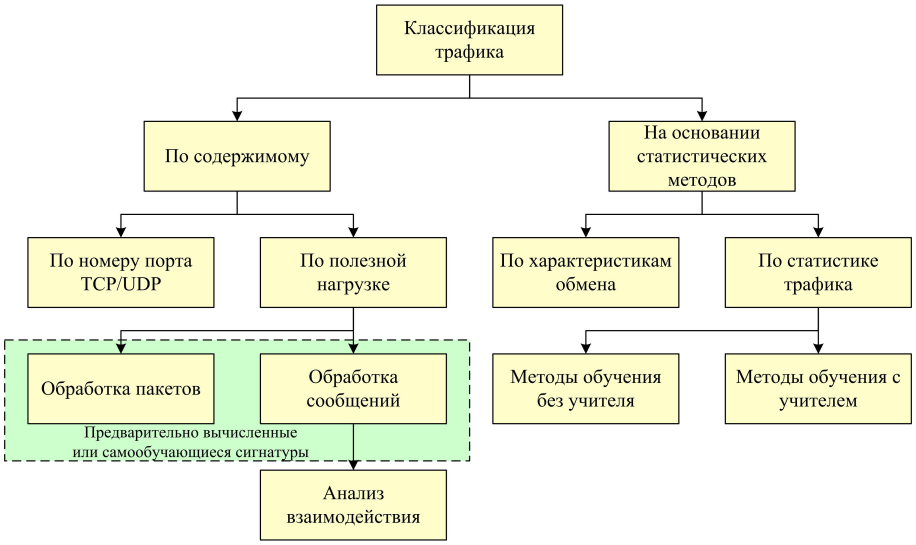


Рис. 1: Методы классификации IP-трафика физических каналов сети Интернет

Каждый из них обладает своими достоинствами и недостатками. Рассмотрим методы, приведенные в классификации, несколько подробнее.

2. Классификация IP-трафика на основе номеров портов

Протоколы TCP и UDP обеспечивают мультиплексирование множества потоков между точками IP-сети посредством использования номеров портов. В большинстве случаев приложения используют для передачи пакетов определенные порты. Задача классификатора при этом заключается

в определении номера порта, например, из заголовка TCP или UDP пакета. Если порт присутствует в списке зарегистрированных портов IANA [2], то можно определить протокол, к которому относится данный пакет. Достоинствами метода являются простота реализации и высокая скорость работы.

Однако такой подход имеет и ряд ограничений. Во-первых, некоторые приложения могут не иметь своих портов, зарегистрированных в IANA. Приложения могут использовать порты, отличные от зарегистрированных в IANA, чтобы обойти ограничения контроля доступа в ОС. В некоторых случаях порты назначаются динамически по мере надобности. Также один и тот же порт для передачи пакетов могут использовать разные программы [1].

3. Классификация IP-трафика на основе полезной нагрузки

В отличие от вышеописанного подхода, где присутствует полная зависимость от номеров портов, для классификации трафика на основе полезной нагрузки требуются дополнительные сведения об используемом протоколе, получаемые с помощью восстановления состояния сеанса и из информации прикладного уровня, содержащейся в каждом пакете.

В [3] приводятся подходы к классификации трафика на основе полезной нагрузки в зависимости от методов обработки и методов проверки (рисунок 2).

При классификации трафика выделяют четыре различных уровня проверки. Первый уровень проверки основан на сигнатуре, его целью является поиск ряда сигнатур в полезной нагрузке прикладного уровня. Метод на основе сигнатур основан на проверке соответствия полезной нагрузки (или ее части) сигнатуре, определенной для данного протокола. Сигнатуры, как правило, задаются регулярными выражениями, однако они могут включать некоторые элементы управления, например, проверку по длине полезной нагрузки.

Второй уровень проверки – синтаксический. Он может рассматриваться как более точная версия сигнатурной проверки, поскольку направлен на проверку корректности переданных данных с синтаксической точки зрения (например, предполагается, что полезная нагрузка протокола HTTP должна содержать HTTP-заголовки). В этом случае необходимо декодировать все поля, содержащиеся в сообщении, и гарантировать, что сообщение является корректно сформированным.

Третий уровень проверки – соответствие протоколу. К примеру, на данном уровне осуществляется контроль, что на запрос HTTP GET от клиента следует действительно ответ от сервера. Такая форма проверки является более точной, так как реальное поведение протокола может быть проверено в соответствии со спецификацией.

Четвертый уровень проверки относится к семантике данных, т.е. возможности проверить, действительно ли объект изображения, передавае-

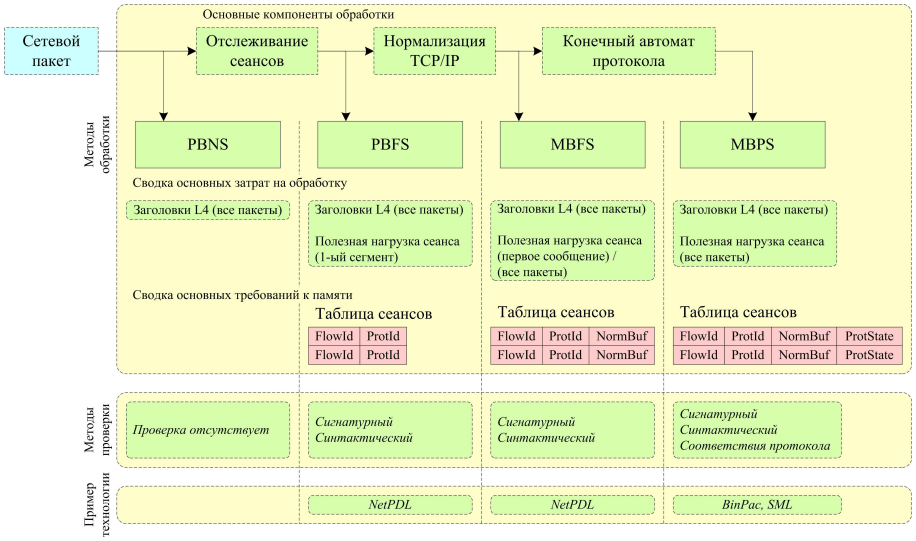


Рис. 2: Классификация трафика на основе полезной нагрузки

мый по протоколу HTTP, является изображением, или его содержимое является чем-либо другим. Эта проверка является чрезвычайно полезной для обнаружения механизма „умных туннелей“, в которых приложение использует другой протокол для транспортировки данных [3].

При классификации трафика по методу обработки выделяют следующие методы: PBNS (Packet Based, No State – обработка пакетов без хранения информации о состоянии сеанса), PBFS (Packet Based, Per Flow State – обработка пакетов с сохранением состояния сеанса), MBFS (Message Based, Per Flow State – обработка сообщений с сохранением состояния сеанса), MBPS (Message Based, Per Protocol State – обработка сообщений с сохранением состояния протокола).

Метод PBNS является простейшим методом из представленных. Он работает, проверяя значения некоторых полей, присутствующих в каждом пакете, таких, как поля TCP/UDP портов. Этот метод очень прост с точки зрения вычислений (должны быть обработаны только заголовки пакета до транспортного уровня эталонной модели взаимодействия открытых систем), для него не требуется хранить состояния, но его точность на текущем трафике является недостаточной.

Второй метод – PBFS – требует хранения таблицы сеансов, в которой каждая запись включает идентификатор сеанса (кортеж из 5 полей – IP-

адреса источника/назначения, транспортный протокол и порты источника/назначения) и соответствующий идентификатор протокола прикладного уровня. Каждая запись в таблице занимает по нескольку десятков байт. Этот метод поддерживает реализацию механизма проверки на основе данных прикладного уровня, хотя это и ограничивается анализом пакетов.

Третий метод – MBFS – работает с сообщениями, а не с пакетами. Для этого метода требуется модуль нормализации пакетов TCP/IP. В принципе, технологии на основе MBFS могут выполнить те же проверки, что и PBFS, но работают с сообщениями, из чего следует, что его средства контроля могут быть распространены на все сообщение полностью, а не только на первый сегмент с данными. В этом случае увеличиваются требования к размеру памяти, потому что для каждого сеанса должна быть сохранена дополнительная информации о состоянии (например, порядковый номер TCP-пакета), а также должны быть выделены буферы памяти, требуемые для работы нормализатора TCP/IP. Все эти параметры сильно зависят от природы трафика, т.е. от количества фрагментированных пакетов и некорректных TCP-сессий.

Четвертый метод – MBPS – делает еще один шаг вперед и интерпретирует то, что передает и получает каждое приложение. Обработчик MBPS понимает не только семантику сообщения, но и различные этапы обмена сообщениями, потому что имеет полное представление о конечном автомате протокола. Требования к размерам памяти становятся еще большими, так как надо учитывать не только состояние транспортного сеанса, но и состояние всех сессий уровня приложений. Кроме того, предъявляются самые высокие требования к производительности, потому что для проверки состояния протокола требуется обработка всех данных прикладного уровня.

Существующие технологии могут не укладываться полностью в эту классификацию, поскольку одна и та же технология может относиться к нескольким категориям. В таком случае, технологии обычно разделяют на технологии на основе анализа пакетов (NetPDL [4], NBAR [5]) и на основе анализа потоков (SML [6], BinPac [7]). Однако, эти технологии в различных случаях могут вести себя в зависимости от реализации, например, NetPDL и NBAR могут относиться либо к PBFS, либо к MBFS в зависимости от наличия нормализатора TCP/IP в цепочке обработки данных [3].

4. Классификация IP-трафика на основе статистических методов

В статистических методах существуют два разных подхода: поведенческие алгоритмы и статистические алгоритмы сетевого и транспортного уровней.

Концепция поведенческого алгоритма была разработана в [8]. Основная цель метода состоит в том, чтобы определить, какие приложения создают определенные потоки трафика. Анализируя, как в рамках сети взаимодей-

ствуют хосты, можно определить, какие виды приложений запущены на хосте.

Подход статистических методов опирается на статистические характеристики трафика для идентификации приложения. В основе этих методов лежит предположение, что сетевой трафик обладает статистическими характеристиками, которые являются уникальными для определенных классов приложений и позволяют разделить различные исходные приложения [1].

Статистические алгоритмы в зависимости от подхода к классификации можно разделить на две группы: методы классификации или обучение с учителем и методы кластеризации или обучение без учителя.

4.1. Применение методов машинного обучения для классификации IP-трафика. Класс обычно идентифицирует IP-трафик, сформированный приложением или группой приложений. В качестве признаков обычно выступают числовые атрибуты, вычисленные на основании множества пакетов, принадлежащих потоку. Не все признаки одинаково применимы для классификации, поэтому на практике классификаторы выбирают наименьшее множество признаков, которое приведет к эффективному разделению.

Оптимальный подход к алгоритму обучения с учителем должен предусматривать предварительно классифицированные образцы двух типов IP-трафика: трафика, соответствующего классу, который хотим позднее идентифицировать в сети, и трафика от других приложений, которые, возможно, встретятся в будущем (часто называемый трафиком помехи).

Сначала собирается смесь „трасс трафика“, включающая в себя как трафик интересующего нас приложения, так и трафик других приложений. Далее следует обработка статистики потоков – вычисление статистических свойств этих потоков. Следующий необязательный этап – осуществление выборки данных, разработанный, чтобы уменьшить пространство поиска для обучающего алгоритма, когда он сталкивается с очень большими обучающими выборками (трассами трафика). На этом этапе из подмножества экземпляров различных классов приложений извлекается статистика и передается классификатору для использования в процессе обучения.

Желательно произвести фильтрацию/выбор признаков, чтобы ограничить число признаков, действительно используемых при обучении классификатора, и, таким образом, создать модель классификации.

Для оценки точности результатов, полученных на этапе обучения, может быть применена перекрестная проверка. Однако если исходный набор данных будет состоять из IP-пакетов, собранных в то же самое время и в той же самой точке наблюдения в сети, то результаты перекрестной проверки, скорее всего, переоценят точность классификатора. В идеальном случае исходный набор данных должен содержать смесь трафика, собранного в разное время и разных точках сети, или должны использоваться

полностью независимо собранные обучающие и тестирующие наборы данных [1].

4.2. Обучение с учителем в сравнении с обучением без учителя. Классификация IP-трафика обычно выражается в идентификации трафика, принадлежащего известным приложениям, внутри потоков IP-пакетов. Ключевой задачей является определение связей между классами IP-трафика и приложениями, его генерирующих.

Для алгоритма машинного обучения с учителем требуется фаза обучения, которая требует априорной классификации потоков и соответствующих классам обучающих выборок. По этой причине машинное обучение с учителем может быть привлекательным для идентификации одного или группы интересующих приложений. Однако классификатор алгоритма машинного обучения с учителем лучше всего работает тогда, когда он обучен на примерах для всех классов трафика, которые ожидаются на практике, в противном же случае его производительность может ухудшиться или результаты могут быть искажены.

При оценке алгоритма обучения с учителем на практике целесообразно рассмотреть, как классификатор будет поступать с адекватными учебными примерами, когда наступит необходимость переобучения и как пользователь будет обнаруживать приложения нового типа.

Может показаться, что автоматическое определение классов посредством распознавания „естественных“шаблонов (кластеров) в наборе данных является одним из преимуществ алгоритма машинного обучения без учителя. Тем не менее, полученные кластеры по-прежнему необходимо маркировать для того, чтобы новые экземпляры могли быть правильно соотнесены с приложениями.

При оценке алгоритма обучения без учителя на практике целесообразно рассмотреть, как могут быть промаркированы кластеры (соотнесены с конкретными приложениями), как будет обновляться маркировка по мере обнаружения новых приложений, и оптимальное количество кластеров (баланс точности, стоимости маркировки и поиска метки, вычислительной сложности) [1].

5. Заключение

В статье рассмотрены современные подходы к классификации IP-трафика физических каналов сети Интернет. Обозначены основные методы, применяемые для классификации IP-трафика, приведены их достоинства и недостатки. Подробно рассмотрена классификация IP-трафика на основе полезной нагрузки. При рассмотрении классификации трафика на основе статистических методов приведено сравнение способов машинного обучения. Предложено применение классификации трафика для объединения физических каналов сети Интернет в группы со схожими характеристиками. Определение критериев принадлежности каналов одной группе является дальнейшим направлением исследований.

ЛИТЕРАТУРА

1. T. T.T. Nguyen, G. Armitage. A Survey of Techniques for Internet Traffic Classification using Machine Learning // Communications Surveys & Tutorials, IEEE 2008, V. 10 (4), P. 56-76.
2. Internet Assigned Numbers Authority (IANA). Service Name and Transport Protocol Port Number Registry, `http://www.iana.org/assignments/port-numbers`
3. F. Risso, A. Baldini, M. Baldi, P. Monclus, O.Morandi. Lightweight, Session-Based Traffic Classification // Proceedings of the IEEE International Conference on Communications (ICC 2008) - Advances in Networks & Internet Symposium, Beijing, China, May 2008.
4. M. Baldi, F. Risso. NetPDL: An Extensible XML-Based Language for Packet Header Description // Elsevier Computer Networks Journal (COMNET), Volume 50, Issue 5, Pages 688-706, April 2006.
5. Cisco Systems. Network Based Application Recognition (NBAR), `http://www.cisco.com/c/en/us/products/ios-nx-os-software/network-based-application-recognition-nbar/index.html`
6. O. Reviv. Inside network programming with SML // EE Times, August 2003.
7. R. Pang, V. Paxson, R. Sommer, L. Peterson. Binpac: a yacc for writing application protocol parsers // Proceedings of the 6th ACM SIGCOMM on Internet Measurement, pages 289-300, Rio de Janeiro, Brazil, October 2006.
8. T. Karagiannis, K. Papagiannaki, and M. Faloutsos. Blinc: Multilevel traffic classification in the dark // Proceedings of ACM SIGCOMM, Philadelphia, PA, August, 2005.

EVALUATION OF FUNCTIONALITY'S EFFICIENCY IN PRIORITY TELECOMMUNICATION NETWORKS WITH HETEROGENEOUS TRAFFIC

I.V.Kalinin[1], *L.A. Muravyeva-Vitkovskaya*[2]

[1,2] ITMO University, Sablinskaya str. 14, 197101, Saint-Petersburg, Russia
[1]kalinin@cs.ifmo.ru, [2]muravyeva-vitkovskaya@cs.ifmo.ru

Abstract

Analytic dependencies for message delivery time computation from sender node to recipient node in priority telecommunication networks with heterogeneous traffic are described in this article.

Keywords: computer network, calculation of intensity, message delivery time, message flow, queuing system.

1. Introduction

Telecommunication network's (TCN) specificity is conditioned on the following factors: variety of network technologies and architectures, variety of QoS requirements for different data types (e.g. the main important text file transmitting index is delivery reliability, guarantees the data loss and distortion absence in received file, and as for audio and video data the main important index is delay jitter for received data packets); traffic heterogeneity, where traffic is controlled by different methods (LAN accessing methods, routing algorithms, establishing connections methods and etc.), aimed at preserving overload and blocking in networks and provide required QoS.

Contemporary TCNs are characterized by variety of provided services, by increasing number of users and amount of transmitted data; by raising level of user's QoS requirements. TCN's requirements can be achieved by selecting structural and functional organization, including such issues as choosing specific technology of data transmitting and processing, defining the most rational communication network topology, choosing network equipment, traffic management mechanisms and etc.

Now days due to intensive growing number of users and applications multiserver networks become more and more common with it's dozens of traffic types [1], caused by new information technology's implementation and using various type applications: Internet, VoIP (Voice over IP), videoconferencing, enterprise resource planning (ERP), client relations management (CRM) and etc. Thus the traffic's heterogeneity [1, 2] is one of the TCN's accents. Traffic's heterogeneity lies in transmitting different types of data packets over TCN (video and audio data packets, speech data packets, text data packets and etc.),

where different delivery requirements exist [3]. This circumstance must be kept in network administrator's mind for increasing using TCN's resources efficiency. Traffic's prioritizing is one of the ways to distribute network resources according to existing priorities.

Building networks using switches allows enabling network technology independent traffic's prioritization for increasing user traffic's servicing quality. This new capability in comparison with networks built only upon hubs is consequence of buffering data frames in switches before sending them to another port. Usually switch keeps not one but some queues for every input and output ports, where every queue has it's own processing priority.

Queue processing algorithms are the one of the basic methods of QoS in network elements.

Solving TCN design problems assumes using efficient models and mathematical methods, which give an opportunity to make qualitative and quantitative analysis of TCN's functioning characteristics depending on structural, functional and load parameters. In the TCN's analyzing process one of the major characteristics being determined is delivery time from sender to recipient. Results, represented in [4], give on opportunity to calculate only average delivery time for various message types. But in practice not average time but probability of prompt delivery from sender to recipient for different message types is more interesting (e.g. operative, service, dialog, file message types). Moreover, it's necessary to keep in mind a possibility of using priority techniques for information flows management, where priority techniques are based on general discipline with mixed priorities.

To solve a given task an open queue network with heterogeneous request flow can be used. Let's illustrate method of analyzing and getting relatively simple results assuming nodes having non-priority information flows management methods.

2. Method

Problem definition. Let messages of K types circulate in TCN with n nodes. Let us assume that abonent, connected to sender node h, generates markovian exponential flow of k-type messages with $\lambda_k(0\,|h,l)$ intensity to destination recipient node l, this kind of messages will be called "$(h,\ l)$-messages of k-type" for short. Also let us assume that probability $\pi_k(i,j\,|h,l)$ of transferring $(h,\ l)$-messages of k-type from node i to connected node j is defined based on chosen routing algorithm. For each $(h,\ l)$-direction these probabilities $\pi_k(i,j\,|h,l)$ form transfer probability matrix which describes possible transfer routes from node h to node l $(i,\ j,\ h,\ l = \overline{1,\ n};\ k = \overline{1,\ K})$.

Message processing duration in node i, represents processing time in node and transferring time to neighbor node, will be assumed exponentially distributed with central tendency equals to $b_k(i)$. Different type messages are serviced in nodes in incoming order. It is necessary to determine transfer-

ring time distribution law for k-type messages routed from node h to node l ($k = \overline{1,\ K}$; $i,\ j,\ h,\ l = \overline{1,\ n}$).

TCN's analysis method is based on network decomposition and led to computation of separate open queue network's nodes like computation of queue system of $M_K/M_K/1$ type with non-priority request processing. This method gives exact results in case of $b_h(i) = b(i)$ ($i = \overline{1,n}$) for all $k = \overline{1,\ K}$ and approximate otherwise. In addition, measure of result's inaccuracy is decreasing with increasing number of routes and number of types of messages, circulating in TCN by these routes, and also with decreasing difference in message processing duration for different message types in node. It is necessary to determine message flow intensity in every network's node to decompose open queue network.

Message flow intensity computation. Message flow intensity for (h, l)-messages of k-type in node j is defined by the following set of linear equations:

$$\lambda_k(j|h,l) = \sum_{i=0}^{n} \lambda_k(i|h,l)\ \pi_k(i,j|h,l) \quad (j = \overline{0,n}), \tag{1}$$

where for all $i,j,h,l = \overline{1,n}$ $\quad \pi_k(0,0|h,l) = 0$;

$$\pi_k(0,j|h,l) = \begin{cases} 1, & j = h; \\ 0, & j \neq h; \end{cases} \qquad \pi_k(i,0|h,l) = \begin{cases} 1, & i = l; \\ 0, & i \neq l. \end{cases}$$

Thus summary intensity of k-type message flow into node j is

$$\lambda_k(j) = \sum_{h=1}^{n} \sum_{l=1}^{n} \lambda_k(j|h,l) \quad (j = \overline{0,n}), \tag{2}$$

where $\lambda_k(0)$ - is intensity of k-type messages incoming into the network.

Transfer probabilities in the open queue network can be calculated in terms of obtained intensities:

$$p_k(0,0) = 0;$$

$$p_k(i,0) = \sum_{h=1}^{n} \lambda_k(i|h,i)/\lambda_k(i) \quad (i = \overline{1,n});$$

$$p_k(0,j) = \sum_{l=1}^{n} \lambda_k(0|j,l)/\lambda_k(0) \quad (j = \overline{1,n}); \tag{3}$$

$$p_k(i,j) = \sum_{h=1}^{n} \sum_{l=1}^{n} \lambda_k(i|h,l)\ \pi_k(i,j|h,l)/\lambda_k(i)\lambda \quad (i,j = \overline{1,n}).$$

Message flow intensities $\lambda_k(j)$ are connected by the obvious dependency:

$$\lambda_k(j) = \sum_{i=0}^{n} \lambda_k(i)\ .p_k(i,j) \quad (j = \overline{0,n}). \tag{4}$$

The open queue network's node *j=0* in expressions (1) - (4) corresponds to outer environment - the source of incoming and the sink of returning messages.

Message delivery time computation. First, let's determine stay duration of k-type message in data transmission node *i,* considering this node as queue system of $M_K/M_K/1$ type, which is receiving K exponential message flows with intensities $\lambda_k(i)$ $(k=\overline{1,K};\ i=\overline{1,n})$. In case of non-priority message processing in node i the Laplace transformation for probability density of stay duration of k-type message is defining in the following way [5]:

$$U_k^*(i,s)=\frac{[1-R(i)][1+s\,b_k(i)]^{-1}\ s}{s-\Lambda(i)+\sum_{r=1}^{K}\lambda_r(i)/[1+s\,b_r(i)]}, \tag{5}$$

where $\Lambda(i)=\sum_{k=1}^{K}\lambda_k(i);\ R(i)=\sum_{k=1}^{K}\lambda_k(i)\,b_k(i);\ i=\overline{1,n};\ k=\overline{1,K};\ s>0$.

The Laplace transformation $V_k^*(h,l,s)$ for probability density of delivery time of k-type messages from node h to node l defines by the following set of equations:

$$V_k^*(h,l,s)=U_k^*(h,s)\sum_{j=1}^{n}\pi_k(h,j|h,l)\,V_k^*(j,l,s)\quad(h,l=\overline{1,n}), \tag{6}$$

where $V_k^*(h,h,s)=U_k^*(h,s)$.

Two first initial moments are determined from the following set of equations:

$$V_k(h,l)=U_k(h)+\sum_{j=1}^{n}\pi_k(h,j|h,l)\,V_k(j,l)\quad(h,l=\overline{1,n}); \tag{7}$$

$$\begin{aligned}&V_k^{(2)}(h,l)=U_k^{(2)}(h)+2[V_k(h,l)-U_k(h)]\,U_k(h)+\\&+\textstyle\sum_{j=1}^{n}\pi_k(h,j|h,l)\,V_k^{(2)}(j,l)\quad(h,l=\overline{1,n}),\end{aligned} \tag{8}$$

where $U_k(h)$ and $U_k^{(2)}(h)$ — respectively first and second initial moments of stay duration of k-type messages in node *h.* These moments are defined by derivation (5) by s at point *s=0.*

3. Results

It's possible to determine various probabilistic and pertaining to time telecommunication system's characteristics, specifically, probability of message's prompt delivery [6], using Laplace transformations $V_k^*(h,l,s)$ or moments $V_k(h,l)$ and $V_k^{(2)}(h,l)$. Prompt delivery probability equals value of delivery time Laplace transformation calculated with $s=s_k$ if message's aging function is exponential and average aging time for k-type messages equals $1/s$.

Table 1

Direction *(h, l)*	Route *(i, j)*	$p_k(i, j \mid h, l)$	
		$h = 1$	$h = 2$
1,4	1,2	0,9	0,3
	1,3	0,1	0,7
	2,3	0,1	0,7
	2,4	0,9	0,3
2,1	2,1	0,9	0,3
	2,3	0,1	0,7
2,4	2,3	0,1	0,7
	2,4	0,9	0,3
3,1	3,1	0,9	0,3
	3,2	0,1	0,7
3,4	3,2	0,1	0,7
	3,4	0,9	0,3
4,1	4,2	0,1	0,7
	4,3	0,9	0,3

For $(i,j|h,l) = (1,2|1,2);\ (1,3|1,3);\ (3,4|1,4)\ (2,3|2,3);\ (3,1|2,1);$

$(3,4|2,4); (2,1|3,1); (2,4|3,4); (3,2|3,2); (2,1|4,1); (3,4|2,4); (3,1|4,1);$

$(4,2|4,2); (4,3|4,3)$ **and** $h = \overline{1,2}:\ \pi_k(i,j|h,l) = 1.$

Example. Let's examine TCN, containing four nodes and two types of messages circulating in it. Markovian exponential flow of k-type messages generates with $\lambda_k(0|h,l)$ intensity from abonents, connected to sender node h, to destination recipient node l.

$$\lambda_1(0|h,l) = \begin{cases} 0,02\ if\ h < l \\ 0,04\ if\ h > l \end{cases};\lambda_2(0|h,l) = \begin{cases} 0,03\ if\ h < l \\ 0,01\ if\ h > l \end{cases};\quad h,l = \overline{1,4}.$$

Processing durations are the same and equals 2 s. for all messages in every node. Delivery times for k-type messages from sender node h to destination node l are constrained by $\hat{V}_k(h,l) = 50s$ for all $k = \overline{1,2};\ h,l = \overline{1,4}$, where $h \neq l$. Average aging time for k-type message equals 10 *s.* $(k = \overline{1,2})$. Probabilities$\pi_k(i,j|h,l)$, shown in table 1, are defined based upon chosen routing algorithms where $\pi_k(0,h|h,l) = \pi_k(l,0|h,l) = 1$ $(i,j = \overline{1,4};\ h,l = \overline{1,4};\ k = \overline{1,2})$. For this TCN we have message's prompt delivery probabilities $P(V_k(h,l) < \hat{V}_k(h,l))$ $(h,l = \overline{1,4};\ k = \overline{1,2})$, shown in table 2, and prompt delivery probabilities, shown in table 3, granting aging functioning.

Table 2

K	H	l			
		1	2	3	4
1	1	1,000	0,996	0,989	0,984
	2	0,994	1,000	0,973	0,994
	3	0,985	0,973	1,000	0,985
	4	0,973	0,996	0,989	1,000
2	1	1,000	0,996	0,989	0,704
	2	0,764	1,000	0,073	0,764
	3	0,749	0,973	1,000	0,749
	4	0,712	0,996	0,989	1,000

TCN's characteristic's computation method, based on decomposition, also can be used in case of priority message management methods used in TCN's nodes. Stay durations for nodes can be determined as described in [7] in case of using mixed priorities message management methods. Final results, computed by this method, are approximate, because message flows of different types at output, and thus at input, differ from exponential in case of priority management methods. However result's inaccuracies lie in acceptable for engineering computations limits as was discovered in research of flow's characters and their influencing on results in wide range of parameters corresponding to real systems.

Table 3

K	h	L			
		1	2	3	4
1	1	0,667	0,374	0,327	0,233
	2	0,354	0,561	0,275	0,354
	3	0,312	0,275	0,490	0,312
	4	0,211	0,374	0,327	0,667
2	1	0,667	0,374	0,327	0,109
	2	0,174	0,561	0,275	0,174
	3	0,158	0,275	0,490	0,158
	4	0,113	0,374	0,327	0,667

4. Conclusions

Found results can be used for solving TCN's optimization problem, which lies in routing algorithm determination (transfer probabilities $\pi_k(i, j|h, l)$) and in assigning priorities to messages of different types, providing specified message delivery time.

Described method of TCN's computation is implemented in program system.

REFERENCES

1. Jakubovich D. Network traffic optimization. // Communications networks and systems, 2001. – V.10. Pp. 92–97

2. Aliev T.I. Characteristics of mixed priorities message servicing disciplines // Izv. vuzov. Priborostroenie, 2014, V. 4(57), pp. 30–35.
3. Goldstein B.S., Pinchuk A.V., Suhovickiy A.L. VoIP. – M.: Radio and communication, 2001.
4. Basharin G.P., Tolmachev A.L. Queue network's theory and its application to analysis of information-calculating systems // Science and engineering summaries. Probability theory. Mathematical statistics. Theoretical cybernetics. – M.: VINITI T. 21. 1983. – p. 3–119.
5. Kleinrock L. Computaion systems with queues: Russian translation. – M.: Mir, 1979.
6. Zaharov G.P. Research methods of data networks. – M.: Radio and communication, 1982.
7. Aliev T.I., Muravyeva-Vitkovskaya L.A. Prioritetnye strategii upravleniya trafikom v multiservisnykh komp'yuternykh setyakh [Priority-based strategies of traffic management in multiservice computer networks] // Izv. vuzov. Priborostroenie, 2011, vol. 54, no. 6, pp. 44–48.

COMPARATIVE ANALYSIS OF SIMULATION TOOLS FOR BODY AREA NETWORKS

I. Khromov, A. Petukhov
National Research University Higher School of Economics, 34 ul. Tallinskaya, Moscow 123458, Russia

The paper presented the structural model of the wireless network node and provides an overview of existing open modeling systems of wireless networks. We give a detailed comparative analysis showing advantages and disadvantages of each approach and propose the most appropriate simulation system.

СРАВНИТЕЛЬНЫЙ АНАЛИЗ СИСТЕМ МОДЕЛИРОВАНИЯ БЕСПРОВОДНЫХ НАТЕЛЬНЫХ СЕТЕЙ

И. Хромов, А. Петухов

Национальный исследовательский университет «Высшая школа экономики», МИЭМ НИУ ВШЭ. 123458, Москва, ул. Таллинская, д. 34.
ignx@yandex.ru, petukhov-aa08@yandex.ru

Аннотация

В работе рассмотрена структурная модель узла, дается обзор существующих открытых систем моделирования беспроводных сетей. Мы провели подробный сравнительный анализ, показывающий преимущества и недостатки каждого подхода, и предлагаем наиболее подходящую систему моделирования.

Ключевые слова: беспроводная сеть, моделирование, модель, нательная сеть, симулятор.

1. Введение

Нательные сети — класс современных цифровых персональных сетей, работающих вблизи, либо непосредственно через тело человека [1].

Технологические достижения в настоящее время позволяют реализовывать нательные сети по низкой цене с высокой эффективностью. Эти новые устройства могут использоваться в следующих областях: спортивной, военной, обеспечение безопасности, медицинской и в области развлечений. В основном, это маленькие, портативные системы с автономным питанием.

В марте 2007 года была создана рабочая группа TG6, в результате работы которой был разработан стандарт IEEE 802.15.6, в котором определены

уровень доступа к среде (MAC) и физический уровнь (PHY) связи, его часто называют Wireless Body Area Network (WBAN). Работа над стандартом связи IEEE 802.15.6 сейчас успешно завершена, и производители оборудования поддерживают стандарт соответствующими устройствами.

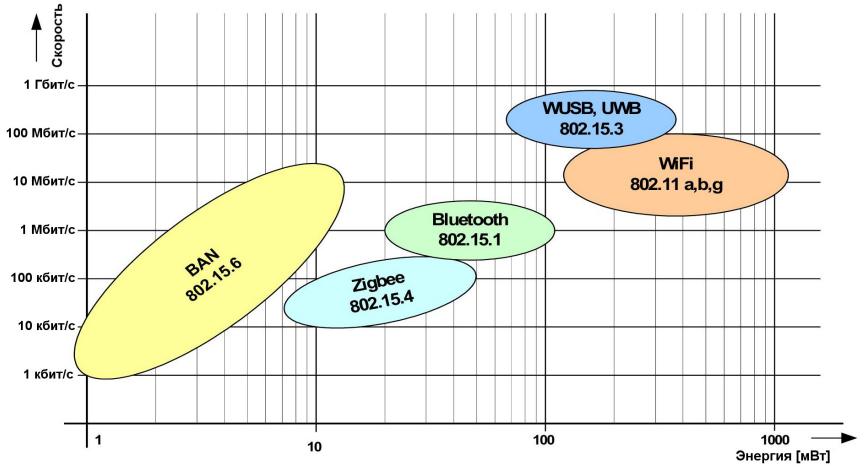


Рис. 1: Стандарты связи в беспроводных нательных сетях.

Передача данных в нательных сетях согласно стандарту IEEE 802.15.6 осуществляется двумя способами: по беспроводным каналам связи или непосредственно через тело человека. Несмотря на то, что средства моделирования беспроводных сетей (БС) на сегодняшний день хорошо проработаны, работа устройств на малой мощности вблизи тела человека предъявляет к ним новые требования, связанные с физикой распространения сигнала [2].

2. Структурная модель узла

В этом разделе описывается структурная модель узла, полученная из [3] и [4]. Эта модель подходит для большинства инструментов оценки, используемых в исследованиях беспроводных нательных сетей (БНС). Рассмотрим структурную модель узла с целью анализа и проанализируем характеристики, которые мы можем моделировать. На рисунке 1 представлена структурная модель узла БНС. Рассмотрим компоненты, из которых она состоит [5]:

- Узлы (nodes). Каждый узел — это физическое устройство, отслеживающее набор физических переменных. Связь между узлами осуществляется через общий канал радиосвязи. Стек протоколов контролирует коммуникации. В отличие от классических моделей сети сенсорные

модели включают вторую группу компонентов — уровень физического узла, который соединен с окружающей средой.

- Окружающая среда (environment). Основное различие между классической и БНС моделями — это дополнительный компонент «окружающая среда». Этот компонент модели генерирует и распространяет события (events). Событиями, как правило, являются физические величины, на которые датчики реагируют. Датчики представляют собой устройства, которые измеряют физические, химические, электрические или оптические воздействия.
- Радиомодуль (transceiver) характеризует распространение радиосигналов между узлами в сети. Очень подробные модели используют компонент «местность» (terrain), связанный с окружающей средой и компонентами радиоканала. Компонент местность учитывается при вычислении распространения радиоволн.
- Микроконтроллер (microcontroller). К нему подключены узлы сбора данных. Он может запросить датчики о событии. Использование узлов сбора данных зависит от применения.

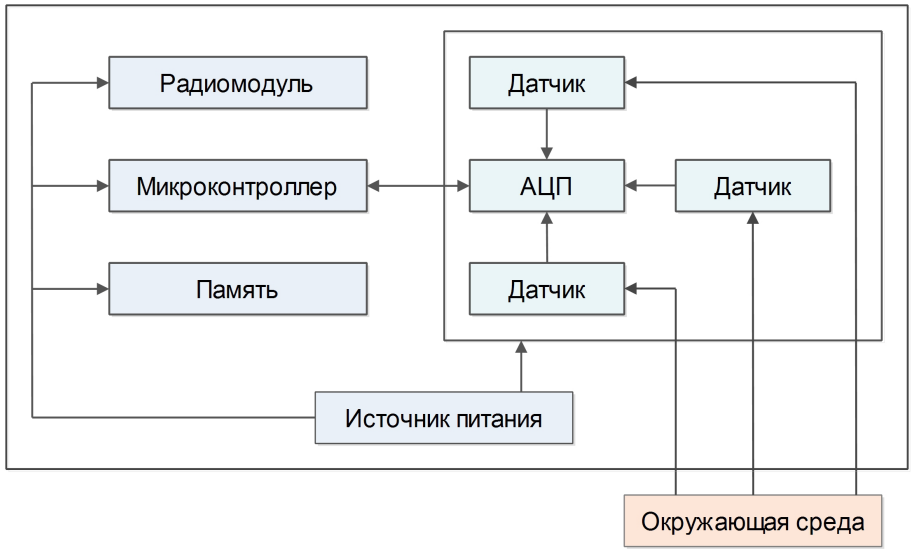


Рис. 2: Структурная модель узла.

3. Системы имитационного моделирования беспроводных сетей

Как правило, ключевыми свойствами для выбора подходящей среды моделирования являются:

- возможность повторного использования и доступность;

- производительность и масштабируемость;
- поддержка богатых семантикой языков сценариев для описания экспериментов и результатов процесса;
- поддержка графики, отладки и трассировки.

Также существует ряд особенностей моделирования нательных сетей:

- мобильность узлов;
- энергопотребление;
- количество датчиков;
- измеряемые физические величины;
- параметры радиомодели..

В этом разделе представлены используемые среды моделирования беспроводных сетей, рассмотрены их основные особенности и вопросы реализации.

3.1. NS-2. NS-2 является дискретным средством имитации событий ориентированным на сетевые исследования. Он имитирует TCP, маршрутизацию, многоадресные протоколы проводных и беспроводных сетей. Симулятор был написан на C++ и объектно-ориентированной версии Tcl, которая называется OTcl (Object Tool Command Language) [6]. К достоинству этого симулятора можно отнести то, что первое время NS-2 моделировал только статические сети TCP/IP, но позже стал поддерживать беспроводные сети, это позволило моделировать мобильные ad-hoc сети и беспроводные сенсорные сети [7]. Недостатками являются: последняя версия вышла в ноябре 2011 года и использование языка OTcl.

3.2. NS-3. NS-3 представляет собой дискретно-событийный симулятор для интернет-систем. Он ориентирован в первую очередь для использования в научных исследованиях и образовательных целях. NS-3 является открытым программным обеспечением, в соответствии с лицензией GNU GPLv2 [8]. Как и его предшественник, NS-3 использует C++ для реализации имитационных моделей. Достоинства:

- NS-3 не использует скрипты OTcl для управления моделированием;
- кроме улучшения производительности был расширен набор его функций.

Преимущества NS-3 перед NS-2: модульность и гибкость трассировки. NS-3 способен имитировать большое количество узлов (20000 и более). Он позволяет исследователям изучать протоколы Интернета и крупномасштабных систем в контролируемой среде. Недостатки: в NS-3 не хватает визуализации окружающей среды (IDE) и мало распространенных моделей.

3.3. OMNeT++. OMNeT++ представляет собой расширяемую, модульную, на основе компонентов C++ платформу, в первую очередь для построения сетевых симуляторов. Предметно-ориентированная функциональность, например, поддержка сенсорных сетей, беспроводных одноранговых сетей, интернет-протоколов, моделирование производительности, фотонных сетей и т.д., обеспечивается дополнительными моделями, разработанными в качестве независимых проектов. Он предлагает графическую среду IDE на основе Eclipse и много других инструментов [9]. OMNeT++ является бесплатным для академического и некоммерческого использования, это широко используемая платформа в мировом научном сообществе. Коммерческие пользователи должны получить лицензию от omnest.com [10]. Достоинства: существуют расширения для моделирования в режиме реального времени, эмуляции сети, альтернативных языков программирования (Java, C#), интеграции с базами данных, SystemC интеграция и некоторые другие функции. Недостатком является отсутствие прямой поддержки моделирования БНС. Для исправления этого недостатка был разработан симулятор Castalia.

3.4. Castalia. Castalia является симулятором с открытым исходным кодом для моделирования БС. Он основан на платформе OMNeT++ и разработан для сетей, построенных на маломощных встроенных устройствах, таких как беспроводные сенсорные узлы [11]. Это программное обеспечение используется исследователями и разработчиками, чтобы проверить свои распределенные алгоритмы и протоколы в реалистичном беспроводном канале, с реалистичным поведением узла. В Castalia узлы модулей OMNeT++ не соединяются друг с другом напрямую, а соединяются через модуль беспроводного канала. Когда узел отправляет пакет, он проходит через беспроводный канал, который затем решает, какие узлы должны получить пакет. Благодаря своему точному моделированию и относительной простоте использования, Castalia получил широкое признание в исследовательском сообществе БС [5]. Недостатком является то, что при установке на операционную систему Windows нельзя представить результаты моделирования в графической форме.

3.5. MiXiM. MiXiM основан на платформе OMNeT++, и создан для моделирования мобильных и стационарных БС (беспроводные сенсорные сети, сети мониторинга тела, ad-hoc сети, автомобильные сети и т.д.). Он предлагает точные модели распространения радиоволн, оценки помех, энергопотребления радиомодуля и беспроводных MAC-протоколов (например, Zigbee) [12]. MiXiM позволяет разработчику применять мощные и многофункциональные инструменты для того, чтобы облегчить моделирование и анализ производительности беспроводных сетей. В то же время структура и дизайн MiXiM такие, что он пытается скрыть сложность моделирования и предоставляет разработчику простой в использовании интерфейс. MiXiM позволяет моделировать 2D и 3D окружающую среду, например, дома или стены. Расположением всех узлов можно управлять с помощью

диспетчера подключений [5]. Недостатком является то, что он не реализует верхние сетевые (сетевой и прикладной) уровни. Этот недостаток исправляется объединением симулятора MiXiM и платформы INET (проект Mixnet).

3.6. TOSSIM. Симулятор разработан для узлов использующих операционную систему TinyOS. Его можно использовать с графическим интерфейсом TinyViz. TOSSIM был разработан для маломощных беспроводных устройств [13]. Преимущества использования [14]:

- простота использования — компиляция исходного кода непосредственно из TinyOS уменьшает сложность и количество ошибок;
- точность — эмулирует оборудование на уровне компонентов и имитирует сеть на уровне битов;
- масштабируемость — возможность использования до тысяч узлов;
- полнота — показывает всю систему поведение и все взаимодействия между отдельными компонентами.

Недостатки:

- не включает моделирование энергопотребления;
- можно улучшить для того, чтобы запускать несколько приложений одновременно;
- применим только для приложений на платформе TinyOS.

Сравнение инструментов моделирования приводится в таблице 1.

	NS-2	NS-3	OMNeT++ Castalia	OMNeT++ MiXiM	TOSSIM
Основные					
Точность моделирования	−	−	+	−	−
Производительность	−	+	+	+	−
Поддержка отладки и трассировки	−	+	+	+	+
Распространенность	+	−	+	−	−
Вспомогательные					
Установка на ОС Windows	−	−	+	−	−
Наличие учебных материалов	−	−	+	−	−
Аппаратные модели	−	−	−	−	+

Таблица 1: Сравнение систем моделирования.

4. Моделирование точности нательной системы сбора данных

В большинстве нательных систем снимаются показания с датчиков (медицинские, датчики поворота, сгиба и т.д.). В связи с этим возникает до-

полнительная задача оценки и прогнозирования точности подобных систем.

Каждое измерительное устройство обладает такими параметрами как точность, разрешающая способность, быстродействие и время отклика. Эти параметры часто влияют друг на друга и могут варьироваться при сборе, оцифровке и передаче данных с сенсоров. Появляется проблема определения точности измерений всей нательной системы сбора данных, состоящей из множества узлов и датчиков. Для ее решения необходимо построить собственную модель, отслеживающую изменения указанных параметров на каждом ее узле, так как существующие системы моделирования этого не позволяют.

Для построения подобной модели необходимо выделить влияющие на точность компоненты системы и типы связи между ними. Таковыми являются датчики (аналоговые и цифровые), микроконтроллеры (снимающие, записывающие и обрабатывающие данные) и аналого-цифровые преобразователи (оцифровывающие аналоговый сигнал). Тип связи между ними может быть проводным (интерфейсы SPI, I2C и др.) и беспроводным (ZigBee, Bluetooth, Wi-Fi и т.д.). Далее для каждого компонента и типа связи определяются соответствующие параметры, которые могут каким-либо образом повлиять на данные с датчиков.

Таким образом, имея набор компонентов системы, их типы связи и параметры, можно исследовать изменения точности в каждом пакете данных с каждого датчика на протяжении всего пути. Такую математическую модель можно написать на C++ и использовать совместно с другими системами моделирования. Это позволит в дальнейшем проектировать нательные системы сбора данных со сложной иерархией и большим количеством измерительных датчиков.

5. Заключение

В работе рассмотрена структурная модель узла, дается обзор существующих открытых систем моделирования беспроводных сетей. Предложены критерии выбора системы моделирования, наиболее подходящей для исследования БНС по сравнению с другими. Моделирования самого по себе недостаточно, чтобы иметь точные результаты, необходим испытательный стенд, для получения результатов в режиме реального времени. Тем не менее, современные средства моделирования позволяют получать результаты близкие к реальности. У каждой системы моделирования есть свои преимущества и недостатки, это дает потенциал для их развития. В результате работы для исследования нательных сетей был выбран симулятор Castalia. Он соответствует наибольшему числу критериев представленных в таблице 1 и его использование позволяет наиболее точно моделировать беспроводные нательные сети.

Данное научное исследование (исследовательский проект № 14–05–0064) выполняется при поддержке Программы «Научный фонд НИУ ВШЭ» в 2014/2015 гг.

ЛИТЕРАТУРА

1. Juraj Miček, Ondrej Karpiš, Peter Ševčík, Body Area Network Analysis and Application Areas // International Journal of Engineering Research and Development, Volume 6, Issue 8, pp. 22–26, April 2013.
2. IEEE Standard Assoication (2012). IEEE Standard for Local and Metropolitan Area Networks — Part 15.6: Wireless Body Area Networks.
3. S. Park, A. Savvides, M. B. Srivastava. SensorSim: A Simulation Framework for Sensor Networks, In Proc. ACM Modeling, Analysis and Simulation of Wireless and Mobile Systems (MSWiM 2000), Boston, MA, pp. 104–111, August 2000.
4. Sobeih, W. Chen, J. C. Hou, L. Kung, N. Li, H. Lim, H. Tyan, H. Zhang, J-Sim: A Simulation and Emulation Environment for Wireless Sensor Networks, In Proc. Annual Simulation Symposium (ANSS 2005), San Diego, CA, pp. 175–187, April 2005.
5. B. Sai Chand, K. Raghava Rao, S. Sreedhar Babu, Exploration of New Simulation Tools for Wireless Sensor Networks // International Journal of Science and Research (IJSR), Volume 2 Issue 4, April 2013.
6. NS-2 [Электронный ресурс]. URL: http://nsnam.isi.edu/nsnam/index.php/Main_Page/ (дата обращения: 16.10.2014).
7. Восков Л.С., Галкин А.А. Средства имитационного моделирования отдельных событий и состояний беспроводных сенсорных сетей // Качество. Инновации. Образование. 2010. №6. С. 37–43.
8. NS-3 [Электронный ресурс]. URL: http://www.nsnam.org/ (дата обращения: 09.11.2014).
9. Хромов И.А., Краюшкин В.В. Сравнительный анализ систем моделирования компьютерных сетей [Текст] // Научно-техническая конференция студентов, аспирантов и молодых специалистов НИУ ВШЭ. Материалы конференции. — М. ∼: МИЭМ НИУ ВШЭ, 2014. — С. 132.
10. OMNeT++ [Электронный ресурс]. URL: http://www.omnetpp.org/ (дата обращения: 17.11.2014).
11. Castalia Wireless Sensor Network Simulator [Электронный ресурс]. URL: http://castalia.research.nicta.com.au/index.php/en/ (дата обращения: 19.11.2014).
12. MiXiM project [Электронный ресурс]. URL: http://mixim.sourceforge.net/ (дата обращения: 23.11.2014).
13. TinyOS [Электронный ресурс]. URL: http://www.tinyos.net/ (дата обращения: 10.12.2014).
14. TOSSIM [Электронный ресурс]. URL: http://tinyos.stanford.edu/tinyos-wiki/index.php/TOSSIM/ (дата обращения: 19.12.2014).

METHODS OF TEST FLYING UBIQUITOUS SENSOR NETWORKS

R. Kirichek[1], *V. Kulik*[2]
The Bonch-Bruevich Saint-Petersburg State University of Telecommunications, Saint-Petersburg, Russia
kirichek@sut.ru[1], vaklicr@gmail.com[2]

Abstract

Public flying ubiquitous sensor networks is a new application of the Internet of Things. The main purpose of Public Flying Ubiquitous Sensor Networks (FUSN-P) is data collection from the sensor fields and delivery for further analysis using public unmanned aerial vehicles (UAV-P) to the base station. The article describes the approach of testing the wireless link IEEE 802.15.4 between UAV-P and terrestrial segment FUSN-P. It considers the structure of tests and describes full-scale experiment using the developed software. It defines the rational distance between the UAV-P and the terrestrial segment FUSN-P.

Keywords: Flying ubiquitous sensor networks, public unmanned aerial vehicles, Delay-Tolerant Networks, Testing

1. FUSN ideology

Increasing of the number of sensor nodes USN which are used for monitoring objects that are distributed over a large area, such as vineyards, seismically dangerous objects and other border strip immediately identified these networks from the rest of the self-organizing networks in a separate class [1, 2] - the terms of wireless sensor networks [3], machine-to-machine [4] and sensory field have appeared. Typically, the sensor fields are situated in remote areas that lack the communications channels with a public network for transmitting data and suggest independent power sensor nodes. If these channels are realized (satellite communication, GPRS / 3G, LTE, etc.), the power consumption which is required for data transmission will lead to reduction of the life cycle of the sensor node [5, 6]. Considering the need to collect data from the remote sensor fields and the possibility of applying the public unmanned aerial vehicles for this purpose, a new application of the Internet of Things - (Public Flying Ubiquitous Sensor Networks (FUSN-P) has been formed [7, 8]. Flying ubiquitous sensor networks differ from the well-known class of flying Ad Hoc networks [9, 10, 11] both the interaction with the terrestrial networks and advantageous use of UAV-P's, which requires the use of UAV-P's for flight FUSN-P on pre-elect path [12]. Such networks require the presence of two segments: terrestrial one and flying one, which interact with each other on the basis of protocols LLN [13, 14]: ZigBee, 6LoWPAN, RPL, Bluetooth Low Energy [15] and others.

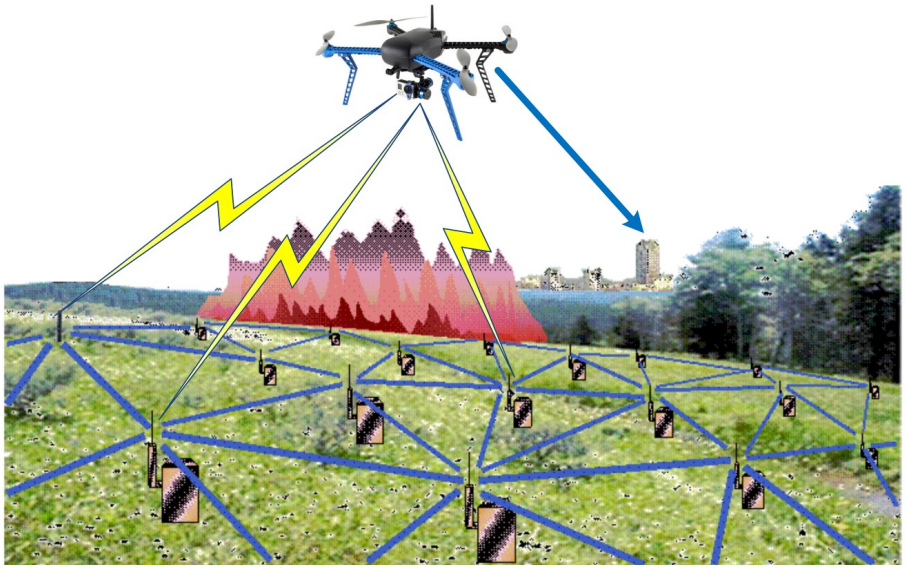


Fig. 1: Structure of the Flying Ubiquitous Sensor Networks.

The terrestrial segment is a distributed network of self-organizing sensor nodes USN, and the flying segment is represented by one or several UAV-P's (Fig. 1).

The problems which are associated with the interaction of terrestrial and flying segments are virtually unexplored. Given that the UAV-P has several radios for different purposes on board: transmission of telemetry, video streaming, interaction with sensory nodes, receiving commands from the control panel and other, there is a possibility of errors due to interference of the signals. For this reason, the data corruption can occur in the physical layer, resulting in delays and losses during the data transmission. In this regard, the development of test methods for FUSN-P is very relevant.

It should be noted that despite the fact that FUSN-P has a large number of features, it is a network and for the testing based on the experience [16] the organizing of a model network [17] is reasonable.

2. Problems of interaction between segments FUSN

Model UAV-P can be implemented on the basis of different flying platforms. There are public unmanned aerial vehicles, manned aircraft types and helicopter types (helicopters and multicopters are the aircraft with four or more rotors with the main rotors). To solve the problems of data collection from the sensor fields using UAV-P and delivery of data to the gateway to the public network, it is advisable to use public multicopters with programming

algorithms fly, installation of the sensor nodes in the given area and optimizing trajectory for data collection. Multicopters can take off and land vertically so that they enable to fly in the areas of unprepared platform which is necessary for the launch and landing. Unlike the airplanes, multicopters have no minimum speed to avoid stalling and can easily hang at the given point. Moreover they are flexible to maneuver in out-of-the-way places (woods, vineyards, etc.). Time of the multicopter autonomous flight is usually from 5 to 35 minutes. To begin piloting the high level of skills is not required.

The structure of a typical multicopter is considered in Figure 2.

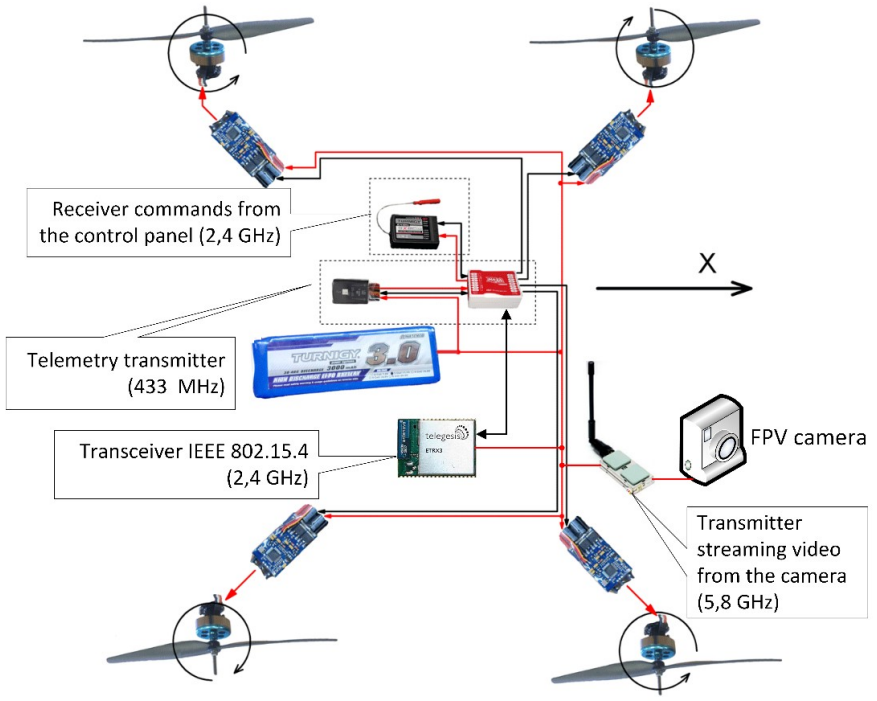


Fig. 2: The structure of a typical radio modules for multicopter FUSN-P.

As it is seen in Figure 2, there are several radios which operate in different frequency bands on board multicopter. For passing parameters of telemetry (the sending is up to 500 parameters at once) there is a radio channel 433 MHz for Europe and 915 MHz for the US to broadcast streaming video from cameras UAV-P, for example, First Person View - flights using 5.8 GHz radio channel. For interaction with the UAV-P terrestrial segment different technologies and protocols are used, but in this article we consider only the technology which is described in standard IEEE 802.15.4, running at 2.4 GHz, as the most common

basic framework for the protocols USN. It is worth noting that of all the above-mentioned radios the channel between the UAV-P and the remote control is mandatory because there can be the emergency of the need for the manual flight control.

According to the overview of the most critical systems of the UAV-P, the errors in the radio channel can lead to some fatal consequences until the fall of UAV-P.In Table 1 there are the possible problems that can appear in the mutual influence of radio channels on each other.

	Delay	**Packet loss**	**Full blocking transmission**
Streaming video from cameras UAV (5,8 GHz)	Artifacts while watching the streaming video on the ground station	Artifacts while watching the streaming video on the ground station	The lack of visual control of the UAV movement on a given route. Inability to detour obstacles on the way
Radio Control channel (2,4 GHz)	Failure to execute commands to change a course or altitude. The possibility of collision with an obstacle.	The possibility of collision with an obstacle.	Fall of the UAV.
FUSN Channel (2,4 GHz)	Increase of a number of the unserved sensor nodes.	The incompleteness of the data which is collected from the sensor fields.	The lack of data from the sensor fields.
Telemetry Channel (433 MHz)	Late delivery of information about the parameters of the flight of the UAV.	Brief interruption of displaying the information about the parameters of flight.	Inability to control the parameters of the flight (only possible visual control in the manual mode)

Table 1: Description of the control problems due to the UAV-P misrepresentation of different radio channels.

In order to identify the mutual influence of different radio channels on each other and as a consequence - the emergence of delays and loss in the data transmission in the Internet of Things laboratory SPbSUT [17], a hardware

and software system for automated testing of the interaction of two segments FUSN -P is developed.

3. Problems of interaction between segments FUSN

To test the radio-channel " UAV-P-sensor node " under the influence of other radio stations the client-server interaction between the flying segment and terrestrial segment FUSN-P based standard IEEE 802.15.4 was organized.

Designed complex consists of three applications:

- The client application (installed on the sensor node);
- The server application (installed on UAV-P's);
- The analyzer of test results (the portable workstation).

While testing, the sensor node with a pre-installed client application switches to send broadcast frames at a predetermined communication channel (Figure 3).

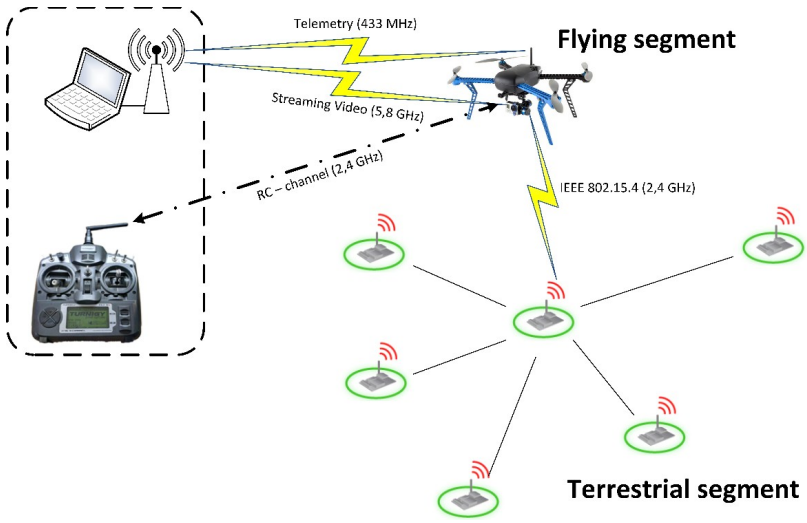


Fig. 3: Schematic testing of FUSN-P.

IEEE 802.15.4 radio was established on multicopter with the pre-installed server application that interacts with the flight controller UAV-P - Pixhawk. During the testing, the server switches to read the data from the radio module and the data broadcast via the UART port on a flight controller. After the test the multicopter returned to the base where it synchronized with the computer to transfer the measurement results. Results Analyzer that is installed on a laptop collects the data from multicopter in the automated mode. Data is written to the database and on the user's request the averaged results of each parameter for each test were formed (Figure 3). During the test IEEE 802.15.4 radios manufactured by Silicon Labs EM357 is used.

To assess the quality of the communication channel "UAV-P-sensor node" a number of parameters has been selected that fully reflects the mutual influence of different radio channels on each other on the physical level:

- Packet Error Rate (PER) — the percentage of the received packets which contain the errors;
- Link Quality Indicator (LQI) — indicator of the quality of the communication channel which displays the "noise pollution" of the communication channel;
- Reciver Strength Signal Indicator (RSSI) — the measure of the received signal strength in dBm;
- Correlated Packets Error (Correlated Packets) — a number of packets that have been subjected to correlation.

All of the above mentioned parameters are monitored at the link layer. In the case of their occurrence there were delays of the data processing, storage of unserved applications in the buffer and packet loss that is unacceptable for the real-time systems. For a comprehensive study of radio "UAV-sensor node" the following structure of the tests was formed:

1) Influence test of the communication channel radio module which operates in the frequency range of 2.4 GHz on the channel IEEE 802.15.4, which operates in the frequency range 2.4 GHz.
2) Influence test of the radio communication channels of the module (2.4 GHz) and the module that broadcasts video streaming from the camera to multicopter which operates in the frequency range of 5.8 GHz, in the channel IEEE 802.15.4 (2.4 GHz).
3) Influence test of the radio communication channels of the module (2.4 GHz) and the telemetry unit which operates in the frequency range of 433 MHz, on the communication channel IEEE 802.15.4 (2.4 GHz).
4) Influence test of the radio communication channels of the module (2.4 GHz), telemetry (433 MHz) and module of the streaming video from a camera (5.8 GHz) on the channel IEEE 802.15.4 (2.4 GHz).

All test parameters are fixed depending on the altitude multicopter: 5 meters, 10 meters, 15 meters, 20 meters, 25 meters and 30 meters respectively. The number of test packets that are sent from the terrestrial to flying FUSN-P segment was 256. At the time of the relevant test the multicopter hung over the sensor node at a specified height for 1 minute.

4. Analysis of test results

During the entire sequence of tests general and partial dependences were revealed. General dependences:

- When the remote radio control UAV-P's is switched on, there is a significant weakening of the power of the received signal with the increasing altitude. Similar results, but with a weaker level of the received signal, is observed in all other experiments.

- While the simultaneous operation of all the radio during the flight of the multicopter, there is a sharp decline in the power of the received signal.
- RSSI signal enables to indirectly estimate the distance to the terrestrial segment of the network in the aerial survey. Figure 4 shows the received signal strength (RSSI) of the distance to the flying segment FUSN-P.

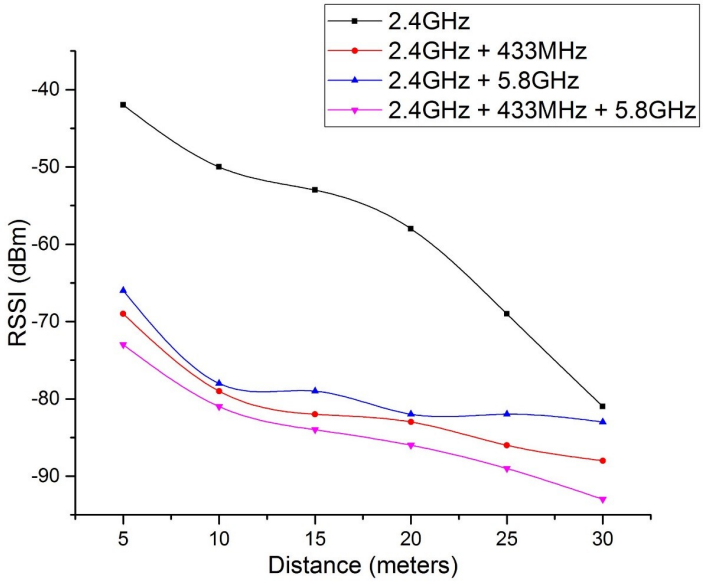


Fig. 4: Dependence of the received signal power to a flying distance segment FUSN-P.

As it is shown in Figure 4, in cases of gradual removal of the terrestrial segment FUSN-P the deterioration in the quality of the communication channel is observed. While the distance from the sensor node to the client is increasing, the power of received signal decreases so that it increases the mutual interference of radio channels with each other and the inability of the receiver to restore the distorted symbols. At a distance of 30 meters it is almost impossible to evaluate the quality of communication.

It is worth noting that the influence test of the channels UAV-P and the radio telemetry on IEEE 802.15.4 there are more "noisy" links.Figure 5 shows the link quality indicator (LQI) of the distance to the flying segment FUSN-P.

When testing the influence of all the radio channels UAV-P on IEEE 802.15.4 channel by increasing the distance between the flying and terrestrial segment FUSN-P, there is a significant an increase in the percentage of packets which are received with errors. For example, at a distance of 20 meters, this value reaches 63% of the total number of sent packets (Figure 6). Thus, with such an amount of errors the collection of data from the terrestrial segment is not

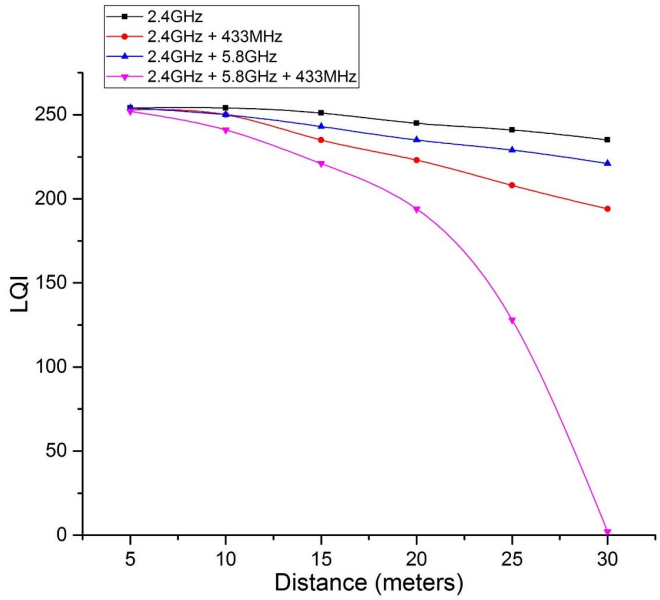


Fig. 5: Dependence of the link quality indicator on the distance to the flying segment FUSN.

possible. Flight altitude UAVs is above 10 meters which is critical for FUSN, in all trials as the loss of more than 5% - 10% leads to considerable delays in the delivery of packages.

While testing with a simultaneous operation of all the radio channels in the channel IEEE 802.15.4, there are a large number of packets which are correlated with an increase in the distance between the UAV-P and the ground segment FUSN (Figure 7). A large number of correlated packets do not quickly and reliably transmit data to the terrestrial segment, which is critical for FUSN-P because of limited battery power on the multicopter. As a result, dependencies can be shown to confirm the conclusion which is obtained in the analysis of the parameter PER: the distance between the terrestrial and flying segments FUSN-P should not be more than 10 meters for all kinds of tests.

5. Conclusion

- The article has considered a test set of the link between the UAV-P and the terrestrial segment FUSN-P. The structure of the tests was suggested and the hardware and software system for automated testing radio FUSN-P service channels under the influence of the UAV-P was developed.
- The distance between the ground and flying segments FUSN-P should not exceed 10 meters for the sustainable functioning FUSN-P.

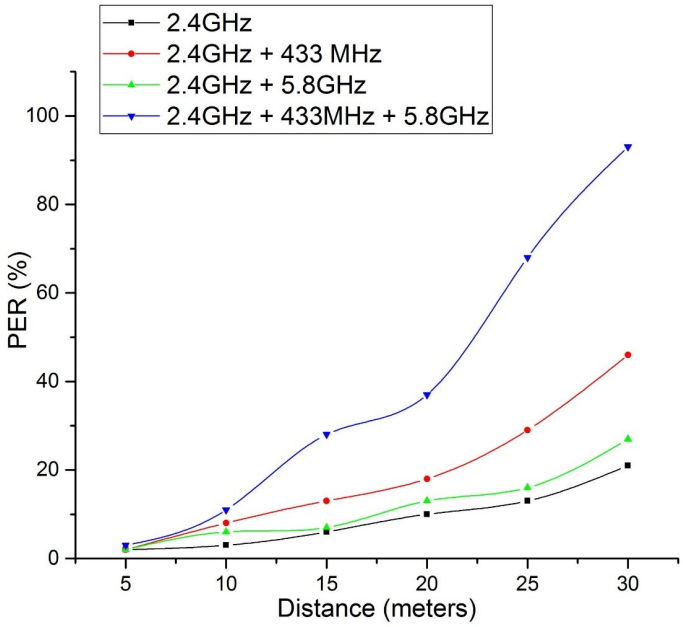


Fig. 6: The dependence of the received packets with errors on the distance to the flying segment FUSN-P.

- In the future we are planning to develop the test specifications and to consider the influence of the errors on the functioning of punctures for low power and lossy networks: ZigBee, 6LoWPAN, RPL etc.

Acknowledgments. The reported study was supported by RFBR, research project No15 07-09431a "Development of the principles of construction and methods of self-organization for Flying Ubiquitous Sensor Networks".

REFERENCES

1. O. Younis and S. Fahmy. Distributed clustering in ad-hoc sensor networks: A hybrid, energyefficient approach. Proceedings, IEEE INFOCOM, Hong Kong, China, 2004.
2. Vinel, A., Vishnevsky, V., Koucheryavy, Y. A simple analytical model for the periodic broadcasting in vehicular ad -hoc networks, 2008 IEEE Globecom Workshops, GLOBECOM 2008
3. I.F. Akyildiz, M.C. Vuran, O.B. Akan, W. Su. Wireless Sensor Networks: A Survey revisited. Computer Networks Journal, 2005
4. Gerasimenko, M. , Petrov, V., Galinina, O., Andreev, S., Koucheryavy, Y. Impact of machine-type communications on energy and delay perfor-

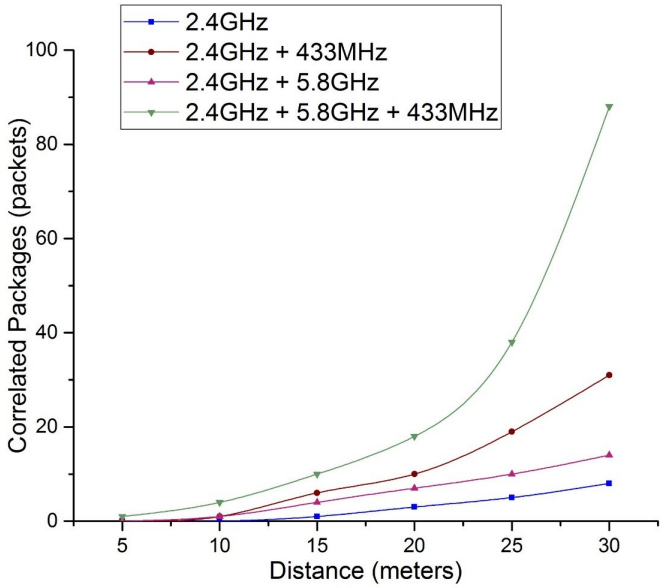


Fig. 7: The dependence of the number of packets received with errors by the distance to the flying segment FUSN.

mance of random access channel in LTE-advanced, European Transactions on Telecommunications, Volume 24, Issue 4, June 2013

5. A. Koucheryavy, A. Salim. Cluster head selection for homogeneous Wireless Sensor Networks. Proceedings, International Conference on Advanced Communication Technology, 2009. ICACT 2009. Phoenix Park, Korea
6. A. Koucheryavy, A. Salim. Prediction-based Clustering Algorithm for Mobile Wireless Sensor Networks. Proceedings, International Conference on Advanced Communication Technology, 2010. ICACT 2010. Phoenix Park, Korea
7. R.Kirichek, A.Paramonov, A.Koucheryavy. Flying Ubiquitous Sensor Networks as a Quening System. Proceedings, International Conference on Advanced Communication Technology, 2015. ICACT 2015, Phoenix Park, Korea.
8. A.Koucheryvy, A.Vladyko, R.Kirichek. State of the Art and Research Challenges for Public Flying Ubiquitous Sensor Networks. The 15th International Conference on Internet of Things, Smart Spaces, and Next Generation Networks and Systems. NEW2AN 2015. LNCS, Springer, Heidelberg, 2015 (accepted).
9. I.Bekmezci, O.K.Sahingoz, S.Temel. Flying Ad-Hoc Networks: A Survey. Ad Hoc Networks, Elsevier, v.11, issue 3, May 2013.

10. O.K.Sahingoz. Networking Model in Flying Ad Hoc Networks (FANETs): Concepts and Challenges. Journal of Intelligent Robotics Systems. V.74, issue 1-2, Springer, 2014.
11. D.Rosario, Z.Zhao, T.Braun, E.Cerqueira, A.Santos. A Comparative Analysis of Beaconless Opportunistic Routing Protocols for Video Dissemination over Flying Ad-hoc Networks. The 14th International Conference on Internet of Things, Smart Spaces, and Next Generation Networks and Systems. NEW2AN 2014. LNCS 8638, Springer, Heidelberg, 2014
12. R.Kirichek, A.Paramonov, K.Vareldzhyan. Optimization of the UAV's Motion Trajectory in Flying Ubiquitous Sensor Networks (FUSN). The 15th International Conference on Internet of Things, Smart Spaces, and Next Generation Networks and Systems. NEW2AN 2015. LNCS, Springer, Heidelberg, 2015 (accepted).
13. X.Liu, J.Guo, G.Bhatti, P.Orlik, k.Parsons. Load Balanced Routing for Low Power And Lossy Networks. Wireless Communications and Networking Conference (WCNC), Proceedings. 7-10April, 2013, Snanghai, China, 2013.
14. C.-A.La, L.-O.Varga, M.Heusse, A.Duda. Energy-efficient Multi-hop Broadcasting in Low Power and Lossy Networks. MSWiM'14, Proceedings of the 17th ACM International Conference on Modelling, Analysis and Simulation of Wireless and Mobile Systems. September, 21-26, 2014, Montreal, Canada.
15. A.Muthanna, A.Prokopiev, A.Koucheryavy. Comparision of Protocols for Ubiquitous Wireless Sensor Networks. 6th International Conference on Ultra Modern Telecommunications and Control Systems. ICUMT 2014, St. Petersburg, Russia, 6-8 October 2014, Proceedings.
16. Recommendation Q.3900. Methods of testing and model network architecture for NGN technical means testing as applied to public telecommunication networks. ITU-T, Geneva, July 2006.
17. Kirichek R., Koucheryavy A. Internet of Things Laboratory Test Bed // International Conference on Wireless Communication, Networking and Application. WCNA 2014. — LNEE v.348. — Heidelberg: Springer, 2014. (accepted).

ESTIMATION QUALITY PARAMETERS OF TRANSFERRING IMAGE AND VOICE DATA OVER ZIGBEE IN TRANSPARENT MODE

R. Kirichek[1], *M. Makolkina*[2], *J. Sene*[3], *V. Takhtuev*[4]
The Bonch-Bruevich Saint-Petersburg State University of Telecommunications, Saint-Petersburg, Russia
kirichek@sut.ru[1], makolkina@list.ru[2], senejeanvalery@yahoo.fr[3], sbst@yandex.ru[4]

Abstract

Wireless Sensor Networks, operating on the basis of IEEE 802.15.4, becoming more popularity in all sphere of life. The new direction of using ZigBee Networks is Flying Ubiquitous Sensor Networks. Data communication with terrestrial segment of network is becoming more effectively than the others technologies, due to self-organizing function and low power consumption.

Keywords: Ubiquitous Sensor Networks, multimedia data, ZigBee, QoE, Flying Ubiquitous Sensor Networks (FUSN)

1. Introduction

Today, Ubiquitous Sensor Networks (USN) are deeply integrated in the everyday life of an ordinary person. Research in the field of capability, network security, traffic characteristics are actively conducted by large number of scientists in different countries of the world [1, 2, 3]. Poses some interesting problems FUSN integration and concept of the Internet of Things [4, 5] are creating some interesting problems, for which the USN is a technological basis.

Recently a new direction of development in USN is Flying Ubiquitous Sensor Networks (FUSN) [6] becoming popular. The scope of FUSN is beneficial for industry, agriculture, transport and monitoring of the functioning over various objects.

FUSN is a part of USN, thereby it has a similar principles of organization and architecture. Additionally, this networks have peer-to-peer or hierarchical (cluster) models. Also, there are 2 segments of FUSN: terrestrial and flying.

Since a part of the FUSN USN, it has a similar organization principles and architecture. Such networks can be peer or hierarchical (cluster). Also allocate 2 segments FUSN: ground and flying, which in turn can be peer or hierarchical.

FUSNs are realized by ZigBee specification, 6 LoWPAN, RPL, Bluetooth Low Energy (BLE) and so on [7, 8], which allowing self-organizing networks with clusters topologies, and low cost realization, transferring small amounts of information and characterized low power consumption.

The self-organizing in Flying Ubiquitous Sensor Networks (FUSN) is automatic creation of network topology, self-acting connection of new nodes, automatic choice of routes the packet network without human intervention and this

is the one of typical side of ZigBee specification. The main feature of FUSN is a large coverage of sensors fields for terrestrial segment. Therefore, is considered the possibility of clustering of the terrestrial segment and the using of UAVs as a main node.

The public unmanned aerial vehicles (UAV-P) can be used for FUSN creation and FUSN-P is the network abbreviation in this case.

The voice and video transmission from terrestrial segment to UAV-P is the next investigation task because often it may be the only chance to pass the necessary information to the area of terrestrial sensor fields. There is a positive experience of transfer of voice data over the protocol ZigBee [9]. In connection with this problem the comparison of voice quality and video quality of experience during their transmission from the terrestrial segment for years using different protocols (ZigBee, 6LoWPAN, RPL), and different languages is of great interest for FUSN-P.

2. Goal of Investigation

To assess the quality of voice in different languages and videos from the terrestrial segment, we have chosen the ZigBee protocol, whereas it is commonly known, has many applications in various fields. It is planned to conduct researches and compare with the others protocols.

ZigBee networks are creating of base nodes of 3 types: coordinators, routers and end points [10]. Coordinator is generating network, forming and functioning as the control center and network trust center - setting security policy defining the settings in the process of joining devices to the network, responsible for security keys. Router is transferring packets of data, realizing dynamic routing, restoring routes on network congestion or failure of any device. Routers connect to the coordinator or others routers on the forming of network, and can attaching child devices — routers or end points. Router works in continuous mode, has permanent power consumption and can provide up to 32 end devices. End point can send and receive packets, but does not translate and route. End points can connect to the coordinator or router, but cannot have child devices. Additionally, end points can be converted into sleep mode to conserve battery power. Developers of this specification have documented of transferring small packets of data, mostly text and packets insensitive for delays.

Despite on this, the goal was analyzed opportunities of transferring multimedia data over ZigBee with appropriate quality. Investigation possibilities of transmitting data such as voice, video, image will expand the range of services to end users on the basis of WPAN networks.

3. Experimental Evaluation

To achieve the goal, the following tasks were allocated: investigation of existing algorithms and approaches to transferring data via FUSN; development laboratory test bed of transferring image data and voice; the experiment of broadcast image and voice data; evaluation of the quality obtained results over

the method of assessing the quality of perception, according ITU recommendations [11, 12].

It is known, that developed solutions for voice channel through ZigBee have already existed and studied, but the commercial development perform for specific hardware and software technology are not universal, and the solutions do not involve an assessment of transmission quality, according to the existing standards.

For practical implementation and experimentation was assembled laboratory test bed based on debugging kits by company Silicon Labs, which was

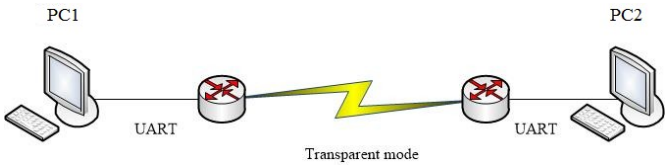


Fig. 1: Laboratory test bed for image transferring.

based on ZigBee modules by Telegesis. Two devices of ETRX3 were selected, satisfying the research conditions, that allow managing the network through the AT command and quickly establish a connection via asynchronous receiver-transmitter (UART). Since, the task to translate the voice and the image, test beds have to include different components, besides the main part of research — the ZigBee network. The wireless network was installed between the two computers, which have simulated a transparent channel (Transparent mode). Scheme of laboratory test bed shown in Fig. 1.

The initial data were selected as an image size of 37 KB, it is obvious that the resolution of the image does not have large, but it is more than enough to recognize objects and analyze their actions. After event from software information is input to the ZigBee ETRX3. The image then is transmitted to the coordinator via the wireless network. The next step is transferring bytes to the computer through a UART interface that can operate at a speed of 115200 bits per second, using the hardware flow control. ETRX3 modules can transmit data in two modes. The first approach is "AT+DMODE" is simulate transparent mode working. This method is transmitted data without acknowledgment, thus a higher data rate is 12.7 Kbit per second. The second method allows transferring numbered packets and confirms the correct data sending. It helps to recover lost packets; this method is "AT+SCAST". Thereby, for analyze network possibilities we needed to take into account different parameters, such as indicators of quality perception, delays and losses, thus a second mode have selected for laboratory test bed —"AT+SCAST". If the loss had occurred, the resulting image would have damaged and as a result, would have not displayed. "AT+SCAST" allows to broadcast data from the router to the coordinator, which marked in the network as a Sink. This module is part of the

network ZigBee, which performing function of receiving all transmitted data over the network. "AT+SCAST" provides correctly transfer and receive image without distortion, but the speed of this process is much slower than the first mode. The speed was 7.5 Kbit per second. The original image was transmitted in 39 seconds.The next experiment has a direction on voice over ZigBee network; the scheme of laboratory test bed shown on Fig. 2.

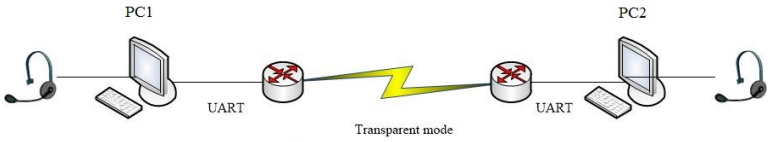


Fig. 2: Laboratory test bed for image transferring.

The goal of this experiment was to transmit voice in real time. The parameters of the experiment are the same as in the previous case, but this test bed has included a microphone for input information and a speaker for outputting information. Each experiment was performed at 15 seconds, the experts were speaking into a microphone in different languages, and the voice is transmitted to an analog-digital converter and then converted into the PCM format. A method for transmitting data via network has selected "AT+DMODE", because this type of traffic is not sensitive to losses, and more dependent on the speed of transmission and delays. The sampling rate was 8000 Hz with 8 quantization bits, and was sending the data to the transmission buffer every 10 ms. Therefore, one packet contained the 80 bits. These characteristics ensured low sound quality; moreover collected on the receiving side for full audio output was failed. Thereby, test bed was using voice compression codec. It was chosen lossy compression algorithm A-Law, compliant with G.711 [13]. Due to the compression of traffic was able to increase quantization up to 16-bit for filling each packet of 80 bits. After transferring data to the PCM format at the transmitting side of test bed traffic was compressed and then was passed via transparent channel. The receiver side decompressed data and output to the speaker, where Assistant assessed results.

One of the important case, connecting with realization network solutions, is provisioning quality of service for each service. Additionally, the requirements for network transmission parameters are specific to different applications and types of traffic. Recommendation ITU-T Y.1540 [14] is identified the following characteristics of the network as the most important in terms of their impact on the quality of service: throughput, network reliability, delays and delay variability, loss ratio, network survivability.

This research is analyzed the influence of network characteristics on the quality of voice and image transmission over a wireless sensor network, according to the existing recommendations. This will determine the further expedi-

ency of using the ZigBee specification for the transmission of multimedia traffic in similar technologies.

Most interesting is the transfer of the image, because today most services are focused on the visualization of information. But the widespread usage of Ubiquitous Sensor Networks suggests that image transmission over ZigBee becomes ordinary technology. For example, if sensor movement in the house or in a protected area, you will be interested to get the picture, what or who caused the alarm. If it ran the hare or other animal, there is no necessity to call special services.

Obviously, the existing Sensor Networks are problematic to transmit image or video with acceptable levels of quality of service. But for the implementation of applications such as "smart infrastructure", the values of some parameters will be enough.

For assessment the quality of the image and voice transmission was chosen methodology proposed in the ITU-T Recommendation R.913. As a subjective evaluation method has chosen method of Absolute Category Rating (ACR). This method uses a categorical estimation. The test sequence evaluated according to the established scale of assessments. The advantage of the method is the ability to evaluate the ACR only received test sequence on the receive side without the etalons, that gets you closer to the real conditions of the network and the estimated of end-users.

The experiment evaluated parameters such as transmission speed, the number of losses, the quality of speech recognition in multiple languages. Evaluations were conducted by four experts. Call duration was 15 seconds. The five experiments for each language provided. The decision to performing experiments on the transmission of speech using fragments in different languages was decided on the basis of the fact that the quality of speech in each language demands different requirements. Subjective assessment may vary from one language to another, due to the fact that languages differ in sound, some more melodic, some more clear and easy to recognize and understand, etc. That is why the experiments were conducted: Russian, English, French, Arabic and Belarusian languages. To reduce the chance of exhibiting lower valuation expert to foreign speech in the expert group included at least one native speaker of each language. Expert estimates are determined according to the following five-point scale: 5 — excellent; 4 — good; 3 — is acceptable; 2 — bad; 1 — unacceptable.

The first stage of the results analysis was calculation of the average rating for each demonstration by the formula (1).

$$\overline{u}_{jkr}=\frac{1}{N}\sum_{i=1}^{N} u_{ijkr} \tag{1}$$

Table 1 illustrates an example of evaluation first fragment of speech in English language.

As can be seen, the results of evaluation fragment English higher than Russian at approximately equal transmission conditions.

Experiment	**Losses, %**	**Speed, bps**	**Expert 1**	**Expert 2**	**Expert 3**	**Expert 4**
1	11,4	353	3,0	3,0	3,5	3,0
2	12,0	312	2,5	3,0	3,0	3,0
3	10,1	412	3,5	4,0	3,5	4,0
4	11,3	368	3,5	3,5	3,5	3,5
5	12,2	302	3,0	3,0	2,5	3,0
Average	11,4	349	3,1	3,3	3,2	3,3

Table 1: Results in English.

This is related to the subjective perception of information by experts. The group of experts contained foreign students who take English familiar, thus there was some tendency inflated estimates, caused by human characteristics.

For further processing of the initial data was calculated the confidence interval by the formula (2), which is derived from the standard deviation of the estimates (3) and the size of each sample, according to ITU BT.500-13. For example, is calculated for the English and Russian language.

$$\delta_{jkr} = 1,96\frac{S_jkr}{\sqrt{N}} \tag{2}$$

$$S_{jkr} = \sqrt{\sum_{i=1}^{N}\frac{(\overline{u}_jkr - u_ijkr)^2}{N-1}} \tag{3}$$

Similarly, the confidence intervals were calculated for the other languages. The results are combined in Table 2.

Language	**Rating**	**Losses, %**	**Speed, bps**	**Confidence interval**
Russian	3,2	11,4	383	(3,06; 3,38)
English	3,4	11,4	349	(3,26; 3,54)
French	3,3	11,4	348	(3,22; 3,48)
Arabic	3,0	11,1	331	(2,81; 3,19)
Belarussian	3,2	11,2	352	(3,04; 3,36)

Table 2: Evaluation MOS for five languages.

Evaluation of subjective methods performed to research the connection between objective indicators of the network and the subjective perception of information by users based on changes during transmission. For displaying of the results using the logistic curve approximation based on the function (4). The result is shown in Figure 3.

$$\sigma = \frac{5}{1 + e^{\frac{-t-t_0}{B}}} \tag{4}$$

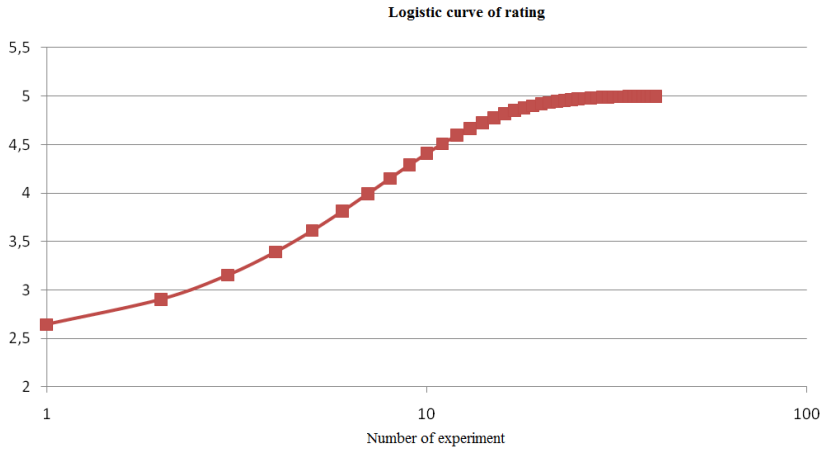


Fig. 3: Logistic curve approximation.

The next step in the analysis was to determine the relationship between the results. To determine the correlation of subjective assessments transmission quality, lossy and speed of transmission was calculated the correlation coefficient. Based on the fact, that the coefficient of correlation between subjective assessment and loss - 0.904, and between the speed of transmission - 0.318, it can be concluded that the estimates are more dependent on the number of lost packets, rather than on the speed of transmission over the network ZigBee.

4. Conclusions

According to the results of the study, the following conclusions:

- In case of low speed and high losses, quality of voice transmission significantly inferior of traditional communication networks. The number of losses provides a large impact of transferring data via ZigBee.
- At that moment transferring video over ZigBee is problematic in that case of low bandwidth of channel. The transmission of low resolution images, makes rational requirements for latency and jitter.
- The compression of voice by codecs with losses of significant information increases probability of packet delivery. On the other hand, if packet losses in network, user will lose significant number of data. The choice of codec with enhanced compression algorithm will improve results.

Thus, the scope of the using ZigBee networks for transmitting multimedia information is specific. These networks are useful where requirements to the quality of voice and image transmission rate low. The main advantage of using a networking standard ZigBee is a low cost, high autonomy, simple creation and survivability of these networks, which allows us to consider the networks for the transmission of multimedia data in the future. Video and voice are the only one way to communicate over FUSN in some cases. Especially, if it

realized in the countryside or outback. Whereas quality of transferring voice and image is satisfactory for all tested languages, then this direction can take important place in FUSN.

Acknowledgments. The reported study was supported by RFBR, research project No 15 07-09431a "Development of the principles of construction and methods of self-organization for Flying Ubiquitous Sensor Networks".

REFERENCES

1. Koucheryavy, A., Prokopiev, A.: Ubiquitous Sensor Networks Traffic Models for Telemetry Applications. In: Balandin, S., Koucheryavy, Y., Hu, H. (eds.) NEW2AN/ruSMART 2011. LNCS, vol. 6869, pp. 287-294. Springer, Heidelberg (2011).
2. Koucheryavy A., Vybornova A. "Ubiquitous Sensor Networks Traffic Models for Medical and Tracking Applications" in The 12th International Conference on Next Generation Wired/Wireless Networking NEW2AN 2012. Aug. 2012 Saint-Petersburg. Springer LNCS 7469, pp. 338-346. Springer, Heidelberg (2012).
3. ITU-T Recommendation Y.2060. Overview of Internet of Things. Geneva, February 2012, Geneva.
4. A.Iera, C.Floerkemeier, J.Mitsugi, G.Morabito. "The Internet of Things." IEEE Wireless Communications. Dec. 2010, v.17
5. R.Kirichek, A.Paramonov, A.Koucheryavy. Flying Ubiquitous Sensor Networks as a Quening System. Proceedings, International Conference on Advanced Communication Technology (ICACT 2015), Phoenix Park, Korea.
6. IEEE Standard for Local and metropolitan area networks–Part 15.4: Low-Rate Wireless Personal Area Networks (LR-WPANs):IEEE 802.15.4.–2011.
7. RPL: IPv6 Routing Protocol for Low-Power and Lossy Networks: RFC 6550. — 03.2012.
8. Touloupis, A.Meliones, S.Apostolacos. Speech codes for high-quality voice over ZigBee applications: Evaluation and implementation challenges. IEEE Communications Magazine, v.50, issue 4, April 2012.
9. ZigBee Specfication, January 17, 2008. ZigBee Alliance.
10. ITU-R Recommendation BT.500-13. Methodology for the Subjective Assessment of the Quality of Television Pictures. Geneva, January, 2012.
11. ITU-T Recommendation P.910. Subjective video quality assessment methods for multimedia applications. Geneva, April, 2008.
12. ITU-T Recommendation G.711. Pulse code modulation (PCM) of voice frequencies. Amendment 1: New Annex A on lossless encoding of PCM frames (1998) Amd.1 (08/2009).
13. ITU-T. Recommendation Y.1540. Internet protocol data communication service — IP packet transfer and availability performance parameters. Geneva, November, 2007.
14. ITU-T Recommendation P.913. Methods for the subjective assessment of video quality, audio quality and audiovisual quality of Internet video and distribution quality television in any environment. Geneva, January, 2014.

QUEUEING SYSTEM WITH TIMER AND RESERVED SERVER

Klimenok V., Shumchenya V.
Belarusian State University, Minsk, Republic of Belarus

Abstract

In the paper, we investigate a single server queue with waiting room and a reserve server which can be used for the modeling of energy saving schemes in some telecommunication networks. An arriving customer is served by the main server until the end of the service or until the timer expires. In latter case, the reserve server joins to the service so that the service rate increases. This allows to avoid too much delay in the system with reasonable energy saving.

СИСТЕМА МАССОВОГО ОБСЛУЖИВАНИЯ С ТАЙМЕРОМ И РЕЗЕРВНЫМ ОБСЛУЖИВАЮЩИМ ПРИБОРОМ

В. Клименок[1], *В. Шумченя*[2]
[1,2] Белорусский государственный университет, Минск, Беларусь
[1] klimenok@bsu.by, [2] fridenn@tut.by

Аннотация

В статье исследуется однолинейная система массового обслуживания с ожиданием и резервным прибором, которая может быть использована при моделировании схем энергосбережения в некоторых реальных системах передачи и обработки информации. Любая заявка, поступившая в систему, обслуживается основным прибором до тех пор, пока не обслужится до конца либо пока не сработает установленный в начале обслуживания таймер. Если заявка еще не обслужилась, а таймер сработал, то к обслуживанию данной заявки присоединяется резервный прибор и скорость передачи увеличивается. Это позволяет избежать слишком больших задержек в системе в условиях разумной экономии энергии.
Ключевые слова: система массового обслуживания, резервный прибор, таймер, стационарное распределение, время пребывания

1. Введение

Проблемы, связанные с энергосбережением во многих реальных системах, в частности, в центрах обработки данных при облачных вычислениях, могут решаться путем резервирования, с дальнейшим адаптивным подключением, обслуживающих ресурсов. В ненадежных системах передачи

данных наличие резервных каналов позволяет повысить качество передачи. Вследствие стохастического характера обработки и передачи информации актуальным является математическое моделирование систем с резервированием в рамках теории массового обслуживания. Упомянем лишь некоторые публикации в данной области.

В публикациях [1]–[4] рассмотрены математические модели гибридных систем связи, состоящих из ненадежного FSO (Free Space Optics) канала и резервного абсолютно надежного радиоканала либо из FSO канала и резервного канала миллиметрового диапазона, которые могут выходить из строя в непересекающихся интервалах времени. Статьи [5], [6] посвящены задаче минимизации потребления энергии в центрах обработки данных при сохранении допустимого уровня обслуживания клиентов. Здесь предполагается, что при увеличении загрузки системы до некоторого порогового уровня к работе могут подключаться приборы из резерва и при снижении загрузки до некотого другого уровня эти приборы снова переходят в холодный резерв. В [7] соответствующие модели расширены на случай нетерпеливых клиентов, которые уходят из системы необслуженными, если время пребывания их в очереди превышает некоторую случайную величину. Аналогичная модель рассмотрена в [8].

В данной статье исследуется однолинейная система массового обслуживания, которую также можно использовать при решении задачи нахождения копромисса между потреблением энергии и качеством обслуживания за счет использования резервного прибора, который подключается к обслуживанию заявки в случае превышения предельного времени ее пребывания на приборе. Данную систему можно также рассматривать как модель ненадежной системы, где при выходе из строя основного прибора дообслуживание заявки производится резервным прибором.

2. Описание системы

Рассматривается однолинейная система массового обслуживания с бесконечным буфером, в которую поступает стационарный пуассоновский поток заявок интенсивности λ.

Время обслуживания заявки имеет фазовое распределение (PH – Phase type distribution) с неприводимым представлением $(\boldsymbol{\beta}, S)$. Это означает следующее. Время обслуживания интерпретируется как время, за которое цепь Маркова m_t, $t \geq 0$, с пространством состояний $\{1, \ldots, M+1\}$ достигнет поглощающего состояния $M+1$. Переходы цепи m_t, $t \geq 0$, с пространством состояний $\{1, \ldots, M\}$ задаются субгенератором S, а интенсивности переходов в поглощающее состояние задаются вектором $\boldsymbol{S}_0 = -S\boldsymbol{e}$. Когда обслуживание начинается, состояние процесса m_t, $t \geq 0$, выбирается из множества $\{1, \ldots, M\}$ на основании вероятностного вектора-строки $\boldsymbol{\beta}$. Полагаем, что матрица $S + S_0\boldsymbol{\beta}$ является неприводимой. Интенсивность обслуживания задается как $\mu = -(\boldsymbol{\beta}S^{-1}\mathbf{e})^{-1}$.

Кроме основного обслуживающего прибора в системе имееется резервный прибор. Резервный прибор подключается к обслуживанию текущей заявки, если время обслуживания этой заявки превышает некоторое предельное время, которое определяется как случайная величина (таймер), имеющая PH распределение с неприводимым представлением $(\boldsymbol{\tau}, T)$ и пространством состояний управляющего процесса $(1, 2, \ldots, R)$. Интенсивность таймера определяется как $\kappa = -(\boldsymbol{\tau} T^{-1}\mathbf{e})^{-1}$, $\boldsymbol{T}_0 = -T\mathbf{e}$. При подключении к обслуживанию резервного прибора оба прибора начинают совместное обслуживание заявки, которое продолжается с той фазы, на которой сработал таймер. Можно также интерпретировать подключение резервного прибора как замену основного при его поломке. Во всяком случае, чтобы отразить тот факт, что обслуживание заявки при подключении резервного прибора продолжается с текущей фазы и должно пройти все оставшиеся фазы, предполагаем, что в момент окончания таймера субгенератор S меняется на $\tilde{S} = \alpha S$. Здесь значение параметра α (меньше, равно или больше единицы) выбирается в зависимости от того, что предполагается: оставшееся время обслуживания в среднем меньше, равно или больше такового при обслуживании только основным прибором.

3. Процесс, описывающий функционирование системы

Пусть в момент времени $t, t \geq 0$,

- i_t - количество заявок в системе, $i_t \geq 0$,
- $r_t = \begin{cases} 0, \text{ если резервный прибор не задействован в момент } t, \\ 1, \text{если резерв. прибор помогает обслуживать заявку в момент } t; \end{cases}$
- m_t - состояние управляющего процесса обслуживания на основном приборе, работающем без помощи резервного прибора, $m_t = \overline{1, M}$;
- η_t - состояние управляющего процесса таймера на основном приборе, работающем без поддержки, $\eta_t = \overline{1, R}$;
- $\tilde{m}_t$ - состояние управляющего процесса обслуживания на основном приборе, работающем с поддержкой резервного прибора.

Процесс изменения состояний системы описывается регулярной неприводимой цепью Маркова $\xi_t, t \geq 0$, с непрерывным временем и пространством состояний

$$\Omega = \{(0);\ (i, 0, m, \eta),\ i \geq 1, m = \overline{1, M}, \eta = \overline{1, R};\ (i, 1, \tilde{m}),\ i \geq 1, \tilde{m} = \overline{1, M}.\}$$

Нетрудно видеть, что число состояний, входящих в любое подмножество со значением $i > 0$, равно $MR + M$.

Далее будем предполагать, что состояния цепи $\xi_t, t \geq 0$, упорядочены следующим образом. При фиксированных значениях компонент i, r упорядочим состояния в лексикографическом порядке возрастания остальных компонент. Обозначим полученные упорядоченные множества как $\Omega_{i,r}$, и все множество состояний Ω упорядочим следующим образом:

$$(0), \Omega_{1,0},\ \Omega_{1,1},\ \Omega_{2,0},\ \Omega_{2,1},\ \Omega_{3,0},\ \Omega_{3,1} \ldots$$

Теорема 1. *Инфинитезимальный генератор Q цепи Маркова ξ_t, $t \geq 0$, имеет следующую блочную структуру*

$$Q = \begin{pmatrix} Q_{0,0} & Q_{0,1} & O & O & \cdots \\ Q_{1,0} & Q_0 & Q_1 & O & \cdots \\ O & Q_{-1} & Q_0 & Q_1 & \cdots \\ O & O & Q_{-1} & Q_0 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix},$$

где

$$Q_{0,0} = -\lambda, \quad Q_{0,1} = (\lambda\boldsymbol{\beta} \otimes \boldsymbol{\tau} \,|\, O_{1\times M}), \quad Q_{1,0} = \begin{pmatrix} \boldsymbol{S}_0 \otimes \mathbf{e}_R \\ \tilde{\boldsymbol{S}}_0 \end{pmatrix},$$

$$Q_{-1} = \begin{pmatrix} \boldsymbol{S}_0\boldsymbol{\beta} \otimes \mathbf{e}_R\boldsymbol{\tau} & O_{MR\times M} \\ \tilde{\boldsymbol{S}}_0\boldsymbol{\beta} \otimes \boldsymbol{\tau} & O_M \end{pmatrix}, \quad Q_0 = \begin{pmatrix} -\lambda I_{MR} + S \oplus T & I_M \otimes \boldsymbol{T}_0 \\ O_{M\times MR} & -\lambda I_M + \tilde{S} \end{pmatrix},$$

$$Q_1 = \lambda I_{M(R+1)}.$$

Доказательство теоремы выполняется путем анализа вероятностей переходов рассматриваемой цепи на бесконечно малом интервале времени.

Из вида генератора следует, что исследуемая цепь принадлежит классу векторных процессов гибели и размножения, см. например, [9].

4. Стационарное распределение. Вероятностные характеристики системы

Теорема 2. *Необходимым и достаточным условием существования стационарного распределения цепи Маркова ξ_t является выполнение неравенства*

$$\lambda < \mathbf{x}\boldsymbol{\mu}, \tag{1}$$

где вектор $\mathbf{x}$ определяется как единственное решение системы линейных алгебраических уравнений

$$\mathbf{x}(S \oplus T)[I - \mathbf{e}(\boldsymbol{\beta} \otimes \boldsymbol{\tau})] = \mathbf{0}, \tag{2}$$

$$\mathbf{x}\mathbf{e} - \mathbf{x}(I_M \otimes \boldsymbol{T}_0)\tilde{S}^{-1}\mathbf{e} = 1, \tag{3}$$

а величина $\boldsymbol{\mu}$ вычисляется как $\boldsymbol{\mu} = -(S \oplus T)\mathbf{e}$.

Доказательство. Поскольку исследуемая цепь является векторным процессом гибели и размножения, то, согласно [9], необходимым и достаточным условием существования ее стационарного распределения является выполнение неравенства

$$\mathbf{z}Q_{-1}\mathbf{e} > \mathbf{z}Q_1\mathbf{e}, \tag{4}$$

где вектор $\mathbf{z}$ является единственным решением системы линейных алгебраических уравнений

$$\mathbf{z}(Q_{-1} + Q_0 + Q_1) = \mathbf{0},\ \mathbf{z}\mathbf{e} = 1. \tag{5}$$

Представим вектор $\mathbf{z}$ в виде $\mathbf{z} = (\mathbf{x}, \mathbf{y})$, где $\mathbf{x}$ и $\mathbf{y}$ имеют размерности MR и M соответственно. Тогда неравенство (4) запишется в виде

$$\mathbf{x}(\boldsymbol{S}_0 \otimes \mathbf{e}_L) + \mathbf{y}\tilde{\boldsymbol{S}}_0 > \lambda, \tag{6}$$

а система (5) как

$$\mathbf{x}(\boldsymbol{S}_0\boldsymbol{\beta} \otimes \mathbf{e}\boldsymbol{\tau} + S \oplus T) + \mathbf{y}(\tilde{\boldsymbol{S}}_0\boldsymbol{\beta} \otimes \boldsymbol{\tau}) = \mathbf{0}, \tag{7}$$

$$\mathbf{x}(I_M \otimes \boldsymbol{T}_0) + \mathbf{y}\tilde{S} = \mathbf{0}, \tag{8}$$

$$\mathbf{x}\mathbf{e} + \mathbf{y}\mathbf{e} = 1. \tag{9}$$

Выражая из (8) вектор $\mathbf{y}$ через вектор $\mathbf{x}$ и подставляя полученное выражение в неравенство (6) и уравнения (7)-(9), после несложных алгебраических преобразований получаем неравенство (1) и систему (2)-(3). ■

Следствие 1. В случае стационарного пуассоновского потока и экспоненциальных распределений времен обслуживания и таймера условие существования стационарного распределения (1)-(3) сводится к следующему неравенству

$$\lambda < \frac{\alpha\mu}{\alpha\mu + \kappa}(\mu + \kappa).$$

Далее будем предполагать, что неравенство (1) выполняется.

Упорядочим стационарные вероятности в соответствии с определенным выше порядком расположения состояний цепи и сформируем векторы-строки $\boldsymbol{p}_i, i \geq 0$, стационарных вероятностей, соответствующих значению i счетной компоненты.

Векторы $\boldsymbol{p}_i, i \geq 0$, вычисляются с использованием алгоритма, разработанного в [10] для нахождения стационарного распределения многомерных квазитеплицевых цепей Маркова, частным случаем которых являются векторные процессы гибели и размножения.

Вычислив стационарное распределение $\boldsymbol{p}_i$, $i \geq 0$, можно вычислить ряд характеристик производительности рассматриваемой системы. При этом будет полезен следующий результат, позволяющий вычислить факториальные моменты распределения без вычисления бесконечых сумм.

Теорема 3. *Векторная производящая функция* $\boldsymbol{P}(z) = \sum\limits_{i=1}^{\infty} \boldsymbol{p}_i z^i$, $|z| \leq 1$, *удовлетворяет следующему уравнению:*

$$\boldsymbol{P}(z)Q(z) = z[\boldsymbol{p}_1 Q_0 - \boldsymbol{p}_0 \tilde{Q}(z)].$$

Факториальные моменты стационарного распределения могут быть вычислены путем дифференцирования этого уравнения. Однако это сопряжено с проблемой, вызванной тем, что матрица этой системы, $Q(z)$, является вырожденной в точке $z = 1$. Для решения этой проблемы нами разработана следующая вычислительная процедура.

Обозначим через $f^{(n)}(z)$ n-ю производную функции $f(z)$, $n \geq 1$, и $f^{(0)}(z) = f(z)$.

Следствие 2. m-я, $m \geq 0$, производная векторной производящей функции $\boldsymbol{P}(z)$ в точке $z = 1$ вычисляется рекуррентно из системы линейных алгебраических уравнений

$$\begin{cases} \boldsymbol{P}^{(m)}(1)Q(1) = \Gamma^{(m)}(1) - \sum\limits_{l=0}^{m-1} C_m^l \boldsymbol{P}^{(l)}(1)Q^{(m-l)}(1), \\ \boldsymbol{P}^{(m)}(1)Q'(1)\mathbf{e} = \frac{1}{m+1}[\Gamma^{(m+1)}(1) - \sum\limits_{l=0}^{m-1} C_{m+1}^l \boldsymbol{P}^{(l)}(1)Q^{(m+1-l)}(1)]\mathbf{e}. \end{cases} \tag{10}$$

где $$\Gamma^{(m)}(1) = \begin{cases} \boldsymbol{p}_1 Q_0 - \boldsymbol{p}_0 \tilde{Q}(1), & m = 0, \\ \boldsymbol{p}_1 Q_0 - \boldsymbol{p}_0 \tilde{Q}(1) - \boldsymbol{p}_0 \tilde{Q}'(1), & m = 1, \\ -\boldsymbol{p}_0[m\tilde{Q}^{(m-1)}(1) + Q^{(m)}(1)], & m > 1, \end{cases}$$

$$\tilde{Q}^{(m)}(1) = \begin{cases} Q_{0,1}, & \text{если } m = 0, \\ O, & \text{если } m > 0; \end{cases}, \; Q^{(m)}(1) = \begin{cases} Q_{-1} + Q_0 + Q_1, & \text{если } m = 0, \\ Q_0 + 2Q_1, & \text{если } m = 1, \\ 2Q_1, & \text{если } m = 2, \\ O, & \text{если } m > 2. \end{cases}$$

Вычислив стационарное распределение и используя формулу (10), можно вычислить ряд важных характеристик производительности системы. Формулы для вычисления этих характеристик приведены ниже.

- Вероятность того, что основной прибор свободен $P_{idle}^{(1)} = p_0$.
- Среднее число заявок в системе $L = \mathbf{P}'(1)\mathbf{e}$.
- Дисперсия числа заявок в системе $V = \mathbf{P}''(1) + \mathbf{P}'(1) - L^2$.
- Вероятность того, что основной прибор обслуживает заявку без помощи резервного ($P^{(0)}$) и с помощью резервного ($P^{(1)}$).

$$P^{(0)} = \mathbf{P}(1)\begin{pmatrix} \mathbf{e}_{MR} \\ O_M \end{pmatrix}, \quad P^{(1)} = 1 - p_0 - P_0^{(0)}.$$

- Вероятность того, что резервный прибор свободен $P_{idle}^{(2)} = 1 - P_0^{(0)}$.
- Вероятность того, что к обслуживанию произвольной заявки будет подключен резервный прибор (P_{help}) и вероятность того, что основной прибор обслужит заявку самостоятельно ($P_{no-help}$)

$$P_{help} = -(\boldsymbol{\beta} \otimes \boldsymbol{\tau})(S \oplus T)^{-1}(I_M \otimes T)\mathbf{e}, \quad P_{no-help} = 1 - P_{help}.$$

5. Распределение времени пребывания

Пусть $\tilde{v}_t$ – время дообслуживания заявки, находящейся на основном приборе в момент времени t. Пусть также

$$\tilde{V}(0,m,\eta,x)=\lim_{t\to\infty}P\{r_t=0,m_t=m,\eta_t=\eta,\tilde{v}_t<x\},\ m=\overline{1,M},\ \eta=\overline{1,R};$$

$$\tilde{V}(1,\tilde{m},x)=\lim_{t\to\infty}P\{r_t=1,\tilde{m}_t=\tilde{m},\tilde{v}_t<x\},\ \tilde{m}=\overline{1,M},\ x\geq 0.$$

Введем обозначения для преобразований Лапласа- Стилтьеса

$$\tilde{v}(0,m,\eta,u)=\int\limits_0^\infty e^{-ux}d\tilde{V}(0,m,\eta,x),\ \tilde{v}(1,\tilde{m},u)=\int\limits_0^\infty e^{-ux}d\tilde{V}(1,\tilde{m},x),\ Re\,u\geq 0,$$

и векторов- столбцов $\tilde{\mathbf{v}}(0,u)$ и $\tilde{\mathbf{v}}(1,u)$, составленных из этих преобразований путем лексикографического упрядочения компонент m,η в первом случае и компоненты $\tilde{m}$ во втором случае.

Справедлива следующая теорема.

Теорема 4. *Векторы преобразований Лапласа- Стилтьеса времени дообслуживания заявки имеют следующий вид:*

$$\tilde{\mathbf{v}}(0,u)=(uI-S\oplus T)^{-1}[\boldsymbol{S}_0\otimes\mathbf{e}_R+(uI-\tilde{S})^{-1}\tilde{\boldsymbol{S}}_0\otimes\boldsymbol{T}_0], \tag{11}$$

$$\tilde{\mathbf{v}}(1,u)=(uI-\tilde{S})^{-1}\tilde{\mathbf{S}}_0. \tag{12}$$

Доказательство. Используя вероятностную интерпретацию преобразования Лапласа-Стилтьеса, запишем $\tilde{\mathbf{v}}(0,u)$ как

$$\tilde{\mathbf{v}}(0,u)=\int\limits_0^\infty e^{-ut}e^{(S\oplus T)t}(\boldsymbol{S}_0\otimes\mathbf{e}_R)dt+\int\limits_0^\infty e^{-ut}\int\limits_0^t e^{(S\oplus T)x}(I_M\otimes\boldsymbol{T}_0)dxe^{\tilde{S}(t-x)}\tilde{\boldsymbol{S}}_0dt. \tag{13}$$

Вычислим интегралы в правой части (13). Очевидным образом получим выражение (12) для первого слагаемого в (13):

$$\int\limits_0^\infty e^{-ut}e^{(S\oplus T)t}(\boldsymbol{S}_0\otimes\mathbf{e}_R)dt=(uI-S\oplus T)^{-1}(\boldsymbol{S}_0\otimes\mathbf{e}_R). \tag{14}$$

В результате алгебраических преобразований получим следующее выражение для второго слагаемого в (13).

$$\int\limits_0^\infty e^{-ut}\int\limits_0^t e^{(S\oplus T)x}(I_M\otimes\boldsymbol{T}_0)dxe^{\tilde{S}(t-x)}\tilde{\boldsymbol{S}}_0dt=$$

$$= [uI - (S \oplus T)]^{-1}[(uI - \tilde{S})^{-1} \otimes I_R](\tilde{\boldsymbol{S}}_0 \otimes \boldsymbol{T}_0). \quad (15)$$

Подставляя (14)-(15) в (13), получим (11). Формула (12) доказывается очевидным образом. ■

Следствие 3. Векторы средних значений времен дообслуживания заявки имеют следующий вид:

$$\tilde{\mathbf{t}}_0 = -(S \oplus T)^{-1}[I + \tilde{S}^{-1} \otimes T]\mathbf{e},$$
$$\tilde{\mathbf{t}}_1 = -\tilde{S}^{-1}\mathbf{e}.$$

Доказательство следует из формул $\tilde{\mathbf{t}}_0 = -\tilde{\mathbf{v}}'(0,0)$, $\tilde{\mathbf{t}}_1 = -\tilde{\mathbf{v}}'(1,0)$.

Следствие 4. Преобразование Лапласа- Стилтьеса времени обслуживания произвольной заявки вычисляется по формуле

$$v(0,u) = (\boldsymbol{\beta} \otimes \boldsymbol{\tau})\tilde{\mathbf{v}}(0,u).$$

Следствие 5. Среднее значение времени обслуживания произвольной заявки вычисляется по формуле

$$\bar{t} = -(\boldsymbol{\beta} \otimes \boldsymbol{\tau})(S \oplus T)^{-1}[I + \tilde{S}^{-1} \otimes T]\mathbf{e}.$$

Теорема 5. *Преобразование Лапласа- Стилтьеса времени пребывания произвольной заявки в системе имеет вид*

$$v(u) = p_0 v(0,u) + \mathbf{P}(v(0,u)) \begin{pmatrix} \tilde{\mathbf{v}}(0,u) \\ \tilde{\mathbf{v}}(1,u) \end{pmatrix}.$$

Доказательство. Доказательство теоремы следует из формулы полной вероятности с учетом вероятностного смысла преобразования Лапласа- Стилтьеса. ■

Следствие 6. Среднее значение времени пребывания произвольной заявки в системе имеет вид

$$\bar{v} = p_0 \bar{t} + \mathbf{P}(1) \begin{pmatrix} \tilde{\mathbf{t}}_0 \\ \tilde{\mathbf{t}}_1 \end{pmatrix} + L\bar{t}.$$

ЛИТЕРАТУРА

1. Arnon S., Barry J., Karagiannidis G., Schober R., Uysal M. (Eds): Advanced Optical Wireless Communication Systems. Cambridge University Press, 2012.
2. Vishnevsky V.M., Kozyrev D.V., Semenova O.V.: Redundant queueing system with unreliable servers. Proceedings of the 6th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT). Moscow. 2014. P. 383-386.

3. Vishnevsky V.M., Semenova O.V., Sharov S.Yu.: Modeling and analysis of a hybrid communication channel based on free-space optical and radio-frequency Technologies. Automation and Remote Control 72, 345-352 (2013)
4. Sharov S.Yu., Semenova O.V.: Simulation model of wireless channel based on FSO and RF technologies. Distributed Computer and Communication Networks. Theory and Applications (DCCN-2010). P. 368-374. (2010)
5. Mitrani I. Trading power consumption against performance by reserving blocks of servers. Computer performance engineering. Springer Berlin Heidelberg, 2013, P. 1-15.
6. Mitrani I. Managing performance and power consumption in a server farm /Annals of Operation Reasearch. 2013. V.202. P. 121-134.
7. Mitrani I. Service center trade-offs between customers impatience and power consumation /Performance Evaluations. 2011.V. 68. P. 1222-1231.
8. Shwartz C., Pries R., and Tran-Gia P. A queueing analysis of an energy-saving mechanism in data centers / Proceedings of International Conference on Information Networking. 2012. P. 70-75.
9. Neuts M. F. Matrix-geometric solutions in stochastic models. The Johns Hopkins University Press, Baltimore, 1981.
10. Klimenok, V. I., Dudin A.N. Multi-dimensional asymptotically quasi-Toeplitz Markov chains and their application in queueing theory / Queueing Systems. 2006. V.54. P. 245–259.

LOW-PRIORITY QUEUE FLUCTUATIONS IN TANDEM OF QUEUING SYSTEMS UNDER CYCLIC CONTROL WITH PROLONGATIONS

V. M. Kocheganov[1], *A. V. Zorine*[1]

[1] Department of Applied Probability Theory, N. I. Lobachevsky State University of Nizhni Novgorod, Nizhni Novgorod, Russia
kocheganov@gmail.com, zoav1602@gmail.com

Abstract

A tandem of queuing systems is considered. Each system has a high-priority input flow and a low-priority input flow which are conflicting. In the first system, the customers are serviced in the class of cyclic algorithms. The serviced high-priority customers are transferred from the first system to the second one with random delays and become the high-priority input flow of the second system. In the second system, customers are serviced in the class of cyclic algorithms with prolongations. Low-priority customers are serviced when their number exceeds a threshold. A mathematical model is constructed in form of a multidimensional denumerable discrete-time Markov chain. The recurrent relations for partial probability generating functions for the low-priority queue in the second system are found.

Keywords: tandem of controlling queuing systems, cyclic algorithm with prolongations, conflicting flows, multidimensional denumerable discrete-time Markov chain

1. Introduction

Conflicting traffic flows control at a crossroad is one of classical problems in queuing theory. In the literature several algorithms were investigated: fixed duration cyclic algorithm, cyclic algorithm with a loop, cyclic algorithm with changing regimes, etc [1, 2, 3, 4, 5, 6]. However, several (two in our case) consecutive crossroads are of great interest, because in a real-life situation a vehicle having passed one highway intersection finds itself at another one. In other words, an output flow from the first intersection forms an input flow of the second intersection. Hence, the second input flow no longer has an *a priori* known simple probabilistic structure (for example, that of an non-ordinary Poison flow), and knowledge about the service algorithm should be taken into account to deduce formation conditions of the first output flow.

Tandems of intersections were considered by a few authors. In [7] a computer-aided simulation of adjacent intersection was carried out. In [8] a mathematical mode of two intersection in tandem governed by cyclic algorithms was investigates and stability conditions were found. In this paper we assume that the first intersection is governed by a cyclic algorithm while the econd intersection is governed by a cyclic

algorithm with prolongations. In particular, we pay attention to the low-priority queue in the second intersection.

2. The problem settings

Consider a queuing system with a scheme shown in (see Fig. 1). There are four input flows of customers Π_1, Π_2, Π_3, and Π_4 entering the single server queueing system. Customers in the input flow Π_j, $j \in \{1, 2, 3, 4\}$ join a queue O_j with an unlimited capacity. For $j \in \{1, 2, 3\}$ the discipline of the queue O_j is FIFO (First In First Out). Discipline of the queue O_4 will be described later. The input flows Π_1 and Π_3 are generated by an external environment, which has only one state. Each of these flows is a nonordinary Poisson flow. Denote by λ_1 and λ_3 the intensities of bulk arrivals for the flows Π_1 and Π_3 respectively. The probability generating function of number of customers in a bulk in the flow Π_j is

$$f_j(z) = \sum_{\nu=1}^{\infty} p_\nu^{(j)} z^\nu, \quad j \in \{1, 3\}, \tag{1}$$

We assume that $f_j(z)$ converges for any $z \in \mathbb{C}$ such that $|z| < (1 + \varepsilon)$, $\varepsilon > 0$. Here $p_\nu^{(j)}$ is the probability of a bulk size in flow Π_j being exactly $\nu = 0, 1, \ldots$. Having been serviced the customers from O_1 come back to the system as the Π_4 customers. The Π_4 customers in turn after service enter the system as the Π_2 ones. The flows Π_2 and Π_3 are conflicting in the sense that their customers can't be serviced simultaneously. This implies that the problem can't be reduced to a problem with fewer input flows by merging the flows together.

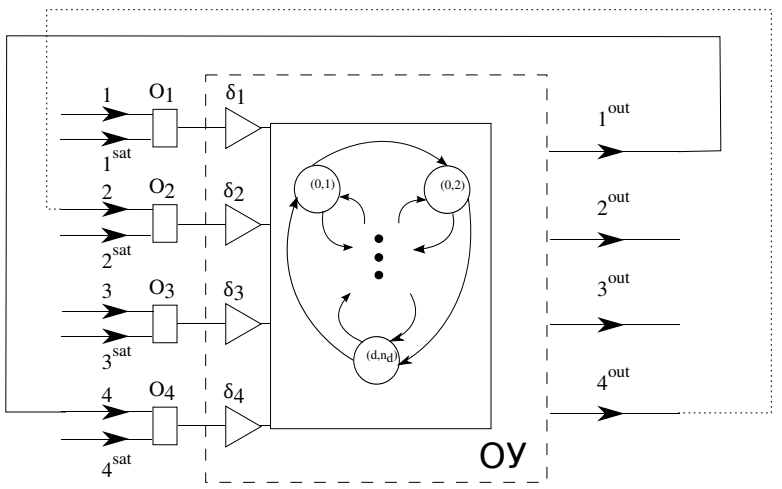


Figure 1: Scheme of the queuing system as a cybernetic control system

In order to describe the server behavior we fix positive integers d, n_0, n_1, …, n_d and we introduce a finite set $\Gamma = \{\Gamma^{(k,r)} \colon k = 0, 1, \ldots, d; r = 1, 2, \ldots n_k\}$ of states server can reside in. At the state $\Gamma^{(k,r)}$ sever stays during constant time $T^{(k,r)}$. Define disjoint subsets Γ^{I}, Γ^{II}, Γ^{III}, and Γ^{IV} of Γ as follows. In the state $\gamma \in \Gamma^{\mathrm{I}}$ only customers from the queues O_1, O_2 and O_4 are serviced. In the state $\gamma \in \Gamma^{\mathrm{II}}$ only customers from the queues O_2 and O_4 are serviced. In the state $\gamma \in \Gamma^{\mathrm{III}}$ only customers from queues O_1, O_3, and O_4 are serviced. In the state $\gamma \in \Gamma^{\mathrm{IV}}$ only customers from queues O_3 and O_4 are serviced. We assume that $\Gamma = \Gamma^{\mathrm{I}} \cup \Gamma^{\mathrm{II}} \cup \Gamma^{\mathrm{III}} \cup \Gamma^{\mathrm{IV}}$. Set also ${}^1\Gamma = \Gamma^{\mathrm{I}} \cup \Gamma^{\mathrm{III}}$, ${}^2\Gamma = \Gamma^{\mathrm{I}} \cup \Gamma^{\mathrm{II}}$, ${}^3\Gamma = \Gamma^{\mathrm{III}} \cup \Gamma^{\mathrm{IV}}$.

The server changes its state according to the following rules. We call a set $C_k = \{\Gamma^{(k,r)} \colon r = 1, 2, \ldots n_k\}$ the k-th cycle, $k = 1, 2, \ldots, d$. For $k = 0$ the state $\Gamma^{(0,r)}$ with $r = 1, 2, \ldots, n_0$ is called a prolongation state. Put $r \oplus_k 1 = r + 1$ for $r < n_k$, and $r \oplus_k 1 = 1$ for $r = n_k$ ($k = 0, 1, \ldots, d$). In the cycle C_k we select a subset C_k^{O} of input states, a subset C_k^{I} of output states, and a subset $C_k^{\mathrm{N}} = C_k \setminus (C_k^{\mathrm{O}} \cup C_k^{\mathrm{I}})$ of neutral states. After the state $\Gamma^{(k,r)} \in C_k \setminus C_k^{\mathrm{O}}$ the server switches to the state $\Gamma^{(k,r\oplus_k 1)}$ within the same cycle C_k. After the state $\Gamma^{(k,r)}$ in C_k^{O} the server switches to the state $\Gamma^{(k,r\oplus_k 1)}$ if number of customers in the queue O_3 at switching instant is greater than a predetermined threshold L. Otherwise, is the number of customers in the queue O_3 is less than or equals L then the new state is the prolongation one $\Gamma^{(0,r_1)}$ where $r_1 = h_1(\Gamma^{(k,r)})$ and $h_1(\cdot)$ is a given mapping of $\bigcup_{k=1}^{d} C_k^{\mathrm{O}}$ into $\{1, 2, \ldots, n_0\}$. After the state $\Gamma^{(0,r)}$ if the number of customers in O_3 is not above L the state of the same type $\Gamma^{(0,r_2)}$ is chosen where $r_2 = h_2(r)$ and $h_2(\cdot)$ is a given mapping of the set $\{1, 2, \ldots, n_0\}$ into itself; in the other case the new state is $\Gamma^{(k,r_3)} \in C_k^{\mathrm{I}}$ where $\Gamma^{(k,r_3)} = h_3(r)$ and $h_3(\cdot)$ is a given mapping of $\{1, 2, \ldots, n_0\}$ to $\bigcup_{k=1}^{d} C_k^{\mathrm{I}}$. We assume that each prolongation state $\Gamma^{(0,r)}$ belongs to the set ${}^2\Gamma$ and that relations $C_k^{\mathrm{O}} \subset {}^2\Gamma$ and $C_k^{\mathrm{I}} \subset {}^3\Gamma$ hold. We also assume that all the cycles have exactly one input and output state. Finally, we assume that all the prolongation states make a cycle, that is $h_2(r) = r \oplus_0 1$. Putting all together, we introduce a function which formalizes the server state changes:

$$h(\Gamma^{(k,r)}, y) = \begin{cases} \Gamma^{(k,r\oplus_k 1)} & \text{if } \Gamma^{(k,r)} \in C_k \setminus C_k^{\mathrm{O}} \text{ or } (\Gamma^{(k,r)} \in C_k^{\mathrm{O}}) \wedge (y > L); \\ \Gamma^{(0,h_1(\Gamma^{(k,r)}))} & \text{if } \Gamma^{(k,r)} \in C_k^{\mathrm{O}} \text{ and } y \leqslant L; \\ \Gamma^{(0,r\oplus_0 1)} & \text{if } k = 0 \text{ and } y \leqslant L; \\ h_3(r) & \text{if } k = 0 \text{ and } y > L. \end{cases} \tag{2}$$

In general, service durations of different customers can be dependent and may have different laws of probability distributions. So, saturation flows will be used to define the service process. A saturation flow Π_j^{sat}, $j \in \{1, 2, 3, 4\}$, is defined as a virtual output flow under the maximum usage of the server and unlimited number of customer in the queue O_j. The saturation flow Π_j^{sat}, $j \in \{1, 2, 3\}$ contains a non-random number $\ell(k, r, j) \geqslant 0$ of customers in the server state $\Gamma^{(k,r)}$. In particular, $\ell(k, r, j) \geqslant 1$ for $\Gamma^{(k,r)} \in {}^j\Gamma$ and $\ell(k, r, j) = 0$ for $\Gamma^{(k,r)} \notin {}^j\Gamma$. Let $\mathbb{Z}_+$ be the set of non-negative integer numbers. If the queue O_4 contains $x \in \mathbb{Z}_+$ customers the saturation flow Π_4^{sat} also contains the x customers. Finally, in the state $\Gamma^{(k,r)}$ every customer from queue O_4

with probability $p_{k,r}$ and independently of others ends servicing and joins Π_2 to go to O_2. With the complementary probability $1 - p_{k,r}$ the customer stays in O_4 until the next time slot. In the next time slot it repeats its attempt to join Π_2 with a proper probability.

A real-life example of just described queuing system is a tandem of two consecutive crossroads (Fig. 2). The input flows are flows of vehicles. The flows Π_1 and

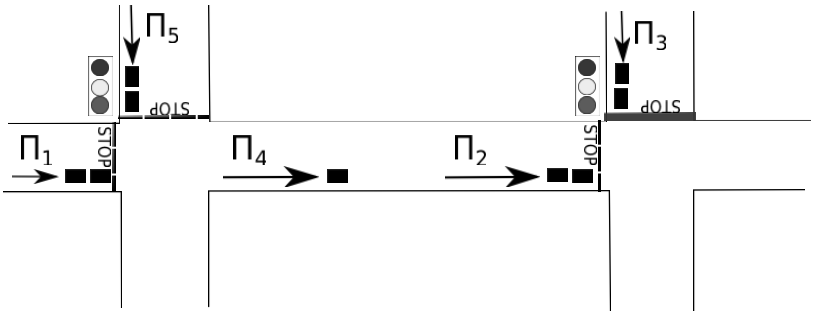


Figure 2: A tandem of crossroads, the physical interpretation of the queuing system under study

Π_5 at the first crossroad are conflicting; Π_2 and Π_3 at the second crossroad are also conflicting. Every vehicle from the flow Π_1 after passing first road intersection joint the flow Π_4 and enters the queue O_4. After some random time interval the vehicle arrives to the next road intersection. Such a pair of crossroads is an instance of a more general queuing model described above.

3. Mathematical model

The queuing system under investigation can be regarded as a cybernetic control system what helps to rigorously construct a formal stochastic model [8]. The scheme of the control system is shown in Fig. 1. There are following blocks present in the scheme: 1) the external environment with one state; 2) input poles of the first type — the input flows Π_1, Π_2, Π_3, and Π_4; 3) input poles of the second type — the saturation flows Π_1^{sat}, Π_2^{sat}, Π_3^{sat}, and Π_4^{sat}; 4) an external memory — the queues O_1, O_2, O_3, and O_4; 5) an information processing device for the external memory — the queue discipline units δ_1, δ_2, δ_3, and δ_4; 6) an internal memory — the server (OY); 7) an information processing device for internal memory — the graph of server state transitions; 8) output poles — the output flows Π_1^{out}, Π_2^{out}, Π_3^{out}, and Π_4^{out}. The coordinate of a block is its number on the scheme.

Let us introduce the following variables and elements along with their value ranges. To fix a discrete time scale consider the epochs $\tau_0 = 0$, τ_1, τ_2, ... when the server changes its state. Let $\Gamma_i \in \Gamma$ be the server state during the interval $(\tau_{i-1}; \tau_i]$, $\varkappa_{j,i} \in \mathbb{Z}_+$ be the number of customers in the queue O_j at the instant τ_i, $\eta_{j,i} \in \mathbb{Z}_+$ be the number

of customers arrived into the queue O_j from the flow Π_j during the interval $(\tau_i;\tau_{i+1}]$, $\xi_{j,i} \in \mathbb{Z}_+$ be the number of customers in the saturation flow Π_j^{sat} during the interval $(\tau_i;\tau_{i+1}]$, $\overline{\xi}_{j,i} \in \mathbb{Z}_+$ be the actual number of serviced customers from the queue O_j during the interval $(\tau_i;\tau_{i+1}]$, $j \in \{1,2,3,4\}$.

The server changes its state according to the following rule:

$$\Gamma_{i+1} = h(\Gamma_i, \varkappa_{3,i}) \tag{3}$$

where the mapping $h(\cdot,\cdot)$ is defined by Formula (2). To determine the duration T_{i+1} of the next time slot it useful to introduce a mapping $h_T(\cdot,\cdot)$ by

$$T_{i+1} = h_T(\Gamma_i, \varkappa_{3,i}) = T^{(k,r)} \quad \text{where } \Gamma^{(k,r)} = \Gamma_{i+1} = h(\Gamma_i, \varkappa_{3,i}).$$

A functional relation

$$\overline{\xi}_{j,i} = \min\{\varkappa_{j,i} + \eta_{j,i}, \xi_{j,i}\}, \quad j \in \{1,2,3\}, \tag{4}$$

between $\overline{\xi}_{j,i}$ and $\varkappa_{j,i}$, $\eta_{j,i}$, $\xi_{j,i}$ describes the service strategy. Further, since

$$\varkappa_{j,i+1} = \varkappa_{j,i} + \eta_{j,i} - \overline{\xi}_{j,i}, \quad j \in \{1,2,3\},$$

and due to (4) it follows that

$$\varkappa_{j,i+1} = \max\{0, \varkappa_{j,i} + \eta_{j,i} - \xi_{j,i}\}, \quad j \in \{1,2,3\}. \tag{5}$$

We also have from the problem settings the following relations for the flow Π_4:

$$\eta_{4,i} = \min\{\xi_{1,i}, \varkappa_{1,i} + \eta_{1,i}\}, \quad \varkappa_{4,i+1} = \varkappa_{4,i} + \eta_{4,i} - \eta_{2,i}, \quad \xi_{4,i} = \varkappa_{4,i}. \tag{6}$$

Put $\varkappa_i = (\varkappa_{1,i}, \varkappa_{2,i}, \varkappa_{3,i}, \varkappa_{4,i})$. The non-local description of the input and saturation flows consists in specifying particular features of the conditional probability distribution of selected discrete components $\eta_i = (\eta_{1,i}, \eta_{2,i}, \eta_{3,i}, \eta_{4,i})$ and $\xi_i = (\xi_{1,i}, \xi_{2,i}, \xi_{3,i}, \xi_{4,i})$ of marked point processes $\{(\tau_i, \nu_i, \eta_i); i \geqslant 0\}$ and $\{(\tau_i, \nu_i, \xi_i); i \geqslant 0\}$ with marks $\nu_i = (\Gamma_i; \varkappa_i)$. Let $\varphi_1(\cdot,\cdot)$ and $\varphi_3(\cdot,\cdot)$ be defined by series expansions

$$\sum_{\nu=0}^{\infty} z^{\nu} \varphi_j(\nu, t) = \exp\{\lambda_j t (f_j(z) - 1)\}$$

with functions where $f_j(z)$ defined by (1), $j \in \{1,3\}$. The function $\varphi_j(\nu, t)$ equals the probability of $\nu = 0, 1, \ldots$ arrivals in the flow Π_j during time $t \geqslant 0$. If $\nu < 0$ the value of $\varphi_j(\nu, t)$ is set to zero. Define function $\psi(\cdot,\cdot,\cdot)$ by

$$\psi(k; y, u) = C_y^k u^k (1-u)^{y-k}.$$

Then $\psi(k; y, p_{k,r})$ is the probability of k arrival from flow Π_2 given the queue O_4 contains y customers and the server state is $\Gamma^{(k,r)}$. For values $k \notin \{0, 1, \ldots, y\}$ the value of $\psi(k; y, u)$ is set to zero.

Let $a = (a_1, a_2, a_3, a_4) \in \mathbb{Z}_+^4$ and $x = (x_1, x_2, x_3, x_4) \in \mathbb{Z}_+^4$. If the mark value is $\nu_i = (\Gamma^{(k,r)}; x)$ then the probability $\varphi(a, k, r, x)$ of simultaneous equalities $\eta_{1,i} = a_1$, $\eta_{2,i} = a_2$, $\eta_{3,i} = a_3$, $\eta_{4,i} = a_4$ according the the problem statement is

$$\varphi_1(a_1, h_T(\Gamma^{(k,r)}, x_3)) \cdot \psi(a_2, x_4, p_{\tilde{k},\tilde{r}}) \cdot \varphi_3(a_3, h_T(\Gamma^{(k,r)}, x_3)) \cdot \delta_{a_4, \min\{\ell(\tilde{k},\tilde{r},1), x_1 + a_1\}}$$

where $\Gamma^{(\tilde{k},\tilde{r})} = h(\Gamma^{(k,r)}, x_3)$ and $\delta_{i,j}$ is the Kroneker's delta:

$$\delta_{i,j} = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{if } i \neq j. \end{cases}$$

Let $b = (b_1, b_2, b_3, b_4) \in \mathbb{Z}_+^4$. The probability $\zeta(b, k, r, x)$ of simultaneous equalities $\xi_{1,i} = b_1$, $\xi_{2,i} = b_2$, $\xi_{3,i} = b_3$, $\xi_{4,i} = b_4$ given the fixed label value $\nu_i = (\Gamma^{(k,r)}; x)$ is

$$\delta_{b_1, \ell(\tilde{k},\tilde{r},1)} \cdot \delta_{b_2, \ell(\tilde{k},\tilde{r},2)} \cdot \delta_{b_3, \ell(\tilde{k},\tilde{r},3)} \cdot \delta_{b_4, x_4}.$$

The assumptions on statistical properties of some blocks and function relations between blocks are not contradicting and sufficient to construct a formal probability model, as the following theorem first proven in [9] demonstrates.

Theorem 1. *Choose* $\gamma_0 = \Gamma^{(k_0,r_0)} \in \Gamma$ *and* $x^0 = (x_{1,0}, x_{2,0}, x_{3,0}, x_{4,0}) \in \mathbb{Z}_+^4$. *There exists a probability space* $(\Omega, \mathcal{F}, \mathbf{P}(\cdot))$, *random variables* $\eta_{j,i} = \eta_{j,i}(\omega)$, $\xi_{j,i} = \xi_{j,i}(\omega)$, $\varkappa_{j,i} = \varkappa_{j,i}(\omega)$, *and random elements* $\Gamma_i = \Gamma_i(\omega)$, $i \geqslant 0$, $j \in \{1, 2, 3, 4\}$ *defined on this space, such that: 1) equalities* $\Gamma_0(\omega) = \gamma_0$ *and* $\varkappa_0(\omega) = x^0$ *hold; 2) relations* (3), (5), (6) *hold; 3) for any a, b,* $x^t = (x_{1,t}, x_{2,t}, x_{3,t}, x_{4,t}) \in \mathbb{Z}_+^4$, $\Gamma^{(k_t,r_t)} \in \Gamma$, $t = 1, 2, \ldots$ *the joint conditional probability distribution of vectors* η_i *and* ξ_i *has the form*

$$\mathbf{P}\left(\left\{\omega\colon \eta_i(\omega) = a, \xi_i(\omega) = b\right\} \middle| \bigcap_{t=0}^{i} \left\{\omega\colon \Gamma_t(\omega) = \Gamma^{(k_t,r_t)}, \varkappa_t(\omega) = x^t\right\}\right) = \varphi(a, k_i, r_i, x^i) \cdot \zeta(b, k_i, r_i, x^i).$$

From now on we focuse on low-priority customers in the queue O_3.

4. The low-priority queue

Here we will consider the stochastic sequence

$$\{(\Gamma_i(\omega), \varkappa_{3,i}(\omega)); i = 0, 1, \ldots\} \tag{7}$$

which includes the number of low-priority customers $\varkappa_{3,i}(\omega)$ in the queue O_3. In this section we will report several results concerning this stochastic sequence.

Theorem 2. *Let* $\Gamma_0(\omega) = \Gamma^{(k,r)} \in \Gamma$ *and* $\varkappa_{3,0}(\omega) = x_{3,0} \in \mathbb{Z}_+$ *be fixed. Then the stochastic sequence* (7) *is a homogeneous denumerable Markov chain.*

Theorem 3. *Let $x_3,\ \tilde{x}_3 \in \mathbb{Z}_+$ and $\Gamma^{(k,r)},\ \Gamma^{(\tilde{k},\tilde{r})} = h(\Gamma^{(k,r)}, x_3) \in \Gamma$. Then the transition probabilities of the Markov chain* (7) *are*

$$\begin{aligned}\mathbf{P}(\{\omega\colon \Gamma_{i+1}(\omega) = \Gamma^{(\tilde{k},\tilde{r})}, \varkappa_{3,i+1}(\omega) = \tilde{x}\} \mid \{\omega\colon \Gamma_i(\omega) = \Gamma^{(k,r)}, \varkappa_{3,i}(\omega) = x\}) \\ = (1 - \delta_{\tilde{x}_3,0}) \cdot \varphi\Big(\tilde{x}_3 + \ell(\tilde{k},\tilde{r},3) - x_3, h_T(\Gamma^{(\tilde{k},\tilde{r})}, x_3)\Big) \\ + \delta_{\tilde{x}_3,0} \sum_{a=0}^{\ell(\tilde{k},\tilde{r},3)-x_3} \varphi_3\Big(a, h_T(\Gamma^{(\tilde{k},\tilde{r})}, x_3)\Big).\end{aligned}$$

The last theorem clarifies which states of the Markov chain $\{(\Gamma_i, \varkappa_{3,i});\ i \geqslant 0\}$ are essential. To make a complete classification we introduce sets

$$S^3_{0,r} = \left\{(\Gamma^{(0,r)}, x_3)\colon\ x_3 \in Z_+,\ L \geqslant x_3 > L - \max_{k=1,2,\ldots,d}\left\{\sum_{t=0}^{n_k} \ell(k,t,3)\right\}\right\},\quad 1 \leqslant r \leqslant n_0,$$

$$S^3_{k,r} = \left\{(\Gamma^{(k,r)}, x_3)\colon\ x_3 \in Z_+,\ x_3 > L - \sum_{t=0}^{r-1} \ell(k,t,3)\right\},\quad 1 \leqslant k \leqslant d,\quad 1 \leqslant r \leqslant n_k.$$

Theorem 4. *The set of essential states of the Markov chain $\{(\Gamma_i, \varkappa_{3,i});\ i \geqslant 0\}$ consists of sets $\bigcup\limits_{1\leqslant r\leqslant n_0} S^3_{0,r}$ and $\bigcup\limits_{\substack{1\leqslant k\leqslant d\\ 1\leqslant r\leqslant n_k}} S^3_{k,r}$.*

As before, let $\Gamma^{(k,r)} \in \Gamma$ and $x_3 \in Z_+$. Denote by $H_{-1}(\Gamma^{(k,r)}, x_3)$ the set of all server states γ such that $h(\gamma, x_3) = \Gamma^{(k,r)}$ and put $r \ominus_k 1 = r - 1$ for $n_k \geqslant r > 0$, and $r \ominus_k 1 = n_k$ for $r = 0$ ($k = 0, 1, \ldots, d$). Then formula (2) makes it possible to define the mapping $H_{-1}(\Gamma^{(k,r)}, x_3)$ explicitly:

$$H_{-1}(\Gamma^{(k,r)}, x_3) = \begin{cases}\{\Gamma^{(k_1,r_1)}, \Gamma^{(0,r\ominus_0 1)}\} & \text{if } (k = 0) \wedge (x_3 \leqslant L),\\ \{\Gamma^{(k,r\ominus_k 1)}, \Gamma^{(0,r_2)}\} & \text{if } (\Gamma^{(k,r)} \in C^{\mathrm{I}}_k) \wedge (x_3 > L),\\ \{\Gamma^{(k,r\ominus_k 1)}\} & \text{if } (\Gamma^{(k,r)} \in C^{\mathrm{O}}_k) \vee (\Gamma^{(k,r)} \in C^{\mathrm{N}}_k);\\ \varnothing & \text{if } (k = 0) \wedge (x_3 > L)\\ & \quad \text{or } (\Gamma^{(k,r)} \in C^{\mathrm{I}}_k) \wedge (x_3 \leqslant L)\end{cases} \tag{8}$$

where $h_1(\Gamma^{(k_1,r_1)}) = r$ and $h_3(r_2) = \Gamma^{(k,r)}$.

Let's define for $\gamma \in \Gamma$ and $x_3 \in Z_+$ values

$$Q_{3,i}(\gamma, x) = \mathbf{P}(\{\omega\colon \Gamma_i(\omega) = \gamma, \varkappa_{3,i}(\omega) = x_3\}).$$

Suppose k and r are such that $\Gamma^{(k,r)} \in \Gamma$. Let's define the partial probability generating functions

$$\mathfrak{M}^{(i)}(k,r,v) = \sum_{w=0}^{\infty} Q_{3,i}(\Gamma^{(k,r)}, w)v^w, \qquad \Phi^{(i)}(k,r,v) = \sum_{x_3=0}^{\infty} \sum_{\gamma\in H_{-1}(\Gamma^{(k,r)},x_3)} Q_{3,i}(\gamma, x_3)v^{x_3},$$

$$q_{k,r}(v) = v^{-\ell(k,r,3)} \sum_{w=0}^{\infty} \varphi_3(w, T^{(k,r)})v^w.$$

Theorem 5. *Let* $\tilde{\gamma} = \Gamma^{(\tilde{k},\tilde{r})} \in \Gamma$. *The following recurrent w.r.t.* $i \geqslant 0$ *relations take place for the partial probability generating functions:*

$$\mathfrak{M}^{(i+1)}(\tilde{k},\tilde{r},v) = q_{\tilde{k},\tilde{r}}(v)\Phi^{(i)}(\tilde{k},\tilde{r},v) + \sum_{x_3=0}^{\ell(\tilde{k},\tilde{r},3)} \sum_{\gamma\in H_{-1}(\tilde{\gamma},x_3)} Q_{3,i}(\gamma,x_3) \sum_{a=0}^{\ell(\tilde{k},\tilde{r},3)-x_3} \varphi_3(a,T^{(\tilde{k},\tilde{r})}) -$$

$$- \sum_{x_3=0}^{\ell(\tilde{k},\tilde{r},3)} \sum_{\gamma\in H_{-1}(\tilde{\gamma},x_3)} Q_{3,i}(\gamma,x_3)v^{x_3-\ell(\tilde{k},\tilde{r},3)} \sum_{w=0}^{\ell(\tilde{k},\tilde{r},3)+1-x_3} \varphi_3(w,T^{(\tilde{k},\tilde{r})})v^w.$$

5. Acknowledgments

This work was fulfilled as a part of State Budget Research and Development program No. 01201456585 "Mathematical modeling and analysis of stochastic evolutionary systems and decision processes" of N.I. Lobachevsky State University of Nizhni Novgorod and was supported by State Program "Promoting the competitiveness among world's leading research and educational centers"

REFERENCES

1. Neimark Yu. I., Fedotkin M. A., Preobrazhenskaja A. M. Operation of an automate with feedback controlling traffic at an intersection // Izvestija of USSR Academy of Sciences, Technical Cybernetic. 1968. No. 5. P. 129–141.
2. Fedotkin M. A. On a class of stable algorithms for control of conflicting flows or arriving airplanes // Problems of control and information theory. 1977. V. 6, No. 1. P. 13–22.
3. Fedotkin M. A. Construction of a model and investigation of nonlinear algorithms for control of intense conflict flows in a system with variable structure of servicing demands. I // Lithuanian mathematical journal. 1977. V. 17, No. 1. P. 129–137.
4. Litvak N. V., Fedotkin M. A. A probabilistic model for the adaptive control of conflict flows // Automation and Remote Control. 2000. V. 61, No. 5. P. 777–784.
5. Proidakova E.V., Fedotkin M.A. Control of output flows in the system with cyclic servicing and readjustments // Automation and remote control. 2008. V. 69, No. 6. P. 993–1002.
6. Afanasyeva L. G., Bulinskaya E. V. Mathematical models of transport systems based on queueing theory // Trudy of Moscow Institute of Physocs and Technology. 2010. No. 4. P.6–21.
7. Yamada K., Lam T. N. Simulation analysis of two adjacent traffic signals // Proceedings of the 17th winter simulation conference. ACM, New York. 1985. P. 454–464.
8. Zorin A.V. Stability of a tandem of queueing systems with Bernoulli noninstantaneous transfer of customers // Theory of Probability and Mathematical Statistics. 2012. V. 84. P. 173-188.
9. Kocheganov V. M., Zorine A.V. Probabilistic model of tandem of queuing systems under cyclic control with prolongations // Proceedings of Internation conference "Probability theory, stochastic processes, mathematical statistics and applications" (Minsk, Feb. 23–26 2015). 2015. P. 94-99.

SELECTING MULTILEVEL STRUCTURE SECURE ACCESS TO RESOURCES EXTERNAL NETWORK

V. Kolomoitcev[1], *V. Bogatyrev*[2]
[1,2] ITMO University, Saint-Petersburg, Russia
[1]dek-s-kornis@yandex.ru, [2]vladimir.bogatyrev@gmail.com

Abstract

The paper presents the evaluation of the effectiveness of the structural organization of the system of multi-level secure access to external network resources. We conducted a comparative analysis and optimization of the access scheme "Direct connection", with of its various forms of implementation during the organization of a secure connection of end node to the internal network resources located in the external network.

Keywords: information protection, unauthorized access, firewalls, networking, fault tolerance, information security, reliability.

1. Introduction

Modern computer networks, both corporate and public, have a complicated structure. In such networks, there are some very serious problems of information security. They may be at risk of unauthorized access, denial-of-service nodes, the loss of transmitted information, as well as threats of violations of privacy that could lead to significant economic and other losses. [1, 2]. Threats can be both external - as a result of remote network attacks, and internal - by various stowing software or hardware. To eliminate the challenges of information security, some measures can be taken and means of information security, located on various levels of the network used. The principles of organization of a secure connection of the corporate network to public network are among the most important elements for ensuring information security. They have a significant impact on the safety and reliability of the network. However, it is worth remembering that the most effective security techniques usually imply some significant costs. In this study, we investigate possibilities of the scheme for the organization secure access to external network resources, taking into account the requirements set out in the guidance documents (legislative and legal documents) on information security. The study is aimed at a choice of rational options for creating protection system, with ensuring minimizing the average residence time in her requests and maintaining its high reliability [3, 4].

2. Object and objectives of the study

The scheme which is regarded in this paper is focused on improving the level of protection devices on the network. The key challenge of the scheme is to organize

secure access to poorly protected and / or uncontrolled portions of the network. This scheme allows reducing the threat of DDoS-attacks, unauthorized access to a network node, listening to the information channel and penetration of malicious software [5]. The scheme under consideration is based on a standard network access scheme to the resources of the external network: the node 'Internal (local) network' - Routers - 'External network (the Internet)'. This approach minimizes the degree of possible reorganization of existing corporate network. This standard scheme is depicted in Figure 1. In the standard scheme of access the end-node of the corporate network to nodes of the external network the protection of this node is based on a built-in means: antivirus protection (AV), a standard firewall (FW), and possible means of protection against unauthorized access. In the standard scheme at the entrance to the network have a router. The measures used in this scheme, leading to the fact that almost all of the work to eliminate threats from the external network rests on the end-node. For mission-critical systems of mentioned above means of protection are not enough. Therefore, we should use the scheme that ensures a comprehensive information security. In the role of such a scheme can be used the scheme 'Direct connection' [5].

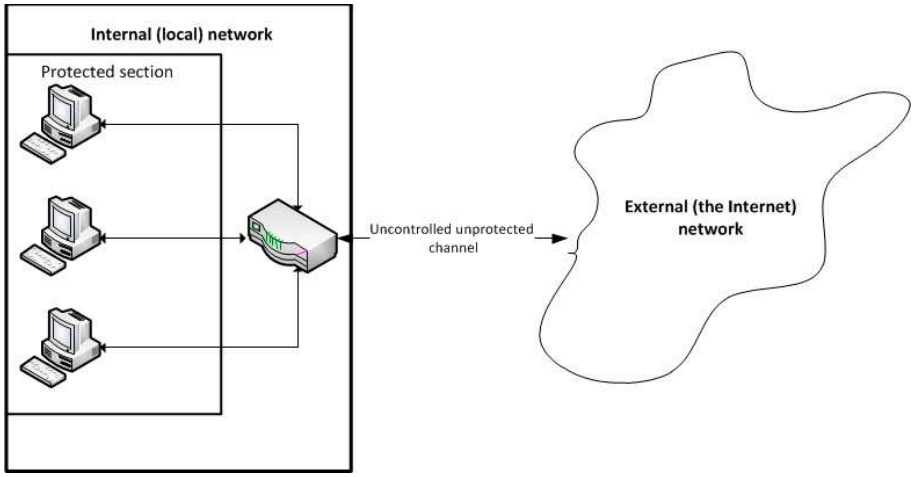


Figure 1: The standard scheme of access node in the external network.

As a result, we are suggested to consider the possibilities of the scheme "Direct connection", in its various physical interpretations, in terms of its reliability and the minimum average residence time of the request in the system. The aim is investigating the possibilities of access scheme 'Direct connection', in its various physical interpretations in terms of ensuring high reliability of the system while minimizing the delays to the service flow.

3. The basic version of the access scheme 'Direct connection'

The use scheme 'Direct connection' involves minimal changes in the architecture of the corporate network, as well as minimal additional financial costs to implement it. The structure of the scheme of 'Direct connection' is presented in Figure 2.

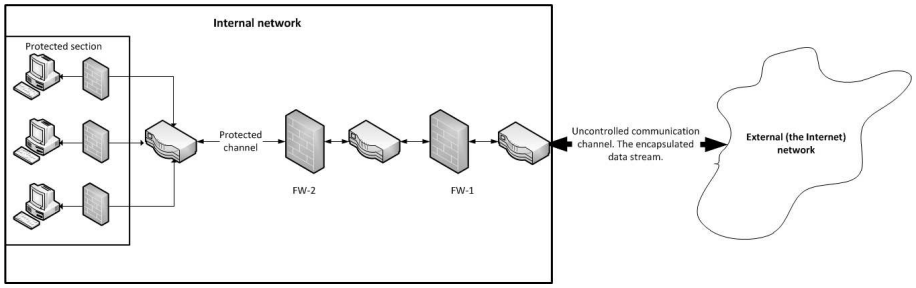


Figure 2: The scheme 'Direct connection'.

In this scheme, at the entrance to the internal network (just after the router) is set firewall with packet-filtering (FW-1) to eliminate spam, reduce the total load of the channel within the network, as well as reduce the risk of DDoS-attacks. In most cases mentioned above router can carry functional firewall with packet-filtering, however, would be more effective to use a separate router and firewall. Firewall with adaptive detailed packet inspection (FW-2) installed behind a FW-1 for a deeper analysis of the contents of packages. [1]. Given the fact that the input of the FW-2 will flow less data than the input of the FW-1, the load on the FW-2 will be smaller and, as a result, higher performance of the network itself. Once the data have passed the FW-2, they are (potentially "clean-data") must be received to the desired end-node. On end-node there are local antivirus (AV) (with personal firewall) installed [1], as well as some systems of protection against unauthorized access (UAA), and some secure data storage in order to reduce the negative effects of potential insider attacks. In this scheme of access, the channel data is to be protected, thereby reducing or even prevents the possibility of influence an intruder on data flowing in the channel. The choice of encryption algorithms and means (for the implementation of various functions in the scheme) carried out according to existing guidelines. To improve the overall network protection from DDoS-attacks, data loss or destruction and other threats, mission critical nodes (both in terms of network architecture, and in terms of data stored on them) should be reserved, and for the data stored on them, backups are created.

4. Ways of construction the network infrastructure of access scheme 'Direct connection'

For qualitative and uninterrupted operation of the network, you must do a backup of system components. Network architecture of the scheme 'Direct connection' has three main components: firewall with packet-filtering, firewall with adaptive detailed

packet inspection and routers that connect all the elements of scheme together. In this scheme there are four possible ways for the construction of scheme: with three, two or one groups of routers on the entire system. Possible ways of constructing a network infrastructure scheme ”Direct connection” presented in Figure 3.

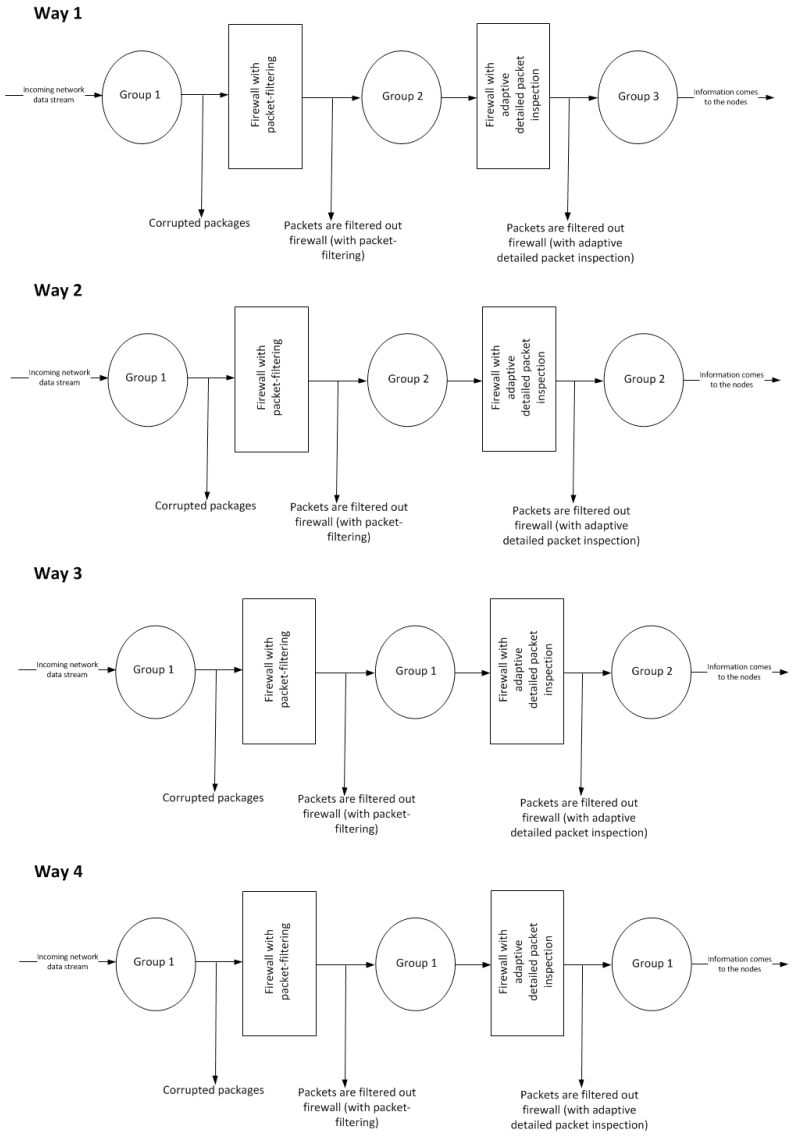


Figure 3: Ways network architecture scheme ’Direct connection’.

5. Estimation of reliability and average residence time of request in the system

Optimization of possible ways scheme "Direct connection" requires a search for multiplicity of redundant nodes in each group. Required to reach the highest possible level of system reliability while maintaining a minimum average residence time of request in the system (hereinafter, ARToRS), given imposed limits on the cost of implementing the system [6]. In this study, we assume that each of the FW and routers eliminates and finds only its share of threats (errors) in the incoming stream. Each node in the network represents the queuing system type M/M/1 with infinite queue. The average residence time of request in such a system is defined as [7]:

$$T = \frac{1/\mu}{1-\rho} = \frac{\nu}{1-\lambda/\mu} = \frac{\nu}{1-\lambda/\nu} \tag{1}$$

Here μ - service rate, and $\rho = \lambda/\mu$ - channel usage, $\nu = 1/\mu$ - average service time a request in a node, λ - arrival rate of requests (the density of the flow). The arrival rate of requests received at each node is divided by N, as it enters for maintenance in the N-nodes. When requests pass through several nodes, the ARToRS is defined as the sum of the residence times in the nodes that are consistently involved in its maintenance. Thus, for the system consisting of set of nodes the total the ARToRS is defined as:

$$T_{total} = \sum_i T_i \tag{2}$$

After passing through the router the input flow is filtered and, thus, the density of the flow on the FW-1 is lower than the router. The same happens with the input flow received at the FW-2. After passing through the FW-1 a certain proportion of the input flow is filtered and to the FW-2 is received smaller input flow. Thus, all four variants of the scheme "Direct connection" the ARToRS is defined as:

$$T_1(\lambda) = \frac{\nu_0}{B_1} + \frac{\nu_1}{B_2} + \frac{\nu_0}{B_3} + \frac{\nu_2}{B_4} + \frac{\nu_0}{B_5} \tag{3}$$

$$T_2(\lambda) = \frac{\nu_0}{B_1} + \frac{\nu_1}{B_2} + \frac{\nu_0}{B_3} + \frac{\nu_2}{B_4} + \frac{\nu_0}{1-\alpha_{t2}\cdot\lambda\cdot\frac{\nu_0}{n_{02}}} \tag{4}$$

$$T_3(\lambda) = \frac{\nu_0}{B_1} + \frac{\nu_1}{B_2} + \frac{\nu_0}{1-\alpha_{t1}\cdot\lambda\cdot\frac{\nu_0}{n_{01}}} + \frac{\nu_2}{B_4} + \frac{\nu_0}{B_5} \tag{5}$$

$$T_4(\lambda) = \frac{\nu_0}{B_1} + \frac{\nu_1}{B_2} + \frac{\nu_0}{1-\alpha_{t1}\cdot\lambda\cdot\frac{\nu_0}{n_{01}}} + \frac{\nu_2}{B_4} + \frac{\nu_0}{1-\alpha_{t2}\cdot\lambda\cdot\frac{\nu_0}{n_{01}}} \tag{6}$$

Here $B_1 = 1-\lambda\cdot\frac{\nu_0}{n_{01}}$; $B_2 = 1-(1-A_0\cdot p_0)\cdot\lambda\cdot\frac{\nu_1}{n_1}$; $B_3 = 1-\alpha_{t1}\cdot\lambda\cdot\frac{\nu_0}{n_{02}}$; $B_4 = 1-\alpha_{t1}\cdot\lambda\cdot\frac{\nu_2}{n_2}$; $B_5 = 1-\alpha_{t2}\cdot\lambda\cdot\frac{\nu_0}{n_{03}}$.

Moreover, $\alpha_{t1} = (1-A_0\cdot p_0)\cdot(1-A_1\cdot p_1)$ and $\alpha_{t2} = \alpha_{t1}\cdot(1-A_2\cdot p_2)$, where $(1-A_i\cdot p_i)$ - the proportion of the filtered input flow of previously placed node. At the same time ν_0, ν_1, ν_2 - average service time of request in routers, FW-1 and FW-2; λ -

the arrival rate of requests; A_0, A_1, A_2 - respectively, the proportion of threats (errors) in the input stream, the router detected with a probability P_0, FW-1 with a probability P_1, FW-2 with a probability P_2; n_{0i} - the number of routers in the i-th group; n_1 - the number of FW-1; n_2 - the number of FW-2; Costs for the implementation of the ways of construction of the scheme, are defined as:

$$C_{1-4} = c_0 \cdot \sum_i n_{0i} + c_1 \cdot n_1 + c_2 \cdot n_2 \tag{7}$$

Here c_0, c_1, c_2 - the cost of router, FW-1, FW-2. In turn, reliability of the proposed schemes is equal:

$$P_1 = P_{01} \cdot Pm1 \cdot P_{02} \cdot P_{m2} \cdot P_{03} \tag{8}$$

$$P_{2-3} = P_{01} \cdot P_{m1} \cdot P_{02} \cdot P_{m2} \tag{9}$$

$$P_1 = P_{01} \cdot Pm1 \cdot P_{m2} \tag{10}$$

Here $P_{m1} = (1 - (1 - r_1)^{n_1})$, $P_{m2} = (1 - (1 - r_2)^{n_2})$.Assuming that the routers in each group are the same: $P_{0i} = (1 - (1 - r_0)^{n_{0i}})$. Here $r_j = e^{-\lambda_j t}$ and λ_0, λ_1, λ_2 - failure rate of routers, FW-1 and FW-2; n_{0i} - the number of routers in the i-th group; n_1 - the number of FW-1; n_2 - the number of FW-2.

Optimization of protection systems includes finding the distribution of each type of node that provides maximum reliability of the entire system considering the limitation of the cost of implementation: $C_1 \leq C$, $C_2 \leq C$,...,$C_4 \leq C$; and compliance steady state conditions of service [8, 9, 10].

Results of reliability calculation , depending on the constraints imposed on the system throughput determined by a known the arrival rate - λ, when $r_0 = 0.7$, $r_1 = 0.8$, $r_2 = 0.9$, are shown in Figure 4.

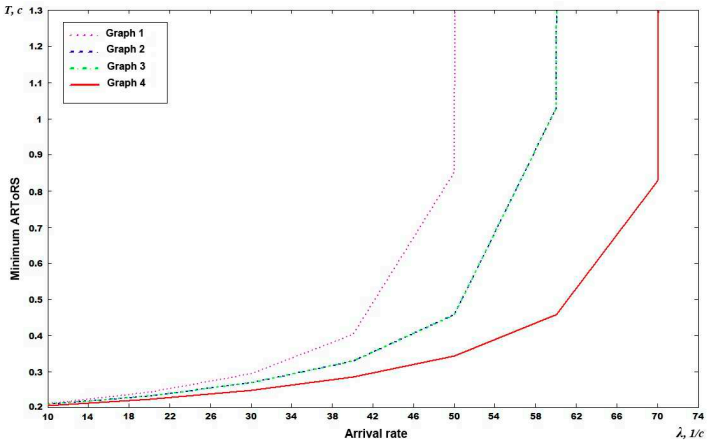


Figure 4: Reliability access schemes based on the arrival rate of requests.

The calculation results are the average residence time in the system of requests according to the formulas (3) - (6) for each type of node defined by the search of the maximum reliability of the designed system, depending on the arrival rate λ, shown in Figure 5. The calculations are performed when $\nu_0 = 0.025$sec, $\nu_1 = 0.04$sec, $\nu_2 = 0.075$sec, $c_0 = 10$cu, $c_1 = 25$cu, $c_2 = 50$cu, $C = 550$cu, $p_0 = 0.95$, $p1 = p2 = 0.899$, $A_0 = 0.05$, $A_1 = 0.1$, $A_2 = 0.25$.

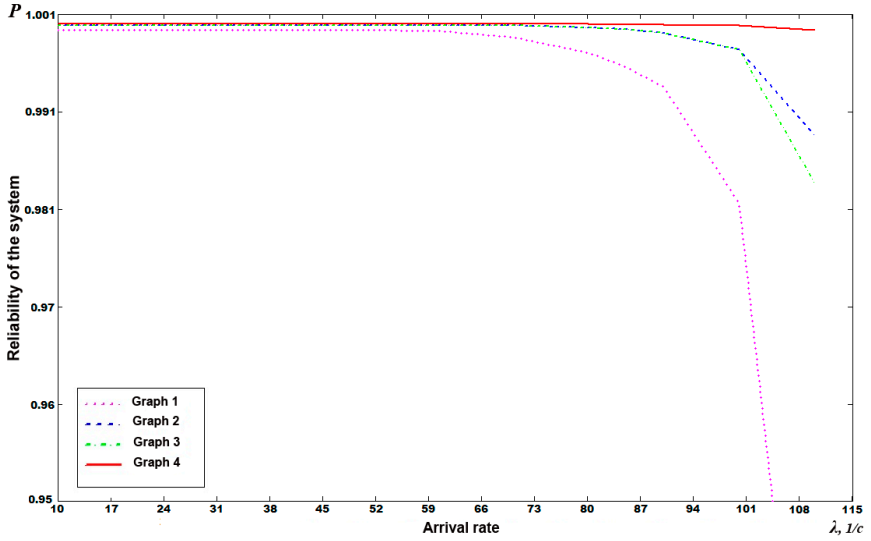


Figure 5: ARToRS at the highest possible reliability of the system.

Sequence numbers of curves in the graphs have a direct accordance with sequence numbers of their defining formulas, namely the graphs 1-4 - ways of constructing scheme 'Direct connection'.

As shown in Figure 4, for small values of the arrival rate, the reliability of each of the ways of the scheme 'Direct connection' is approximately equal. However, with increasing the arrival rate is detected, the fourth way of scheme is more reliable than other ways, and the first - the least reliable. Also from Figure 4 shows that the second and third ways have almost identical scheme reliability.

After analyzing Figure 5, we can say that the best result for the minimum ARToRS also has a fourth way of the scheme 'Direct connection', and the worst - the first way. As is the case with the indicator the system reliability, the Figure 5 shows that ways two and three also possess almost identical values.

As a result, we can conclude that if you want to use the most reliable way of the scheme and at the same time having the lowest value ARToRS, it is best to choose the fourth option. In the case where raises the question of maximizing the system reliability and to select one of two options - the second or third then there is no difference which of them to use.

6. Conclusion

The paper analyzes possibilities of access scheme 'Direct connection' that allow you to organize a secure connection the end-node internal network to resources located in the external network. The study identified the advantages and disadvantages of the scheme "Direct connection", depending on the way of its structure. It has been shown that one of the ways of the scheme 'Direct connection' (using a single group router to the entire access scheme), has a higher level of reliability than others. In addition to that, the same way of construction has the lowest value of the minimum ARToRS than other variants of the scheme. We also show that the second and third variants of schemes are almost identical to each other in terms of reliability and minimum ARToRS. Thus, when implementing the scheme 'Direct connection' is best to use a way of its structure using a common pool of routers for the entire system.

REFERENCES

1. Ingham Kenneth, Forrest Stephanie. A History and Survey of Network Firewalls // University of New Mexico. - 2002.
2. Алиев Т.И. Проектирование систем с приоритетами.// Известия высших учебных заведений. Приборостроение. 2014. Т. 57. № 4. С. 30-35
3. Богатырев В.А. и др Оптимизация вычислительных систем с объединением межсетевых экранов в отказоустойчивые кластеры // Научно-технический вестник информационных технологий, механики и оптики. - 2011. № 6 (76). С. 140-142.
4. Богатырев В.А. Оценка и выбор отказоустойчивых конфигураций межсетевых экранов / Богатырев В.А., Фокин С.Б., Попова М.В. // Научно-технический вестник информационных технологий, механики и оптики. 2011. № 3 (73). С. 139-140.
5. Коломойцев В.С. Сравнительный анализ подходов к организации безопасного подключения узлов корпоративной сети к сети общего доступа // Кибернетика и программирование. — 2015. - № 2. - С.46-58.
6. 6. 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. Vol. 49. No. 1. pp. 46-56. DOI 10.3103/S0146411615010022
7. Вишневский В.М. "Теоретические основы проектирования компьютерных сетей ".- М.: Техносфера, 2003. 512 с.
8. Bogatyrev V.A., Golubev I.Y., Bogatyrev S.V. Optimization and the Process of Task Distribution between Computer System Clusters//Automatic Control and Computer Sciences, IET - 2012, No. 3, pp. 103 – 11
9. Bogatyrev V.A.: Fault tolerance of clusters configurations with direct connection of storage devices // Automatic Control and Computer Sciences. 2011. V. 45. N 6. P. 330-337.
10. Bogatyrev V.A. Exchange of Duplicated Computing Complexes in Fault tolerant Systems // Automatic Control and Computer Sciences. - 2011. - V. 46. - No. 5. - pp. 268-276.

ASYMPTOTIC ANALYSIS OF RETRIAL QUEUE WITH TWO ORBITS UNDER LONG DELAY CONDITION

I. Kononov, E. Fedorova
Tomsk state university, Tomsk, Russia

Abstract

In the paper, the retrial queueing system M2/M2/1 with two orbits is studied by means of asymptotic analysis method under long delay of calls in both orbits. Joint probability distribution of number os calls in orbits is obtained.

АСИМПТОТИЧЕСКИЙ АНАЛИЗ RQ-СИСТЕМ С ДВУМЯ ИСТОЧНИКАМИ ПОВТОРНЫХ ВЫЗОВОВ В УСЛОВИИ БОЛЬШОЙ ЗАДЕРЖКИ

И. Кононов, Е. Федорова

Национальный исследовательский Томский государственный университет, Томск, Россия
develi@sibmail.com, moiskate@mail.ru

Аннотация

В работе исследована математическая модель RQ-системы M2/M2/1 с двумя источниками повторных вызовов методом асимптотического анализа в условии большой задержки заявок в обоих источниках. Найдено совместное распределение вероятностей числа заявок в первом и во втором источниках повторных вызовов.

Ключевые слова: RQ-система, метод асимптотического анализа, два источника повторных вызовов, большая задержка

1. Введение

Модели систем массового обслуживания являются важным инструментом при исследовании различных технических и экономических систем, в том числе информационно-коммуникационных сетей. В теории массового обслуживания обычно выделяют два класса моделей – системы с ожиданием и системы с потерями. Но начиная с 70-х годов стали возникать такие системы, которые требовали рассмотрения моделей, выходящих за рамки

множества классических систем массового обслуживания. В связи с этим возникла потребность исследовать новый класс систем, которые именуются RQ-системами (Retrial Queueing Systems) или системами с повторными вызовами.

RQ-системы особенны тем, что при обращении заявки к обслуживающему прибору в случае, когда прибор был занят, заявка не теряется, а уходит в источник повторных вызовов (ИПВ), откуда она повторно обращается к прибору после некоторой задержки. Системы с повторными вызовами широко применяются при моделировании телекоммуникационных систем, мобильных сотовых радиосетей, компьютерных сетей, call-центров и т.д. [1, 2, 3].

Наиболее широкое исследование систем с повторными вызовами приведено в работах J.R. Artalejo, A. Gomez-Corral [4], Г.И. Фалина [5], А.Н. Дудина [6] и др.

Изучением систем с двумя ИПВ и двумя входящими потоками занимались W.Yang [9], N.Rengnanathan, R. Kalayanaraman , B. Srinivasan [10], K. Avrachenkov, P. Nain, U. Yechiali [11]. В данной статье предлагается использовать метод асимптотического анализа [7, 8] для исследования RQ-систем с двумя ИПВ, суть которого отражена в работе на примере исследования системы $M_2|M_2|1$ с двумя ИПВ.

2. Математическая модель

Рассмотрим RQ-систему (1) с двумя источниками повторных вызовов, на вход которой поступают два простейших потока заявок с интенсивностями λ_1 и λ_2. Если поступившая заявка застает прибор свободным, то она занимает его для обслуживания, время обслуживания каждой заявки распределено по экспоненциальному закону с параметрами μ_1 и μ_2. Если прибор занят, то заявка 1-го типа переходит в первый источник повторных вызовов, а заявка 2-го типа – во второй, где они осуществляют случайную задержку, продолжительность которой имеет экспоненциальное распределение с параметрами σ_1 и σ_1 соответственно. Из ИПВ (это может быть как первый, так и второй) после случайной задержки заявка вновь обращается к прибору. Если прибор свободен, то заявка занимает его для обслуживания, если же он занят, то заявка мгновенно возвращается в свой ИПВ для реализации следующей задержки.

Обозначим $i_1(t)$ – число заявок в 1-м источнике повторных вызовов, а $i_2(t)$ – число заявок во 2-м источнике. Случайный процесс $k(t)$ описывает состояние прибора следующим образом:

$$k(t) = \begin{cases} 0, \text{если прибор свободен,} \\ 1, \text{если на приборе находитчя заявка 1-го типа,} \\ 2, \text{если на приборе находитчя заявка 2-го типа.} \end{cases}$$

Очевидно, что трехмерный процесс $\{k(t), i_1(t), i_2(t)\}$ является марковским. Обозначим $P\{k(t) = k, i_1(t) = i_1, i_2(t) = i_2\} = P_k(i_1, i_2, t)$ – вероят-

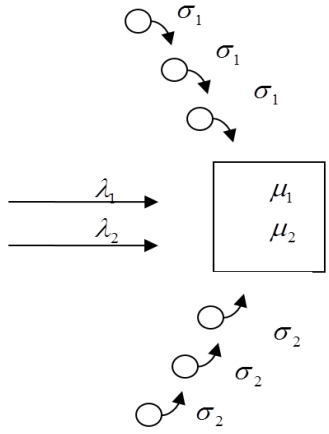


Рис. 1: RQ-система $M_2|M_2|1$

ность того, что прибор в момент времени t находится в состоянии k, в 1-м источнике повторных вызовов i_1 заявок и во 2-м источнике повторных вызовов i_2 заявок. Ставится задача найти совместное распределение вероятностей числа заявок в первом и во втором источниках повторных вызовов. Для распределения вероятностей $P_k(i_1, i_2, t)$ состояний рассматриваемой RQ-системы составим систему дифференциальных уравнений Колмогорова, которая в стационарном режиме примет вид:

$$\begin{cases} \mu_1 P_1(i_1,i_2) + \mu_2 P_2(i_1,i_2) - (\lambda_1+\lambda_2+i_1\sigma_1+i_2\sigma_2)P_0(i_1,i_2) = 0, \\ \lambda_1 P_0(i_1,i_2) + (i_1+1)\sigma_1 P_0(i_1+1,i_2) + \lambda_1 P_1(i_1-1,i_2) + \\ +\lambda_2 P_1(i_1,i_2-1) - (\lambda_1+\lambda_2+\mu_1)P_1(i_1,i_2) = 0, \\ \lambda_2 P_0(i_1,i_2) + (i_2+1)\sigma_2 P_0(i_1,i_2+1) + \lambda_1 P_2(i_1-1,i_2) + \\ +\lambda_2 P_2(i_1,i_2-1) - (\lambda_1+\lambda_2+\mu_2)P_2(i_1,i_2) = 0, \text{ for } i_1,i_2 \geq 0. \end{cases} \quad (1)$$

Перейдем в системе (1) к частичным характеристическим функциям:

$$H_k(u_1,u_2) = \sum_{i_1}\sum_{i_2} e^{ju_1i_1} e^{ju_2i_2} P_k(i_1,i_2), \quad k = \overline{0,2}.$$

где $j = \sqrt{-1}$ – мнимая единица.

Получим следующую систему:

$$\begin{cases} \mu_1 H_1(u_1,u_2)+\mu_2 H_2(u_1,u_2)-(\lambda_1+\lambda_2)H_0(u_1,u_2)+ \\ +j\sigma_1\dfrac{\partial H_0(u_1,u_2)}{\partial u_1}+j\sigma_2\dfrac{\partial H_0(u_1,u_2)}{\partial u_2}=0, \\ \lambda_1 H_0(u_1,u_2)-j\sigma_1 e^{-ju_1}\dfrac{\partial H_0(u_1,u_2)}{\partial u_1}+\lambda_1 e^{ju_1}H_1(u_1,u_2)+ \\ +\lambda_2 e^{ju_2}H_1(u_1,u_2)-(\lambda_1+\lambda_2+\mu_1)H_1(u_1,u_2)=0, \\ \lambda_2 H_0(u_1,u_2)-j\sigma_2 e^{-ju_2}\frac{\partial H_0(u_1,u_2)}{\partial u_2}+\lambda_1 e^{ju_1}H_2(u_1,u_2)+ \\ +\lambda_2 e^{ju_2}H_2(u_1,u_2)-(\lambda_1+\lambda_2+\mu_2)H_2(u_1,u_2)=0. \end{cases} \tag{2}$$

Аналитически данную систему решить не представляется возможным. Будем решать полученную систему (2) методом асимптотического анализа в условии большой задержки заявок в обоих ИПВ, то есть при $\sigma_1,\sigma_2\to 0$.

3. Метод асимптотического анализа в условии большой задержки

3.1. Асимптотика первого порядка. Введем следующие обозначения:

$$\sigma_k=\sigma\gamma_k,\quad \sigma=\varepsilon,\quad u_k=\varepsilon w_k,\quad H_k(u_1,u_2)=F_k(w_1,w_2,\varepsilon). \tag{3}$$

Тогда система (2) примет вид:

$$\begin{cases} \mu_1 F_1(w_1,w_2,\varepsilon)+\mu_2 F_2(w_1,w_2,\varepsilon)-(\lambda_1+\lambda_2)F_0(w_1,w_2,\varepsilon)+ \\ +j\varepsilon\gamma_1\dfrac{\partial F_0(w_1,w_2,\varepsilon)}{\partial(\varepsilon w_1)}+j\varepsilon\gamma_2\dfrac{\partial F_0(w_1,w_2,\varepsilon)}{\partial\varepsilon w_2}=0, \\ \lambda_1 F_0(w_1,w_2,\varepsilon)-j\varepsilon\gamma_1 e^{-j\varepsilon w_1}\dfrac{\partial F_0(w_1,w_2,\varepsilon)}{\partial(\varepsilon w_1)}+\lambda_1 e^{j\varepsilon w_1}F_1(w_1,w_2,\varepsilon)+ \\ +\lambda_2 e^{j\varepsilon w_2}F_1(w_1,w_2,\varepsilon)-(\lambda_1+\lambda_2+\mu_1)F_1(w_1,w_2,\varepsilon)=0, \\ \lambda_2 F_0(w_1,w_2,\varepsilon)-j\varepsilon\gamma_2 e^{-j\varepsilon w_2}\dfrac{\partial F_0(w_1,w_2,\varepsilon)}{\partial(\varepsilon w_2)}+\lambda_1 e^{j\varepsilon w_1}F_2(w_1,w_2,\varepsilon)+ \\ +\lambda_2 e^{j\varepsilon w_2}F_2(w_1,w_2,\varepsilon)-(\lambda_1+\lambda_2+\mu_2)F_2(w_1,w_2,\varepsilon)=0. \end{cases} \tag{4}$$

Нетрудно показать, что справедлива следующая теорема.

Теорема 1. *Предельное значение* $F_k(w_1,w_2)=\lim\limits_{\varepsilon\to 0}F_k(w_1,w_2,\varepsilon)$ *решения системы* (4) *имеет вид*

$$F_k(w_1,w_2)=R_k e^{jw_1x_1+jw_2x_2},$$

где величины R_0,R_1,R_2,x_1,x_2 *определяются по формулам:*

$$R_1=\frac{\lambda_1}{\mu_1},\quad R_2=\frac{\lambda_2}{\mu_2},\quad R_0=1-\left[\frac{\lambda_1}{\mu_1}+\frac{\lambda_2}{\mu_2}\right], \tag{5}$$

$$x_1=\frac{R_1\lambda_1+R_2\lambda_1}{\gamma_1 R_0},\quad x_2=\frac{R_1\lambda_2+R_2\lambda_2}{\gamma_2 R_0}. \tag{6}$$

Тогда из теоремы 1 и замены (3) следует, что

$$H(u_1,u_2)\approx\sum_k F_k(w_1,w_2)=\exp\left\{j\frac{u_1}{\sigma}x_1+j\frac{u_2}{\sigma}x_2\right\}.$$

Полученное равенство будем называть асимптотикой первого порядка.

3.2. Асимптотика второго порядка. В системе (2) осуществим замены:

$$H_k(u_1,u_2)=H_k^{(2)}(u_1,u_2)\exp\left\{ju_1\frac{\lambda_1}{\sigma_1}x+ju_2\frac{\lambda_2}{\sigma_2}x\right\},\tag{7}$$

где $x=\dfrac{1-R_0}{R_0}$.

Тогда получим следующую систему уравнений:

$$\begin{cases}\mu_1H_1^{(2)}(u_1,u_2)+\mu_2H_2^{(2)}(u_1,u_2)-(\lambda_1+\lambda_2)H_0^{(2)}(u_1,u_2)+\\ +j\sigma_1\dfrac{\partial H_0^{(2)}(u_1,u_2)}{\partial u_1}+j\sigma_2\dfrac{\partial H_0^{(2)}(u_1,u_2)}{\partial u_2}-\\ -\lambda_1xH_0^{(2)}(u_1,u_2)-\lambda_2xH_0^{(2)}(u_1,u_2)=0,\\ \lambda_1H_0^{(2)}(u_1,u_2)-j\sigma_1e^{-ju_1}\dfrac{\partial H_0^{(2)}(u_1,u_2)}{\partial u_1}+(\lambda_1e^{ju_1}+\lambda_2e^{ju_2}-\\ -\lambda_1-\lambda_2-\mu_1)H_1^{(2)}(u_1,u_2)+\lambda_1xe^{-ju_1}H_0^{(2)}(u_1,u_2)=0,\\ \lambda_2H_0^{(2)}(u_1,u_2)-j\sigma_2e^{-ju_2}\dfrac{\partial H_0^{(2)}(u_1,u_2)}{\partial u_2}+(\lambda_1e^{ju_1}+\lambda_2e^{ju_2}-\\ -\lambda_1-\lambda_2-\mu_2)H_2^{(2)}(u_1,u_2)+\lambda_2xe^{-ju_2}H_0^{(2)}(u_1,u_2)=0.\end{cases}\tag{8}$$

Введем обозначения:

$$\sigma_k=\gamma_k\sigma,\quad \sigma=\varepsilon^2,\quad u_k=\varepsilon w_k,\quad H_k^{(2)}(u_1,u_2)=F_k^{(2)}(w_1,w_2,\varepsilon).\tag{9}$$

Тогда система (8) примет вид:

$$\begin{cases}\mu_1F_1^{(2)}(w_1,w_2,\varepsilon)+\mu_2F_2^{(2)}(w_1,w_2,\varepsilon)-(\lambda_1+\lambda_2)F_0^{(2)}(w_1,w_2,\varepsilon)+\\ +j\varepsilon\gamma_1\dfrac{\partial F_0^{(2)}(w_1,w_2,\varepsilon)}{\partial w_1}+j\varepsilon\gamma_2\dfrac{\partial F_0^{(2)}(w_1,w_2,\varepsilon)}{\partial w_2}-\\ -\lambda_1xF_0^{(2)}(w_1,w_2,\varepsilon)-\lambda_2xF_0^{(2)}(w_1,w_2,\varepsilon)=0,\\ \lambda_1F_0^{(2)}(w_1,w_2,\varepsilon)-j\varepsilon\gamma_1e^{-j\varepsilon w_1}\dfrac{\partial F_0^{(2)}(w_1,w_2,\varepsilon)}{\partial w_1}+\\ +(\lambda_1e^{j\varepsilon w_1}+\lambda_2e^{j\varepsilon w_2}-\lambda_1-\lambda_2-\mu_1)F_1^{(2)}(w_1,w_2,\varepsilon)+\\ +\lambda_1xe^{-j\varepsilon w_1}F_0^{(2)}(w_1,w_2,\varepsilon)=0,\\ \lambda_2F_0^{(2)}(w_1,w_2,\varepsilon)-j\varepsilon\gamma_2e^{-jw_2}\dfrac{\partial F_0^{(2)}(w_1,w_2,\varepsilon)}{\partial w_2}+\\ +(\lambda_1e^{j\varepsilon w_1}+\lambda_2e^{j\varepsilon w_2}-\lambda_1-\lambda_2-\mu_2)F_2^{(2)}(w_1,w_2,\varepsilon)+\\ +\lambda_2xe^{-j\varepsilon w_2}F_0^{(2)}(w_1,w_2,\varepsilon)=0.\end{cases}\tag{10}$$

Нетрудно показать, что справедлива следующая теорема.

Теорема 2. *Предельное значение* $F_k^{(2)}(w_1, w_2) = \lim\limits_{\varepsilon \to 0} F_k^{(2)}(w_1, w_2, \varepsilon)$ *решения системы* (10) *имеет вид*

$$F_k^{(2)}(w_1, w_2) = R_k \exp\left\{ \frac{(jw_1)^2}{2} q_{11} + \frac{(jw_2)^2}{2} q_{22} + jw_1 jw_2 q_{12} \right\},$$

где R_0, R_1, R_2, x_1, x_2 *вычисляются по формулам* (5), (6), *а величины* q_{11}, q_{12}, q_{22} *определяются из системы уравнений:*

$$\begin{cases} \mu_1 f_1 + \mu_2 f_2 - (\lambda_1 + \lambda_2)(1+x) f_0 - \gamma_1 R_0 q_{11} - \gamma_2 R_0 q_{12} = 0, \\ (\lambda_1 + \lambda_1 x) f_0 - \lambda_1 x R_0 - \mu_1 f_1 + \lambda_1 R_1 + \gamma_1 R_0 q_{11} = 0, \\ (\lambda_2 + \lambda_2 x) f_0 - j\lambda_2 x w_2 R_0 - \mu_2 f_2 + \lambda_1 R_2 + \gamma_2 R_0 q_{12} = 0, \\ \mu_1 g_1 + \mu_2 g_2 - (\lambda_1 + \lambda_2)(1+x) g_0 - \gamma_1 R_0 q_{12} - \gamma_2 R_0 q_{22} = 0, \\ (\lambda_1 + \lambda_1 x) g_0 - \mu_1 g_1 + \lambda_2 R_1 + j\gamma_1 R_0 q_{12} = 0, \\ (\lambda_2 + \lambda_2 x) g_0 - \lambda_2 x R_0 - \mu_2 g_2 + \lambda_2 R_2 + \gamma_2 R_0 q_{22} = 0, \\ -\lambda_1 x f_0 + \lambda_1 f_1 + \lambda_1 f_2 = \gamma_1 R_0 q_{11} - \dfrac{1}{2}\lambda_1 x R_0 - \dfrac{1}{2}\lambda_1 R_1 - \dfrac{1}{2}\lambda_1 R_2, \\ -\lambda_2 x f_0 + \lambda_2 g_1 + \lambda_2 g_2 = \gamma_2 R_0 q_{22} - \dfrac{1}{2}\lambda_2 x R_0 - \dfrac{1}{2}\lambda_2 R_1 - \dfrac{1}{2}\lambda_2 R_2, \\ \lambda_2 f_1 + \lambda_1 g_1 + \lambda_2 f_2 + \lambda_1 g_2 - \lambda_2 x f_0 - \lambda_1 x g_0 = \gamma_1 R_0 q_{12} + \gamma_2 R_0 q_{22}. \end{cases} \tag{11}$$

Возвращаясь к заменам (9), получим, что

$$H^{(2)}(u_1, u_2) \approx \sum_k F_k^{(2)}(w_1, w_2) = \exp\left\{ \frac{(jw_1)^2}{2} q_{11} + \frac{(jw_2)^2}{2} q_{22} + jw_1 jw_2 q_{12} \right\}.$$

Тогда из (7) для допредельной характеристической функции

$$H(u_1, u_2) = M e^{ju_1 i_1(t) + ju_2 i_2(t)}$$

можно записать следующее равенство:

$$H(u_1, u_2) = \exp\left\{ j\frac{u_1}{\sigma} x_1 + j\frac{u_2}{\sigma} x_2 + \frac{(ju_1)^2}{2\sigma} q_{11} + \frac{(ju_2)^2}{2\sigma} q_{22} + \frac{ju_1 ju_2}{\sigma} q_{12} \right\}.$$

То есть двумерный процесс $\{i_1(t), i_2(t)\}$ имеет асимптотически нормальное распределение с математическими ожиданиями x_1/σ и x_2/σ, дисперсиями q_{11}/σ, q_{22}/σ и коэффициентом корреляции q_{12}/σ.

4. Заключение

Таким образом, в работе было проведено исследование RQ-системы $M_2|M_2|1$ с двумя ИПВ методом асимптотического анализа в условии большой задержки. Получено, что двумерный случайный процесс, характеризующий число заявок в каждом ИПВ, асимптотически имеет вид нормального распределения.

ЛИТЕРАТУРА

1. Shneps-Shneppe M. A. The effect of repeated calls on communication system // Proceedings of the 6th International Teletraffic Congress, Munich, 1970.
2. Eldin A., Lind G. Elementary Telephone Traffic Theory. Ericsson Public Telecommunications, 1971.
3. Gosztony G. Repeated call attempts and their effect on traffic engineering // Budavox Telecommunication Review. 1976. №2. P. 49-100.
4. Artalejo J. R., Gomez-Corral A. Retrial Queueing Systems: A Computational Approach. Berlin: Springer, 2008.
5. Falin G. I. Limit theorems for queueing systems with repeated calls // 4th Int. Vilnius Conf. on Probability Theory and Mathematical Statistics, Vilnius, 1985.
6. Deepak T., Dudin A., Joshua V., Krishnamoorthy A. On an M(X)/G/1 Retrial System with Two Types of Search of Customers from the Orbit // Stochastic Analysis and Applications. 2013. V. 31. №1. P. 92-107.
7. Назаров А. А., Терпугов А. Ф. Теория массового обслуживания: учебное пособие. Изд-во НТЛ, Томск, 2010.
8. Назаров А. А., Моисеева С. П. Метод асимптотического анализа в теории массового обслуживания. Изд-во НТЛ, Томск, 2006.
9. Yang W. S., Dug H. M. M/M/c Retrial Queue with Multiclass of Customers // Methodology and Computing in Applied Probability. 2014. V. 16. №4. P. 931Ц949.
10. Rengnanathan N., Kalayanaraman R., Srinivasan B. A finite capacity single server retrial queue with two types of calls // International Journal of Information and Management Sciences. 2002. V. 13, №3. P. 47-56.
11. Avrachenkov K., Nain P., Yechiali U. A retrial system with two input streams and two orbit queues // Queueing System. 2014. V. 77. №1. P. 1-31.

ARCHITECTURE OF SIMULATION OF MOBILE OBJECTS RADIO FREQUENCY IDENTIFICATION SYSTEM

A.A. Larionov[1], *R.E. Ivanov*[1]
[1]V. A. Trapeznikov Institute of Control Sciences of Russian Academy of Sciences, Moscow, Russia

Abstract

Architecture of discrete-event simulation of radio frequency identification of mobile objects by roadside units (RSU) based on UHF RFID EPC Class 1 Generation 2 protocol is given in this paper. The simulation performs high detailed UHF RFID radio protocol simulation in order to gain reliable system performance assessments in the road traffic environment.

Keywords: UHF RFID, EPC Class 1 Generation 2, simulation modeling, OMNeT++

АРХИТЕКТУРА ИМИТАЦИОННОЙ МОДЕЛИ СИСТЕМЫ РАДИОЧАСТОТНОЙ ИДЕНТИФИКАЦИИ МОБИЛЬНЫХ ОБЪЕКТОВ

А.А. Ларионов[1], *Р.Е. Иванов*[1]
[1]Институт проблем управления им. В.А. Трапезникова РАН, Москва, Россия
larioandr@gmail.com, iromcorp@gmail.com

Аннотация

В статье приведена архитектура дискретно-событийной имитационной модели системы радиочастотной идентификации мобильных объектов базовыми станциями на основе протокола UHF RFID EPC Class 1 Generation 2. Особенностью модели является высокая детализация радиопротокола UHF RFID, позволяющая получить достоверные оценки производительности системы в условиях дорожного трафика.

Ключевые слова: UHF RFID, стандарт EPC Class 1 Generation 2, имитационное моделирование, OMNeT++

1. Введение

Вопросы применения технологии радиочастотной идентификации в УКВ-диапазоне (UHF RFID) на транспорте в последнее время получили широкое развитие в исследованиях, связанных с задачами повышения безопасности дорожного движения, создания интеллектуальных транспортных систем, организации взаимодействия машин друг с другом и с сетевой инфраструктурой. Ввиду высокой стоимости оборудования для подобных задач, строительство реалистичных макетов оказывается весьма затратным, из-за чего основными инструментами проектирования и предварительной оценки производительности становятся модели, позволяющие рассчитывать характеристики систем радиочастотной идентификации.

Радиооборудование стандарта EPC Class 1 Generation 2 [1] состоит из активных считывателей и пассивных радиометок. Считыватели создают поле, из которого пассивные метки, не содержащие собственных источников питания, получают эренгию. Считыватели передают меткам команды, получая и обрабатывая которые, метки формируют и передают свои ответы. Каждая метка может содержать несколько банков памяти, в которые можно записать до 512 бит информации. В частности, каждая метка содержит идентификатор EPC, типичная длина которого – от 32 до 96 бит. Значения банков памяти передаются считывателю в ответах на его команды. Для борьбы с коллизиями считыватель использует механизм Slotted ALOHA, выбирая некоторое число временных слотов (Q) и передавая в начале каждого слота специальную команду (Query или QueryRep). Каждая метка, получив команду Query, выбирает случайный слот, в который ей будет передан ответ считывателю (цикл между соседними командами Query называется раундом инвентаризации). если две или более меток передают ответ в одном и том же слоте, происходит коллизия. Для борьбы с коллизиями считыватель может увеличить число слотов (Q), однако при этом также увеличивается длина одного раунда инвентаризации. Типичные расстояния между меткой и считывателем могут достигать 10–15 метров, значения числа слотов - от 2 до 2^{15}.

Поскольку метки получают энергию от считывателя, для успешной передачи ответа требуется не только отсутствие коллизий, но и сохранение метки в поле считывателя в течение всего раунда инвентаризации. В то же время, наличие интерференции из-за многолучевого характера распространения волны, создаваемой считывателем, а также даже кратковременное появление препятствия между считывателем и меткой, может привести к потере энергии меткой и невозможности произвести ей передачу своего идентификатора считывателю. Кроме того, ошибка может произойти если некоторые кадры от считывателя или метки были переданы с ошибками, что весьма вероятно при большом расстоянии между считывателем и меткой, а также большом уровне шума. Ситуация дополнительно осложняется тем, что метки, если их располагать на автомобилях, находятся в области действия считывателя достаточно короткое время, зачастую -

менее одной секунды, в течение которого может произойти ограниченное количество раундов инвентаризации. Из-за перечисленных факторов даже наличие прямой видимости между считывателем и меткой не может гарантировать успешного считывания метки за ограниченное время. Это утверждение было подтверждено в ходе эксперимента, проведенного в городе Казань, в конце 2014-го и начале 2015-го годов, когда на нескольких точках над дорогой размещались RFID-ситыватели, а около тысячи рейсовых автобусов были оснащены RFID-метками. Несмотря на наличие прямой видимости между метками и считывателями, показатель чтения на разных постах составлял от 90% до 95%.

Для того, чтобы учесть и адекватно смоделировать перечисленные факторы, следует рассмотреть, какие особенности стандарта влияют на длительность раундов, вероятность успешной передачи команд считывателя и ответов метки, а также на скорость передачи данных. Оказывается, что таких особенностей очень много. Так, в передаче команд считывателя используется кодирование PIE (Pulse-Interval Encoding), в котором длительности нулей и единиц различаются почти в два раза. При этом длительности преамбул и контрольных сумм могут значительно превышать длительности передачи полезных данных, поэтому на длительность передачи команд считывателя серьезнейшее влияние оказывают сами передаваемые команды, точнее – вид их битовой последовательности. Метки могут использовать различные схемы кодирования (FM-0, Miller-2, Miller-4, Miller-8), выбор которых вкупе с прочими параметрами может увеличивать или сокращать длительность длительность передачи ответа в несколько раз. Как символьная скорость передачи данных, так и длительности периодов ожидания и преамбул определяются набором (базовых) интервалов, наиболее значимые из которых – *Tari* и *RTcal*, *reader-tag calibration*), могут меняться в очень широких пределах (например, *Tari* может принимать значения от 6.25 мкс до 25 мкс). Наконец, прежде, чем получить ответ от метки о содержании хотя бы ее банка EPC, считыватель должен произвести с ней обмен несколькими командами и ответами, а в случае чтения другого банка памяти число команд увеличивается как минимум в два раза.

Целью работы является разработка архитектуры имитационной модели, с помощью которой можно было бы выявить условия, влияющие на снижение вероятности успешного чтения метки в дорожных условиях (результаты, получаемые с помощью этой модели, должны коррелировать с данными, полученными в ходе эксперимента), а также смоделировать систему радиочастотной идентификации, действующую внутри крупного города, в которой данные от точек фиксации передаются в центр обработки данных по существующим телекоммуникационным сетям. Учет второго требования привел к выбору в качестве системы моделирования OMNeT++ [2], в составе которой есть готовые реализации многих стандартных протоколов и технологий передачи данных, а сама система обладает широкими воз-

можностями визуализации эксперимента, сбора и обработки статистики и прочими необходимыми готовыми инструментами, оставаясь существенно проще в использовании, чем более популярная система NS-3 [3]. Учет первого требования привел к тому, что для адекватного моделирования работы системы радиочастотной идентификации в дорожных условиях требуется использовать детальную модель радиопротокола, а также иметь возможность использовать многолучевые модели распространения сигнала. При этом моделирование передачи сигнала должно осуществляться таким образом, чтобы учитывать изменение поля по мере перемещения автомобиля в зоне действия считывтеля. К сожалению, такой модели в OMNeT++ на момент начала работ, равно как и на момент публикации доклада, нет. Фактически, учет перечисленных ранее особенностей стандарта приводит к необходимости разработки модели, по степени детализации сравнимой с реальным считывателем и меткой.

Из-за сложности модели, отдельного внимания заслуживает ее архитектура, подробное описание которой приводится в настоящем докладе. Доклад организован следующим образом: в главе 2 описывается общая архитектура модели, ее крупные модули и их назначение; в главе 3 определяется модель RFID-считывателя, его строение и соответствие протоколу; в главе 4 подробнение описывается механизм моделирования передачи сообщений между активным и пассивным устройствами, приводятся аналитические модели, используемые в системе; в главе 5 излагаются особенности моделирования RFID-метки.

2. Архитектура имитационной модели

Помимо моделирования самого протокола EPC Class 1 Generation 2, возникает необходимость моделировать окружение и сами объекты, участвующие в процессе идентификации. За основу при разработке метода моделирования передачи сигналов между считывателями и метками была взята библиотека Veins [4], которая успешно используется для моделирования WiFi-соединений между мобильными абонентами и базовыми станциями. Каналы связи между считывателями и центром обработки данных моделируются при помощи библиотеки INET [5].

Можно выделить модули четырех типов: модуль сценария (Scenario), модуль базовой станции (Station), модуль идентифицируемого объекта (Object) и модуль центра обработки данных (Server, ЦОД). Модуль сценария содержится в единственном экземпляре; модули остальных типов используются в неограниченном количестве.

Ключевым моментов в процессе моделирования является необходимость задания расстояния между различными объектами, участвующими в процессе идентификации. Таким образом, необходимо задавать и изменять во времени координаты базовых станций и идентифицируемых объектов. Для этого используется модуль Mobility, ассоциированный с каждой станцией или объектом и обеспечивающий доступ к положению модуля на модельной

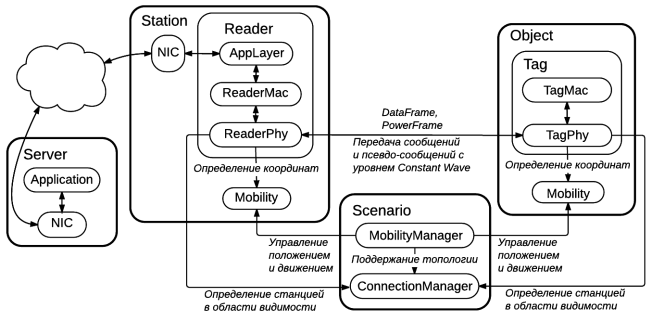


Рис. 1: Архитектура имитационной модели

карте, ее скоростью и направлением движения. Для управления параметрами движения используется менеджер мобильности (MobilityManager), который может как самостоятельно задавать перемещения всех объектов, так и использовать внешние генераторы движения. Примером такого является модель системы идентификации транспортных средств в городских условиях, где движением автомобилей управляет внешний модуль SUMO [6].

Взаимодействие между модулями основано на передаче сообщений. При этом число объектов и станций в процессе моделирования работы протокола может быть очень большим. Чтобы избежать отправки сообщений всем модулям и ограничить их теми, кто оказывается в зоне прямой видимости, используется менеджер соединений (ConnectionManager). Данный модуль поддерживает граф связности, основанный на расстояних между объектами и станциями в данный момент времени. Модуль взаимодействуют с MobilityManager для актуальной топологии. Оба менеджера содержаться в сценарии (Scenario) модели.

Модуль Object является составным и содержит в себе модуль Mobility и модуль Tag (метка). Также является составным модуль Station, содержащий модули Mobility, NIC (Network Interface Card) и Reader (считыватель). Непосредственно в моделирования протокола участвуют только модули считывателя и метки, остальные же модули создают необходимое для работы протокола окружения. Далее, опишем их более подробно.

3. Модель RFID-считывателя

Модель RFID-считывателя состоит и трех уровней (см. рис.2): уровня приложений (Reader APP), канального уровня (Reader MAC) и физического уровня (Reader PHY). На физическом уровне модели производится моделирование передачи и приема радиосигналов, на канальном уровне реализован стандарт EPC Class 1 Generation 2, а уровень приложений осу-

ществляет управление моделью считывателя и взаимодействует с сетью. Следует заметить, что стандарт не определяет различных уровней, а разделение считывателя на уровне было введено исключительно для удобства. Можно сазать, что все моделирование стандарта осуществляется на канальном уровне модели.

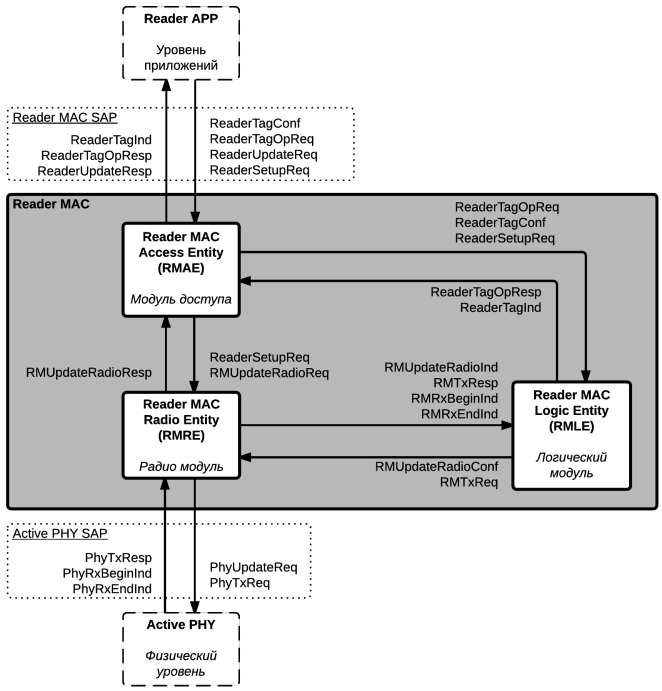


Рис. 2: Архитектура модели считывателя

Задача уровня приложений – моделирование различных сценариев использования считывателя и взаимодействие с иными компонентами модели (базовыми станциями сети передачи данных, центром обработки данных и пр.). Уровень приложений управляет включением и выключением считывателя, производит его настройку и осуществляет прием данных о считанных метках. Кроме того, при чтении очередной метки уровень приложений может запросить канальный уровень выполнить ту или иную операцию над ней, например – прочитать дополнительный банк памяти или записать данные на метку. Также уровень приложение ведет журнал считанных меток, с помощью которого в дальнейшем вычисляется вероятность идентификации объектов. Кроме значений банков памяти, в журнале сохраняется число чтений метки, значения уровней приема от метки.

Канальный уровень модели RFID-считывателя реализует все основные механизмы, описанные в стандарте EPC Class 1 Generation 2. Он осуществ-

ляет управление питанием и выбор несущих частот, ведет инвентаризацию меток и осуществляет дополнительные операции доступа, о которых его запрашивает уровень приложений. На канальном уровне производится формирование информационных кадров, преамбул, вычисление длительности передачи, а также обработка кадров, принятых от меток.

Канальный уровень состоит из трех более простых модулей: модуля доступа (Reader MAC Access Entity, RMAE), логического модуля (Reader MAC Logic Entity, RMLE) и радио-модуля (Reader MAC Radio Entity, RMRE).

Модуль доступа (RMAE) осуществляет взаимодействие с уровнем приложений. Можно сказать, что модуль предоставляет интерфейс (Service Access Point, SAP) для сервисов верхних уровней, в данном случае – для уровня приложений. Основная задача модуля – распределять сообщения, пришедшие сверху, между двумя другими модулями канального уровня, а также передавать сообщения от них наверх.

Логический модуль (RMLE) реализует логический уровень стандарта. его функции:

- Формирование кадров и обработка ответов от меток: все, что касается формирования кажров, вычисления их длительности и добавления преамбул, реализуется модулем RMLE. Он также обрабатывает кадры, которые были успешно получены от меток.
- Реализация цикла инвентаризации: сразу после получения команды включения, модуль начинает слать запросы Query/QueryRep, а также обрабатывать в соответствии со стандартом ответы от меток. При получении идентификатора EPC от метки, информирует об этом уровень приложений через модуль RMAE.
- Выполнение операций доступа к метке: уровень приложений, получив индикацию о прочтении очередной метки, может запросить модуль RMLE выполнить одну или несколько последовательных операций доступа над меткой. По выполнении каждой операции, модуль сообщает уровню приложений ее результат. На момент публикации, были реализованы операции обычного чтения и записи произвольных банков памяти. Следует заметить, что модуль спроектирован таким образом, что добавление новых операций доступа не представляет большой сложности и может при необходимости быть быстро реализовано.

Все процедуры, встроенные в модуль RMLE, предельно точно реализуют соответствующие процедуры стандарта. В частности, при определении длительности передачи кадров рассчитывается длительность символов 0 и 1, их количество в кадре, и на основе этого определяется длительность. Также с высокой точностью моделируются все таймауты, определенные стандартом. Высокая точность позволяет максимально точно смоделировать цикл инвентаризации и ошибки, возникающие вследствие кратковременных потерь энергии метками в течение цикла.

Радио-модуль (RMRE) реализует управление питанием и частотами, FHSS и прочие механизмы, а также взаимодействует с физическим уровнем модели, передавая ему команды изменения мощности передатчика и несущей частоты, запросы передачи кадров, и получает индикации начала и окончания приема кадров. Этот модуль отвечает за ключение и выключение считывателя, производит передачу и прием кадров, а также информирует логический модуль RMLE о событиях начала и окончания приема кадров. Модуль может моделировать работу на одной заданной несущей частоте, а также режим FHSS. Поскольку в любом случае считыватель должен периодически, раз в несколько сотен милисекунд, отключать передатчик, модуль моделирует такое поведение за счет плавного включения и выключения питания через заданные интервалы. Следует заметить, что модуль RMRE работает с относительными значениями энергии, то есть значениями от 0.0 до 1.0, а абсолютная мощность настраивается на физическом уровне.

Канальный уровень сам по себе практически не содержит никаких настроек, а получает все параметры от уровня приложений при включении. Это позволяет строить гибкие эксперименты, в ходе которых приложение может, например, адаптировать настройки считывателя к конкретной среде работы динамически.

4. Модель физического уровня RFID-считывателя

Задача модели физического уровня считывателя — моделирование передачи кадров, поступающих от MAC-уровня, через радиосреду, и моделирование приема кадров от меток. Модель реализуется модулем ActivePhy.

Так считываемые метки пассивные, для питания меток считывателю необходимо излучать постоянную волну (Constant Wave, CW) на протяжении всего времени, когда не ведется передача информационных сообщений. Так как в дискретно-событийной парадигме не предусмотрен процесс непрерывного моделирования, то для передачи метке информации о уровне CW используется псевдосообщение PowerFrame. Получив сообщение запуска от MAC-уровня, ActivePhy включается и начинает рассылку данных сообщений, содержащих данные об излучаемой мощности, которые предназначены для определения метками уровней энергии в местах их нахождения и планирования событий включения и выключения.

Когда MAC-уровню требуется передать кадр, ActivePhy помещает передаваемый кадр в сообщение DataFrame и передает его все меткам, с которыми он в настоящее время связан. Помимо самого передаваемого кадра, это сообщение содержит информацию о физическом представлении сигнала: время отправки, задержку, длительность, мощность, затухание и прочее. если метка передала кадр считывателю, ActivePhy анализирует принятый кадр, на основе многолучевой модели затухания рассчитывает мощность принятого сигнала, вычисляет вероятности ошибок (BER/SER) и на основе полученных данных решает, был ли успешно принят кадр. В

случае успешного приема, кадр передается на обработку MAC-уровню. Работа модуля ActivePhy завершается при получении сообщения останова от MAC-уровня. Помимо этого, MAC-уровень периодически пересылает сообщения управления частотами и питанием, на основе которых ActivePhy расчитывает излучаемую мощность и передает соответствующие изменения в сообщениях PowerFrame.

Наиболее сложная часть работы ActivePhy, — прием сообщения — происходит в несколько этапов. При получении сообщения от метки, модуль вычисляет функцию затухания сигнала на основе расстояния между источником и получателем и получает мощность сигнала на приеме (receiving power). На основе этой мощности вычисляется соотношения сигнал-шум и сигнал-шум-интерференция (SNR и SINR), которое используется для расчета символьной вероятности ошибки приема (SER). Данная вероятность для канала с аддитивным гаусовым шумом (AWGN) рассчитывается по формуле [7]:

$$P_b = 2Q\left(\sqrt{\frac{ME_S}{N_0}}\right)\left[1 - Q\left(\sqrt{\frac{ME_S}{N_0}}\right)\right] \tag{1}$$

где E_S — энергия символа, $N_0/2$ — спектральная плотность мощности шума канала с аддитивным гауссовым шумом, Q — Q-функция, M — порядок модуляции Миллера. Для Отношения E_S/N_0 приблизительно равно отношению сигнал-шум (SNR) $\gamma = \frac{S}{N} \approx \frac{E_S}{N_0}$. Данная формула не учитывает влияние интерференции. Для этого необходимо использовать формулу для канала с релеевским замиранием (Rayleigh fading) [7]:

$$\bar{P}_b = \frac{1}{2} - \frac{1}{\sqrt{1 + 2/(M\bar{\gamma})}} + \frac{2}{\pi}\frac{\arctan\left(\sqrt{1 + 2/(M\bar{\gamma})}\right)}{\sqrt{1 + 2/(M\bar{\gamma})}} \approx \frac{1}{2M\bar{\gamma}}, \tag{2}$$

где $\bar{\gamma}$ – соотношение сигнал-шум-интерференция. На основе ser принимается решение, принят пакет или нет.

Для расчета функции затухания используется модель многолучевого распространения [8]:

$$L_{path} = \left(\frac{\lambda}{4\pi}\right)^2 \left|\sum_{n=0}^{N} G_n \Gamma_n \frac{1}{d_n} e^{-jkd_n}\right|^2, \tag{3}$$

где L_{path} — затухание сигнала при многолучевом распространении, Γ_n — коеффициент отражения n-ого отражающего объекта (включая землю), G_n — коэффициент, обусловленный диаграммой направленности антенны, d_0 — длина прямого пути, d_n — путь n-ого отраженного луча и N — общее число отражений.

В качестве основной модели используется двухлучевая, учитывающая прямое распространение и одно отражение от земли. Но при этом могут применяться модели, учитывающие большое число отражений (что характерно для тунелей, областей под мостом и складских помещений).

5. Модель RFID-метки

Модель метки разбита на два уровня — физический (Tag PHY) и канальный (Tag MAC). Моделирование стандарта осуществляется на канальном уровне, а физический уровень моделирует передачу и прием радиосигнала, а также управление частотами и мощностью. Отдельный уровень приложений в модел и метки не был выделен, поскольку метка является пассивным устройством, работа котрого управляется считывателем, и не требует динамической настройки пользователем.

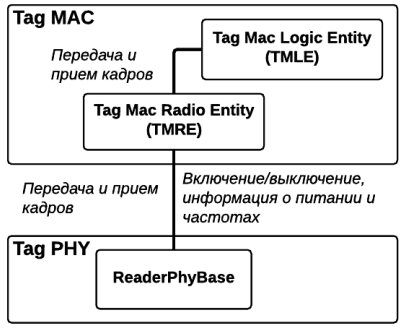


Рис. 3: Архитектура модели метки

Канальный уровень метки имеет модульную структуру и состоит из двух модулей (см. рис. 3): логический модуль(Tag MAC Logic Entity, TMLE) и радио-модуль (Tag MAC Radio Entity, TMRE). Логические модуль (TMLE) реализует логическую часть стандарта EPC Class 1 Generation 2. В его задачи входит формирование кадров, преамбул и суффиксов, определение длительности их передачи, а также обработка кадров-команд, получаемых от считывателя. Информацию о включении и отключении модуль получает от радио-модуля TMRE. Радио-модуль (TMRE) осуществляет взаимодействие с физическим уровнем. От него он узнает о включении и отключении метки из-за изменения внешнего поля, а также о получении кадров от считывателя. Также TMRE производит передачу кадров от логического модуля TMLE физическому уровню. Среди настроек канального уровня метки – размеры и начальные значения банков памяти (EPC, TID, UserMemory).

В отличии от считывателя, пассивная метка не имеет своего источника и использует энергию CW, излучаемую считывателем. если уровень

CW в месте нахождения метки падает слижком низко, метка отключается. Для отправки сообщения, метка моделирует отраженную волну CW (backscatter modulation), меняя свою отражающую способность. Таким образом, мощность излучения метки (TX power) определяется уровнем CW.

На физическом уровне, процесс включения-отключения метки, а также расчет мощности передачи сообщения, моделируется при помощи обмена PowerFrame, описанные выше. Также, как и при прибытии DataFrame, вычисляется затухание CW и рассчитывается мощность "приема" (уровень CW в месте нахождения метки). На основании этой мощности вычисляется времена включения-выключения метки и мощность передачи, с которой метка может отправить свое сообщение DataFrame.

6. Заключение

В работе представлена архитектура имитационной модели, выполненной в системе моделирования OMNeT++, с помощью которой можно определить условия, влияющие на снижение вероятности успешного чтения метки в дорожных условиях, а также смоделировать систему радиочастотной идентификации, действующую внутри крупного города, в которой данные от точек фиксации передаются в центр обработки данных по существующим телекоммуникационным сетям. В рассмотренной модели реализуется детальная модель радиопротокола, а также используются многолучевые модели распространения сигнала. Полученные результаты позволяют выявить причины, снизившие вероятность идентификации автомобилей в проведенном ранее эксперименте, а также определить параметры радиопротокола, повышающие эффективность работы системы.

ЛИТЕРАТУРА

1. "EPC Radio-Frequency Identity Protocols Generation-2 UHF RFID. Specification for RFID Air Interface. Protocol for Communications at 860 MHz – 960 MHz. Version 2.0.1". EPCGlobal Inc., 2015.
2. OMNeT++ Discrete Event Simulator. [web]. URL: https://omnetpp.org/
3. NS-3 Discrete Event Simulator. [web]. URL: https://www.nsnam.org/
4. Vehicles In Network Simulation. [web]. URL: http://veins.car2x.org/
5. INET Framework. [web]. URL: https://inet.omnetpp.org/
6. Institute Of Transportation System. Simulation of Urban MObility. [web]. URL: http://www.dlr.de/ts/en/desktopdefault.aspx/tabid-9883/FIXME .
7. A. Lazaro, D. Girbao, R. Villarino, "Effects of interferences in uhf rfid systems", Progress In Electromagnetic Research, PIER 98, 425-443, 2009.
8. Pavel V. Nikitin, Senior Member, IEEE, and K. V. S. Rao, Senior Member, IEEE, "Antennas and Propagation in UHF RFID Systems".
9. Manar Mohaisen, HeeSeok Yoon, and KyungHi Chang, "Radio Transmission Performance of EPCglobal Gen-2 RFID System", The Graduate School of Information Technology & Telecommunications INHA University, Incheon, Korea.

EFFECTIVENESS ASSESSMENT OF IT-SECURITY MANAGEMENT SYSTEM PROCESSES

I. Livshitz, D. Yurkin, A. Vinel
SPIIRAS, St.Petersburg, Russia
Livshitz.il@yandex.ru, Dvyurkin@yandex.ru, alexey.vinel@gmail.com

Abstract

The relevance of this publication is due to constant attention to the analysis and interpretation of information security management systems (ISMS) implementation results. Analysis of such projects usually takes into account only the minimum requirements, based on known methodological framework - 27000 series of ISO standard. In this study, first of all, a review of current regulatory framework ISO 27001 series was made, and secondly, a practical application of IT-Security metrics was demonstrated, that significantly expands the possibilities for assessing the ISMS effectiveness, and also the recommendations for the formation of IT-Security metrics system were provided, which are directly related to business requirements.

Keywords: Information security (IT-security) information security management system (ISMS), information security assessment, IT-Security metrics, audit.

1. Introduction

There are enough materials currently published supporting the implementation (certification) of management systems (including ISMS), as well as more cautious conservative estimates [1]. It is obvious that the success of several standards (e.g., ISO 9001, 27001, 50001 series) is caused by certain factors, the implementation of which seems appropriate in economic, technical, political and social aspects [2 — 4]. So, we need to propose an assessment - because, firstly, it clearly operates with business objectives, and secondly, it uses the economic criteria and, thirdly, it clearly reflects the problem of forming a coherent system of IT-security metrics linking business goals and evaluation of the implementation of specific projects (including standardization and / or certification). This is well stated in ISO 27000 series for the purposes of implementation of IT-security business requirements. Obviously IT-security metrics system should organically fit the terminology of business and allow stakeholders to objectively evaluate the proposed solutions. The standard approach is to use the "target" standard ISO 27004 for the process of formation, analysis and comparison of IT-security metrics. Below is the technique of the ISMS effectiveness assessment and an example of numerical (quantitative) indicators formation for the analysis of stakeholders (both within the organization and for certain external evaluation).

2. Materials dna methods

The TOP-management need obtain the clearly metrics for ISMS assessment, firstly, as metrics for regular process analysis and, secondly, as basis for financial budget for ISMS improvement. Main focus covers in numerical (quantitative) metrics for IT-security assessment.

3. Problem description

A number of publications reflected approaches to the management of losses in the system of management, organization of the system of internal audits and effective management review. It is shown that the same approach can be applied to integrated management system (IMS). Also there is well-known example of simple IT-Security metrics, which can provide the quantitative estimates as evidence of "utility" for Business. Here it is particularly important to make a comparison directly with the mechanisms of internal audit, which is precisely intended to provide "objective evidence" for senior management in order to make effective management decisions. To solve inconsistencies "blind copy" of the organizational structure, it is advisable to use the method ISMS based on ISO 27001, and propose to form IT-Security target (as well as IT-Security metrics) through priority control vital assets. This approach "breaks" simple copy hierarchical structure formation IT-Security purposes, and introduces, as required in the ISMS, inventory and asset management, which the organization must be protected (see. Fig. 1).

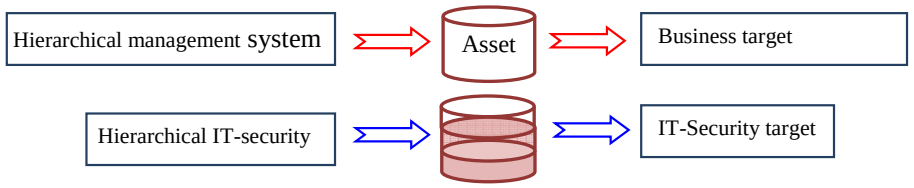


Fig. 1: Dependence between Hierarchical System and Business target

In assessing the results of measurements the targeted requirements for performance should be provided (at the level of the organization), which are directly related to the objectives in the field of information security, in particular - ISMS objectives, goals, application of the measures and means to ensure information security, which should be selected (Sec. 5.4.5 standard [5]). Thus, there is a scheme proposed (see. Fig. 1) complying with ISO 27000 series, and at the same time avoiding "gaps" in the reflection of business objectives for the purposes of information security.

4. Results

Methodology of forming a system of numerical (quantitative) IT-Security metrics corresponding hierarchical system of business objectives of the organi-

zation, which is formed to assess the effectiveness of the ISMS and security of vital important assets of an organization, should consider a number of important points from the perspective of business requirements following important parameters: input value of the economic nature, choice of assets, analysis of the data and dissemination of reports [6, 7]. The following economic indicators can be considered as input variables for the set IT-Security metrics on the basis of set business objectives at different levels of the hierarchy of an organization:

- The costs of providing business (auditing service management);
- Sales (target, current) and the allowable loss (direct or indirect).

Thus, it is rational to speak about static measurements (the "here and now"), as well as to create a forecast model, i.e. form a dynamic evaluations with specific "planning horizon". With regard to the ISMS, this means that the "stakeholders" define those assets that are vital to achieve higher economic performance of business and aiming to provide safety. Also it is necessary to determine such IT-Security metrics that objectively reflect the mismatch (in terms of the applicable standards ISO), for example:

- insolvency plan for risk treatment (eg, threats can bypass existing IT-Security controls);

Measurement method should be based on the attributes of the selected objects measurements, for examples of objects of measurement include:

- state information assets protected measures and means of information security;
- the effectiveness of IT-security processes (including realized in ISMS);

The measurement method may use measurements of objects and attributes from a variety of sources, for example:

- risk analysis, risk assessment and IT-Security risk treatment;
- reports on internal and / or external audits and reports of incidents.

In addition to the requirements of ISO 27000 series, the presented methodology includes a set of steps (see. Table 1: Stages 1 - 4), directly related to the provision of "connections" business goals and objectives of the IT-Security through the list of identified assets to be protected in the organization. Description of steps from proposed methods is shown in Table 1.

5. Conclusions

Presented methodology based on ISO 27004 allows to obtain estimates of ISMS implementation efficiency that are suitable for analysis and managerial decision-making by senior management, using a system of metrics as "work element" in the existing management system of the organization. These estimates can be considered in relation to ISMS (or ISM). To form an estimate of the level of security, it is necessary to form the 'through' system of IT-security metrics, creating an optimal (hierarchical) structure (by processes, subdivisions or the products / services) with taking into account the vital assets. It is necessary

#	Methodology step	ISO 27004
1.	Determine the scope of ISMS certification	–
2.	Determination of the list of protected assets	–
3.	Defining measures of IT-security (from the Statement of Applicability)	–
4.	Defining the implementation measures of IT-security	–
5.	Determination of measurement objects (prefix "O")	+
6.	Defining Attributes	+
7.	Determination of the method of measurement	+
8.	Determination of the main measures	+
9.	Determination of measurement functions	+
10.	The definition of a derivative action measure	+
11.	Determination of the analytical model	+
12.	Defining of indicators	+
13.	Defining criteria decision	+
14.	Determination results of measurements	+

Table 1: Description of steps from proposed IT-Security measure method.

to ensure a constant closed-loop control (PDCA), and strive to implement and monitor PDCA "mini-cycles" at the appropriate levels of management hierarchy.

REFERENCES

1. The ISO Survey of Management System Standard Certifications – 2013.
2. Information technology — Security techniques — Information security management systems — Requirements: ISO/IEC 27001:2013, International Organization for Standardization, 2013.
3. Information technology — Security techniques — Information security management systems — Overview and vocabulary: ISO/IEC 27000:2014, International Organization for Standardization, 2014.
4. Societal security — Business continuity management systems — Requirements: ISO 22301:2012, International Organization for Standardization, 2012.
5. ISO/IEC 20000-1:2011 ''Information technology — Service management — Part 1: Service management system requirements", International Organization for Standardization, 2011.
6. Livshits I.I., The concept of assessing the IT service providers information security level for industrial facilities // SPIIRAS Proceedings. 2014, vol 4, pp. 117 – 135.
7. Livshits I.I., Joint information security audit and availability of information systems problems solving on the basis of international standards ISO // Informatizatiy and Svyaz, 2013 vol. 6, pp. 48 – 52.

THE CONCEPT OF INFORMATION SECURITY PROVIDERS OF IT-SERVICES

I. Livshitz, D. Yurkin, A. Vinel
SPIIRAS, St.Petersburg, Russia
Livshitz.il@yandex.ru, Dvyurkin@yandex.ru, alexey.vinel@gmail.com

Abstract

Currently providers of IT-services are forced to deal with the significantly increased number of threats in information security (IT-Security). Accordingly, it is vital to perform a study of the problems of assessing the IT-Security competence within IT-service provider. In the issue proposed the concept of assurance, which allows to consider the most important for IT-service providers is threats and to propose an approach based on the use of modern risk-oriented ISO standards. The concept of assessing IT-service providers consists of 2 basic principles and several extensions that allow to take into account the performance of specific requirements on IT-security and provide the ability to assess (qualitatively or quantitatively) as part of scheduled inspections (audits).

Keywords: Information security (IT-Security); standard; service; system of information security management (ISMS); statistics; correlation.

1. Introduction

A number of modern publications [1 – 2] addressed the issue of applicability of various management systems to support decision making of senior management, and, as a logical consequence of the provision of quality it services, including IT-security. Now, objectively, for IT-service providers consistently manifested a significant amount of important (critical) threats, due to the emergence of new attack vectors ("targeted attack"), as well as insufficient development and risk management in relation to the previously known threats and vulnerabilities. Accordingly, it is of interest to study the problem of assurance as to the problem of "culture of production" for IT-service providers and assessing their competencies: the existence and level of implementation of standards and methodologies that are applicable for the purposes of establishing and maintaining the required IT-security level.

2. Materials and methods

The question remains about the existing methods of selection of measures and means to ensure IT-security, the degree of adequacy of such measures on the assessment of their effectiveness throughout the life cycle of IT-services – from inception, testing, operation, maintenance. But the questions of "culture", the commitment of the IT-service providers to defined standards, a set of "best practices" and of IT-security standards (on the provider side) – are, objectively, not less important and their evaluation in terms of assurance on the basis of a common methodology is timely and practically demanded. A study poses the following problem is considered based on public ratings and known IT-security standards, the possible competence of IT-services and to elaborate the concept of assessment of the IT-security level.

3. Problem description

The original data has been selected form the public reviews CNews Analytics, Pwc, KPMG, which provides public data on the ratings and performance of IT companies in Russia. From the array data were selected only providers of IT services, i.e. companies, in the "Sphere of activity" which accurately stated "IT-services". Additionally, we examined the public rating of [1-2], which included data of the largest providers of IT-services.

4. Results

The results of the research summary on the evaluation of IT-service providers for the last 3 years (2011 – 2013) and the stated competencies provided in Fig. 1. In accordance with the problem of in this issue – "how to evaluate the IT-service providers level, based on their own competences?", it is interesting to examine the range (possible dependencies) between their own competence of IT-service providers (from the TOP 100 in the rating of Russia) and

their position (including in the dynamics for revenue growth, according to the change of place in the rating).

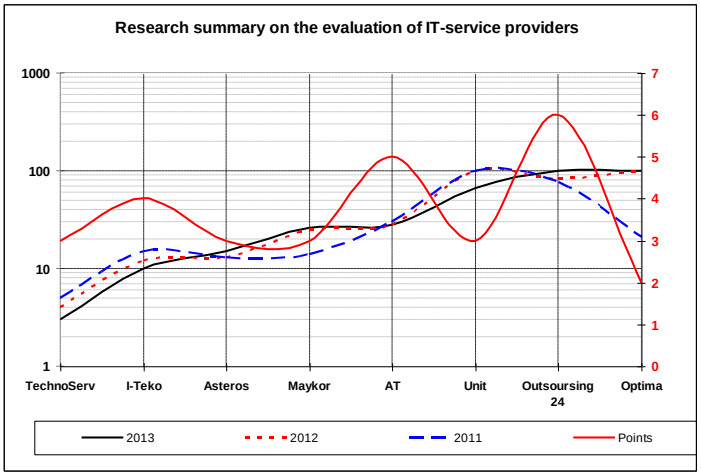


Figure 1: The results of the research summary on the evaluation of IT-service providers

Such issues are proposed to study three:

- Assessment of the correlation "The revenue Growth and the impact of competence";
- Evaluation of correlation "The ranking and influence competencies";
- Evaluation of correlation "The revenue Growth and a change in the rating".

To perform these studies used mathematical research apparatus correlations. Detailed results of studies of the correlation of ratings the ratings of the IT-service providers for the last 3 years (2011 – 2013) and the stated competencies are presented in Fig. 2.

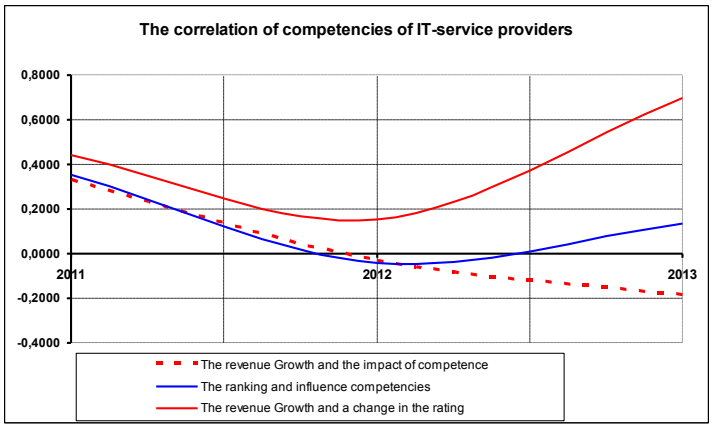


Figure 2: The correlation of competencies of IT-service providers

The results of studies of the dependency of the ratings of the providers of it services (Fig. 1) lead to the following conclusion: factor only in the presence of a large number of their own competencies (over 4 points) does not automatically lead to a high place in the rankings and is also not affected explicitly on revenue growth, but a certain "threshold" of their own competencies (minimum of 3 points, for example: ISO 9001, ISO 20000, ISO 27001) affects explicitly on revenue growth [3]. Analysis of the results of studies of correlations (3-m discrete values) providers of it services (Fig. 2) leads to the following conclusions:

- Correlation "The revenue Growth and influence competencies" demonstrates the transition from a direct correlation to back with approximately equal weak effect in

2011 and 2013 (on a scale of Chedoke less than 0.3). It is interesting the fact of the zero crossing in 2012, both in this and the following correlation;

- Evaluation of correlation "The ranking and the influence of competences" – weak in 2011 and 2013 (on a scale of Chedoke less than 0.3). Seems interesting fact increase direct proportion in 2013, but the influence of the different competencies on revenue growth and rating position requires further review and exceptions "false" correlations.
- Evaluation of correlation "The revenue Growth and rating change" is subject to significant change. This correlation provides a reasonable basis to assume that after 2012, the main influence on the rating gives exactly the revenue growth.

The concept of assessing the IT-security level of IT-service providers consists of 2 basic principles and several extensions (which may reflect a specific requirement for specialization in servicing government, industry and other customers).

1. A basic principle IT-service providers must implement in its operations management system based on national and international standards, minimally adequate for the construction of ISMS and provide the required set of measures and IT-security, adequately identified, evaluated and ranked IT-security risks. This principle can be assessed (qualitatively or quantitatively) through scheduled audits of the ISMS [1, 4].

2. Underlying the principle of sufficient IT-service providers must implement in the control system of the complex of international standards, sufficient to provide services to an agreed level of quality with regard to certain IT-security risks and with the additional requirements of stakeholders. Recommended the establishment the framework, which the implementation of this principle can be assessed (qualitatively or quantitatively) [1, 4].

3. Advanced the principle of "State regulation" – IT-service providers must implement in their management system requirements established by the regulators, IT-security specific procedures, requirements for their availability and reliability, etc. This principle can be assessed (qualitatively or quantitatively) as part of routine inspections by regulators.

4. Advanced the principle of "Industry regulation" IT-service providers must implement in their management system the specific IT-security procedures (NERC, ISAGO, etc.). This principle can be assessed (qualitatively or quantitatively) as part of scheduled inspections of licensees by independent auditors.

5. Advanced the principle of "best practice" IT-service providers must implement in their management system "best practices" adopted for IT-services (ITIL, Cobit, SOX, Basel, COSO, etc.). This principle can be assessed (qualitatively or quantitatively) as part of scheduled inspections on a voluntary basis by independent auditors.

5. Conclusions

The concept of assessing the IT-security level of IT-service consists of 2 basic principles and several extensions that can be tailored to the specific requirements for the specialization in servicing government, industry and other customers with specific IT-security in the provision of IT-services throughout the life cycle, and provide the ability to assess (qualitatively or quantitatively) as part of scheduled inspections (audits).

6. References

1. Livshits I.I. Approaches to solving the problem of losses in integrated management systems // ISO // Informatizatiy and Svyaz, 2013 vol. 1, pp. 55 – 60.
2. Livshits I.I., The concept of assessing the IT service providers information security level for industrial facilities // SPIIRAS Proceedings. 2014, vol 4, pp. 117 – 135;
3. Information technology– Security techniques– Information security management systems – Requirements: ISO/IEC 27001:2013, International Organization for Standardization, 2013;
4. Livshits I.I., Joint information security audit and availability of information systems problems solving on the basis of international standards ISO // Informatizatiy and Svyaz, 2013 vol. 6, pp. 48 – 52.

IT-SECURITY ASSESSMENT IN TELECOMMUNICATION SYSTEMS

I. Livshitz, D. Yurkin, A. Venel
SPIIRAS, St.Petersburg, Russia
Livshitz.il@yandex.ru, Dvyurkin@yandex.ru, alexey.vinel@gmail.com

Abstract

This issue covers the methodological approach to information security (IT-Security) assessment in telecommunication systems (TS) on the basis of technical and economic analysis of their availability. The mathematical expressions have been proposed to calculate the effectiveness of IT-Security management based on security metrics, as well as its cost-effectiveness. The advantages of this methodological approach can be used to obtain a numerical estimate of the data protection level in TS, by means of regular information security management systems (ISMS) audits.

Keywords: Telecommunication systems (TS), information security (IT-security), information security management system (ISMS), information security assessment.

1. Introduction

Currently, the provision of standard mode of operation of technical objects (telecommunications) infrastructure is one of the most important tasks of modern society. The operation of such facilities is based on active use of TS, which is an important component of information systems (IS). The correct TS operation directly addresses the problem of ensuring the IT-security of data passing via technical channels, as well as data stored and processed for specified purposes. Standard 27001 contains the following definition - "the security of information (data): the state of protection of information (data) ensuring its confidentiality, availability and integrity" [1]. When analyzing the factors influencing the availability of TS, the following risk factors are taken into account: the impact of external potential threats (e.g. - intruders), and inherent in any complex system aspects of internal threats. The different approaches can be applied to assess the accessibility threats, but the best practice seems to apply the series of ISO 27001 standard.

2. Materials and methods

The common TS objects have been selected as examples for IT security assessment and analysis. Main focus covers the availability characteristics and special IT-security metrics.

3. Problem description

To ensure the IT-security it is necessary to solve the problem of periodic assessment of the IT-security state, i.e, the degree of compliance with the requirements to ensure the confidentiality, availability and integrity. This problem involves a number of possible options for effective solutions, of which the most advanced, versatile and practical is the use of an ISMS and performing the IT-security assessment by assessing performance. This publication offers a methodical approach to evaluation of data protection in TS based on their availability. The implementation of this approach is based on ISMS use, where the set of requirements involves provision of security, including – availability [1].

Measuring the effectiveness of the ISMS parameters involves the use of IT-security metrics, allowing, among other things, to take into account the actual data availability in a particular TS with periodic evaluation (for example, when performing IT-security audits). Final evaluation metrics calculation of IT-security allow to estimate the overall level of ISMS effectiveness and, therefore, assess the current level of data protection in its current ISMS "configuration" («scope»). These assessments should serve as goals, first of all, for the top management, as sufficient evidence about the choice of optimal composition of tools (measures) to ensure information security, and, secondly, to offer experts, ensuring the safety of TS, the numerical metrics for assessing the effectiveness of individual implemented tools (measures), and IT-security functional subsystems, such as encryption.

4. The main threats for telecommunication systems

The main threats with the most significant consequences for TCS at this stage include:

- Steady increase in the number and size of leaks of sensitive information (commercial, technical, financial and personal data);
- Increasing the value of the consequences of critical infrastructure destruction (availability violation) at the technological level (SCADA, MES);
- Strengthening the ramifications of critical facilities blocking (violation of availability of information systems, data transmission systems).

Therefore, in order to ensure the security of the TS and when performing audits of the ISMS, the impact of appropriate means (measures) to ensure IT-Security «controls» should be measured reliably, the recommended list of which is given in Appendix A of the standard ISO 27001. For the purpose of this publication the following is considered as the most useful for the analysis of the availability of tools (measures) to ensure IT-Security in the TS (the list can be extended if necessary: A.12.7.1, A.13.1.2, A.15.1.3, A.17.1.2) and ISO 22301:2012: «Business continuity» (6.2) and «Business continuity strategy» (п. 8.3).

5. Results

For an effective assessment of IT-security in the TS an integrated approach must be adopted including the formation the set of criteria and metrics related to the ISMS. Obtaining numerical evaluation of data protection is achieved by using IT-security metrics and calculating the impact of an ISMS. It is known that a credible analysis of the ISMS only on the basis of "Statement of Applicability» is more difficult. When the assessment ISMS known difficulties associated with the formalization and analysis requirements for real IS and give statistics (for example, data on IT-security incidents). Should rely only on the facts, but the facts of security incidents and security breaches are rarely publicly known [2, 3].

The ISMS impact assessment is calculated based on presented approach using numerical metrics that take into account separately the IT-security events and IT-security incidents that directly affect the assessment of information availability in TCS. The results of calculation are presented in the charts (Fig. 1, 2).

The calculation of the impact of IT-Security events draw the following formula:

$$K_e = (1 - (S_{curr} / S_{max})) * 100\% \quad (1)$$

where:

K e - success rate for identifying IT-Security events;
S curr - identified the number of IT-Security events in the current configuration «scope»;
S max - maximum number of IT-Security events in the previous period.

The calculation of the impact of IT-Security incidents conduct by the formula:

$$K_i = (1 - (I_{curr} / I_{max})) * 100\% \quad (2)$$

where:

K i - the success rate for identifying IT-Security incidents;
I curr - identified the number of IT-Security incidents in the current configuration «scope»;
I max - maximum possible number of IT-Security incidents in the previous period.

In view of the Formula 1, 2 overall effectiveness of the ISMS is calculated:

$$K(ISMS) = K_e * \alpha + K_i * \beta \quad (3)$$

where:

K (ISMS) - overall effectiveness of the ISMS;
K e - success rate for identifying IT-Security events;
K i - the success rate for identifying IT-Security incidents;
α - weighting factor to determine the importance of identifying the K e;
β - weighting factor to determine the importance of identifying K i.

Example of calculating performance metrics ISMS (formula 3) is shown in Figure 1.

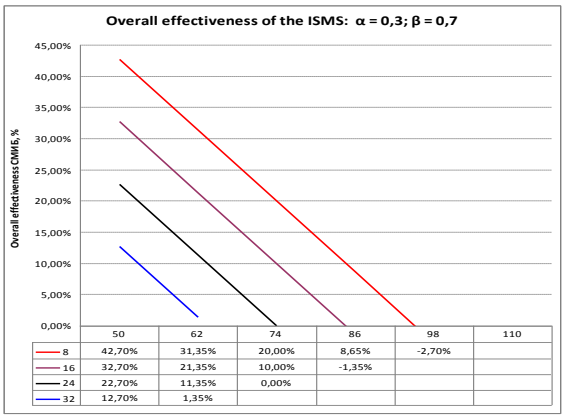


Figure 1: Overall effectiveness of the ISMS

Example of calculating the efficiency is shown in Figure 2.

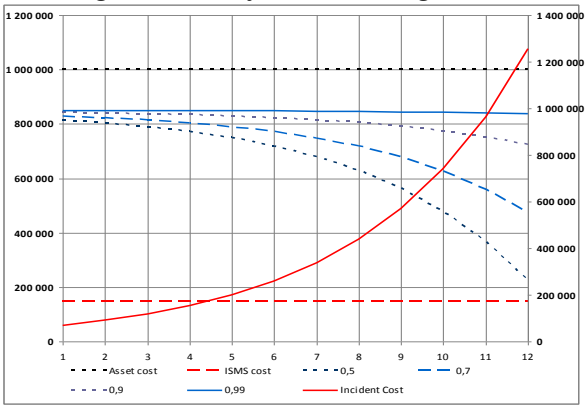


Figure 2: Example of calculating the efficiency of ISMS.

6. Conclusions

The proposed methodological approach to IT-security assessment and data protection in the TS in accordance with international standards (ISO 27001 series, ISO 20000 series and ISO 22301 series) allows to create technical and economic evaluation based on the indicators and the numerical values of their parameters:

- The level of data protection in TS (measured in the currency value of the protected assets of important organizations) implemented ISMS;
- The level of implementation of the requirements of the ISMS (including – availability) on the optimal set of metrics for TS given set of requirements and unique industry characteristics (for example, using weighting factors).

7. References

1. Information technology– Security techniques– Information security management systems – Requirements: ISO/IEC 27001:2013, International Organization for Standardization, 2013;
2. Livshits I.I., The concept of assessing the IT service providers information security level for industrial facilities // SPIIRAS Proceedings. 2014, vol 4, pp. 117 – 135;
3. Livshits I.I., Joint information security audit and availability of information systems problems solving on the basis of international standards ISO // Informatizatiy and Svyaz, 2013 vol. 6, pp. 48 – 52.

CONDITIONAL MONTE CARLO ESTIMATION OF HIGH ACTIVITY PERIOD DURATION IN GAUSSIAN QUEUES

O. Lukashenko[1,2], *E. Morozov*[1,2], *M. Pagano*[3]

[1] Petrozavodsk State University, Petrozavodsk, Russia

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

[3] University of Pisa, Pisa, Italy

lukashenko-oleg@mail.ru, emorozov@karelia.ru, m.pagano@iet.unipi.it

Abstract

Due to the self-similar nature of broadband traffic, the arrival rate can persist on relatively high values for a considerable amount of time. Such a behavior, closely related to the duration of busy periods, has a deep impact on queueing performance in terms of loss probability and distribution of losses. In the paper we consider the probability that the normalized cumulative workload grows at least as the length T of the considered interval in case of Gaussian input traffic. As T increases, the event becomes rare and standard Monte Carlo simulation would require a large number of generated sample paths to get an accurate estimate. To cope with this problem, we propose a variant of the well-known conditional Monte Carlo method, in which conditioning is expressed in terms of the bridge process. We derive the analytical expression of the estimator and verify its effectiveness through simulations.

Keywords: Gaussian processes, Conditional Monte Carlo, bridge process, persistence phenomena

1. Introduction

The self-similar nature of broadband traffic [1] has a deep impact in terms of network dimensioning and Quality of Service (QoS) issues [2]. Indeed, persistence phenomena (known in the literature as Noah effect) imply that the arrival rate can remain on relatively high values for a considerable amount of time. Such a behavior, closely related to the duration of busy periods in the underlying queueing system, negatively affects QoS performance in terms of loss probability and distribution of losses.

In traffic modelling, Gaussian processes have emerged as a flexible and powerful tool, able to take into account the long memory properties of real traffic, while keeping a relatively simple and elegant description. The best known model is Fractional Brownian Motion (FBM), originally proposed by Norros [3], but our method is more general and includes FBM as a special case.

In more detail, in this work we consider a centered Gaussian process with stationary increments $\{X_t,\, t \in \mathbb{R}_+\}$. Let us denote by $v_t := \mathbb{D}X_t$ the variance

of X_t; then the covariance function has the following expression:

$$\Gamma_{s,t} \;=\; \frac{1}{2}\left(v_t + v_s - v_{|t-s|}\right). \tag{1}$$

We are interested in estimation of the following probability

$$\pi(\mathbb{T}) \;:=\; \mathbb{P}\left(\forall t \in \mathbb{T}:\; X_t > t\right), \tag{2}$$

where $\mathbb{T} = [0, T] \subseteq \mathbb{R}_+$. Such probability is closely related to the duration of busy periods and plays an important role in the study of QoS indexes since it takes into account bursts of losses, see [4, 5] for more details.

The main contribution of this work is the application of a variant of the conditional Monte Carlo method for variance reduction, to estimate the target probability (2) when $T \to \infty$. Indeed, in this case the event $\{\forall t \in \mathbb{T}:\; X_t > t\}$ becomes rare and hence standard Monte Carlo requires unacceptable large number of generated sample paths.

2. BMC estimator

Bridge Monte Carlo (BMC) [6, 7] is a special case of conditional Monte Carlo method, particularly suitable for the estimation of the rare event probabilities in a queueing system with Gaussian input.

Let fix $\bar{t} \in \mathbb{T}$ and consider the new process:

$$Y_t = X_t - \psi_t X_{\bar{t}}, \tag{3}$$

where function ψ_t is expressed via covariance function Γ as

$$\psi_t := \frac{\Gamma_{t,\bar{t}}}{\Gamma_{\bar{t},\bar{t}}}\,.$$

The considered probability can be rewritten as

$$\pi = \mathbb{E}\left[\Phi\left(\frac{\overline{Y}}{\sqrt{\Gamma_{\bar{t},\,\bar{t}}}}\right)\right],$$

where Φ denotes the tail distribution of standard normal variable and

$$\overline{Y} := \sup_{t \in T} \frac{t - Y_t}{\psi_t}. \tag{4}$$

Given an i.i.d. sequence $\{\overline{Y}^{(i)},\; i = 1, ..., N\}$ distributed as $\overline{Y}$, the estimator of $\pi(\mathbb{T})$ is defined as follows:

$$\widehat{\pi}_N^{\mathrm{BMC}} \;=\; \frac{1}{N}\sum_{i=1}^{N} \Phi\left(\frac{\overline{Y}^{(i)}}{\sqrt{\Gamma_{\bar{t},\bar{t}}}}\right). \tag{5}$$

3. Simulations

We plan to estimate $\pi(\mathbb{T})$ based both on the estimator (5) and Crude Monte Carlo, and evaluate the efficiency comparing sample variances.

Although the approach presented in the paper only requires that v_t is an increasing function, in the simulation analysis we will consider the following important cases of Gaussian inputs:

1) Fractional Brownian Motion (FBM). In this case $v_t = t^{2H}$, with Hurst parameter $H \in (0,1)$ (in the teletraffic framework usually $H \in (0.5,1)$, corresponding to traffic processes with long range dependence). It has been shown in [8] that FBM arises as the scaled limit process when the cumulative workload is a superposition of on-off sources with mutually independent heavy-tailed on and/or off periods.
2) Sum of independent FBMs with $v_t = \sum_i t^{2H_i}$. The use of this model is also motivated by the fundamental result in [8] in case of heterogeneous on-off sources.
3) Integrated Ornstein-Uhlenbeck process (IOU) with $v_t = t + e^{-t} - 1$. IOU is the Gaussian counterpart of the well-known Anick-Mitra-Sondi fluid model.

Acknowledgments.
This work is supported by Russian Foundation for Basic research, projects 15–07–02341, 15–07–02354,15–07–02360, and also by the Program of strategic development of Petrozavodsk State University.

REFERENCES

1. Will E. Leland, Murad S. Taqqu, Walter Willinger, and Daniel V. Wilson. On the self-similar nature of ethernet traffic (extended version). *IEEE/ACM Trans. Netw.*, 2(1):1–15, February 1994.
2. Ashok Erramilli, Onuttom Narayan, and Walter Willinger. Experimental queueing analysis with long-range dependent packet traffic. *IEEE/ACM Trans. Netw.*, 4(2):209–223, April 1996.
3. I. Norros. On the use of fractional brownian motion in the theory of connectionless networks. *Selected Areas in Communications, IEEE Journal on*, 13(6):953–962, Aug 1995.
4. I. Norros. Busy periods for fractional brownian storage: a large deviation approach. *Adv. in Perf. Anal.*, 2:1–19, 1999.
5. M. Mandjes, I. Norros, and P. Glynn. On convergence to stationarity of fractional brownian storage. *Ann. Appl. Probab.*, 19:1385–1403, 2009.
6. S. Giordano, M. Gubinelli, and M. Pagano. Bridge Monte-Carlo: a novel approach to rare events of Gaussian processes. In *Proc. of the 5th St.Petersburg Workshop on Simulation*, pages 281–286, St. Petersburg, Russia, 2005.
7. S. Giordano, M. Gubinelli, and M. Pagano. Rare events of Gaussian processes: a performance comparison between Bridge Monte-Carlo and Importance Sampling. In *Next Generation Teletraffic and Wired/Wireless Advanced Networking*, pages 269–280, St. Petersburg, Russia, 2007.
8. M. S. Taqqu, W. Willinger, and R. Sherman. Proof of a fundamental result in self-similar traffic modeling. *Computer communication review*, 27:5–23, 1997.

THE SCIENTIFIC PROBLEM OF CORRECTING VIDEO DATA ERRORS, CORRUPTED DURING TRANSMISSION OVER THE INTERNET

A. Minyaev, S. Morkovin
Academy FSO, Orel, Russia

Abstract

The article discusses the problem statement processing malformed video data with reference to the conditions of functioning of systems of network monitoring. Formulated existing contradiction, research problem and research hypothesis.

НАУЧНАЯ ПРОБЛЕМА ИСПРАВЛЕНИЯ ОШИБОК ВИДЕОДАННЫХ, ИСКАЖЕННЫХ ПРИ ПЕРЕДАЧЕ ПО СЕТИ ИНТЕРНЕТ

А. Миняев, С. Морковин
Академия ФСО, Орел, Россия
sec@academ.msk.rsnet.ru

Аннотация

В статье рассматривается постановка задачи обработки искаженных видеоданных применительно к условиям функционирования систем мониторинга каналов связи. Сформулированы существующее противоречие, научная проблема и гипотеза исследования. Предложена постановка задачи на проведение исследования по разработке метода коррекции ошибок в искаженных видеоданных.

Ключевые слова: видеосжатие, видеокодирование, агрегация каналов, маскирование искажений, медиавещание.

1. Введение

В последнее время лидирующие позиции в общем объеме трафика телекоммуникационных систем занимают данные различных систем медиавещения. После прохождения процедур устранения избыточности и упаковки в контейнеры сетевых потоков эти данные имеют строго определенную структуру, существенно неустойчивую к воздействию ошибок в канале связи. Последствия ошибок могут иметь как локальный характер, так и приводить к полной невозможности корректного декодирования видеоданных. Существующие методы восстановления искаженных видеоданных ориентированы на узкий круг критических ситуаций и не могут быть в

полной мере применены для решения задач обработки видеоданных в системах мониторинга каналов связи. В частности, существующие методы не учитывают типичную для современных телекоммуникационных систем ситуацию трансляции данных по различным физическим каналам и невозможности наблюдения всей совокупности физических каналов в системах мониторинга. Это приводит к необходимости разработки и реализации в современных системах мониторинга каналов связи методов и алгоритмов обработки видеоданных в условиях их искажения. Целью настоящей статьи является постановка задачи на разработку метода коррекции ошибок в искаженных видеоданных.

1.1. Искажение видеоданных при передаче по сети Интернет. Трансляция видеоданных в современных системах связи описывается общей моделью системы передачи информации, одним из основных элементов которой является кодер источника сообщений (рис. 1).

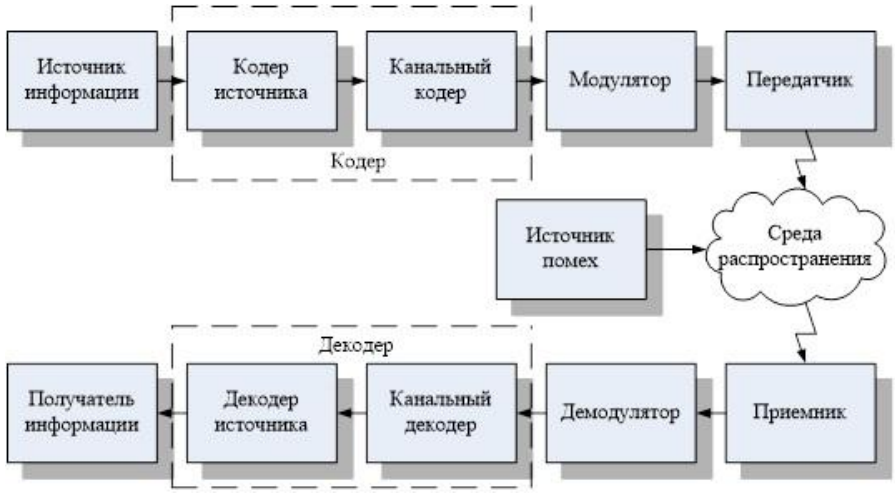


Рис. 1: Модель системы передачи информации.

Основная цель кодера заключается в сокращении избыточности видеоданных (т.е. сжатие видеоданных) на основе учета внутрикадровой и межкадровой корреляции подвижных изображений. Этапы сжатия видеоданных представлены на рисунке 2.

При этом, отправляемая в канал последовательность кадров подвижного изображения имеет строго определенную структуру (рис 3). Первым в последовательности следует так называемый опорный кадр, при сжатии которого используется информация только из пространственной области этого кадра. Кодирование всех последующих кадров (вплоть до следую-

Рис. 2: Этапы сжатия подвижных изображений.

щего опорного) производится относительно предыдущего и последующего кадров на основе учета разницы между предсказанными и действительными коэффициентами векторов движения объектов на кодируемой графической сцене.

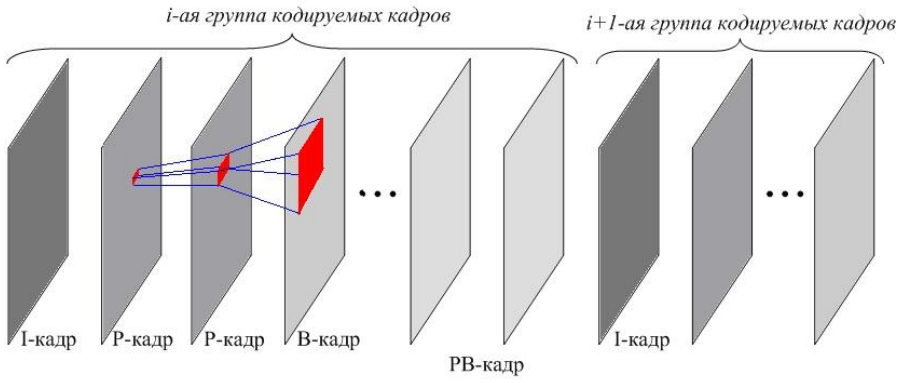


Рис. 3: Распространение единичной ошибки в зависимых кадрах.

Рассматриваемая структура последовательности позволяет достичь существенных коэффициентов сжатия видеоданных, но обладает неустойчивостью к ошибкам в канале связи. Искажение единичного элемента кадра

приводит к искажению всех зависимых от него элементов в последующих кадрах Ц т.е. к распространению ошибки (рис. 4).

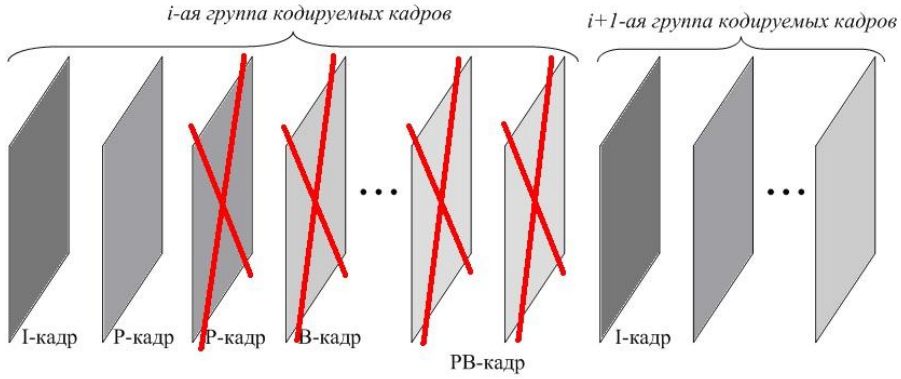


Рис. 4: Распространение ошибки, связанной с потерей промежуточного кадра.

Потеря целого кадра приводит к сбою в работе декодера и невозможности восстановления видеоданных вплоть до следующего опорного кадра (рис. 4).

Классификация известных подходов к восстановлению искаженных видеоданных представлена на рисунке 5. Группы традиционных для систем передачи данных методов и методов, относящихся к классу кодирование, устойчивое к ошибкам, основываются на специальных правилах организации канала связи и не могут быть применены при проектировании систем мониторинга.

Методы маскирования искажений восстанавливают только ограниченные области искаженного кадра и применяются при воздействии одиночных или локализованных групповых ошибок.

Однако, анализ условий функционирования систем мониторинга показывает, что современные системы с коммутацией пакетов зачастую строятся на принципах агрегации каналов, когда данные могут быть переданы по разным физическим каналам. При этом, в случае наблюдения в системе мониторинга не всей совокупности физических каналов, корректное декодирование видеоданных становится невозможным.

Таким образом, возникает противоречие между применением в системах передачи информации процедур сокращения избыточности (сжатия) видеоданных и отсутствием в системе мониторинга методов их корректной обработки. Разрешение противоречия требует проведения исследования по разработке теоретических основ обработки видеоданных применительно к условиям функционирования систем мониторинга.

Рис. 5: Методы борьбы с искажениями видеоданных.

2. Постановка задачи исследования

Научная проблема исследования: разработка теоретических основ обработки видеоданных применительно к системам мониторинга каналов связи.

Решение научной проблемы возможно путем выдвижения гипотезы о существовании взаимосвязей (корреляций) между параметрами и элементами неискаженных и искаженных (отсутствующих) кодовых слов, позволяющих выполнить восстановление синхронизации видеодекодера без участия следующего опорного кадра.

Формализованная постановка задачи исследования представлена на рисунке 6.

Исходные данные:

- исходные видеоданные (последовательность подвижных изображений), поступающие на вход системы передачи информации от источника видеоданных;
- переданная в канал связи последовательность кодовых слов, полученная в результате применения комплекса операторов устранения избыточности исходных видеоданных;
- поступившая на вход системы мониторинга неполная последовательность кодовых слов, подверженная действию шумов канала связи, и особенностям организации канала связи;
- особенности организации канала связи заключаются в использовании нескольких независимых физических каналов, образующих связанную логическую группу передачи последовательности кодовых слов.

Необходимо:

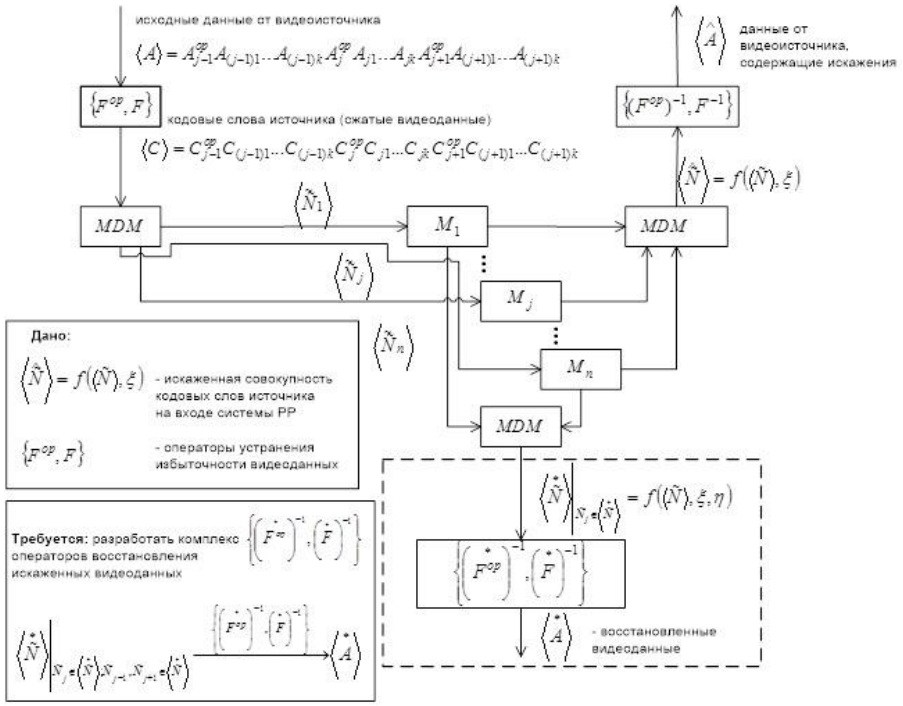


Рис. 6: Постановка задачи исследования.

- разработать комплекс операторов, обеспечивающих обработку и восстановление искаженных видеоданных в условиях функционирования систем мониторинга каналов связи.

Ограничения и допущения:

- характер воздействия канала связи ограничивается отсутствием искажений двух подряд идущих кодовых слов.

3. Заключение

В статье рассмотрена типичная для систем мониторинга каналов связи критическая ситуация наблюдения неполной группы физических каналов, приводящая к невозможности корректного декодирования видеоданных. Выявлено противоречие, сформулирована научная проблема и поставлена задача на проведение исследования, направленного на решение этой проблемы.

ЛИТЕРАТУРА

1. Ян Ричадсон. Видеокодирование H.264 и MPEG-4 - стандарты нового поколения. Техносфера, 2005. - с.47-112
2. ITU-T recommendation by standard of joint video specification (ITU-T Rec. H.264/ISO/IEC 14 496-10 AVC). Joint Video Team (JVT) of ISO/IEC MPEG and ITU-T VCEG, JVTG050, 2003.
3. M. Oliveira, B. Bowen. Fast digital image inpainting. Proc. International Conf. on Visualization, Imaging and Image Processing - Marbella, Spain, 2001. - Pp. 161-232.
4. Lin Liang, Ce Liu. Real-time texture synthesis by paych-based sampling. ACM Trans. Graph. - 2001. - July. Vol. 20, no. 3. - Pp. 67-152.
5. M. Bertalmio, L. Vese. Simultaneous structure and texture image inpainting. Image Processing, IEEE Transactions on. - 200. - Vol. 12, no. 8. - Pp. 882-889.

RECURRENCE CONDITIONS OF MODULATED MAP FLOW OF EVENTS UNDER ITS INCOMPLETE OBSERVABILITY

L. Nezhel'skaya
National Research Tomsk State University, Tomsk, Russia

Abstract

A modulated MAP flow of events with unextendable dead time is considered. It is one of the mathematical models for an input flow of events in digital integral servicing networks (ISDN). The observation conditions for this flow are such that each event generates a period of dead time during which other events from the flow are inaccessible for observation and do not extend the dead time period (unextendable dead time). An explicit form of a probability density of the interval duration between neighboring events in the observed flow is derived. Also an explicit form of a joint probability density of the duration of the two neighboring intervals is obtained. The recurrence conditions of the observed flow are found.

УСЛОВИЯ РЕКУРРЕНТНОСТИ МОДУЛИРОВАННОГО MAP-ПОТОКА СОБЫТИЙ ПРИ ЕГО НЕПОЛНОЙ НАБЛЮДАЕМОСТИ

Л.А. Нежельская

Национальный исследовательский Томский государственный университет, Томск, Россия
ludne@mail.ru

Аннотация

Изучается модулированный MAP-поток событий, являющийся одной из адекватных математических моделей информационных потоков событий в цифровых сетях интегрального обслуживания (ISDN), функционирующий в условиях непродлевающегося мертвого времени. Приводится явный вид плотности вероятностей длительности интервала между моментами наступления соседних событий потока, а также явный вид совместной плотности вероятностей длительности двух соседних интервалов. Приводятся условия рекуррентности потока.

Ключевые слова: модулированный MAP-поток событий, непродлевающееся мертвое время, инфинитезимальные характеристики, совместная плотность вероятностей, рекуррентность потока.

1. Введение

Математические модели теории массового обслуживания широко используются для описания реальных физических, технических и других процессов и систем. Интенсивное развитие компьютерной техники и информационных технологий во многом определило важную сферу приложений теории массового обслуживания – проектирование и создание информационно-вычислительных сетей, телекоммуникационных сетей и т.д. Все это послужило стимулом к созданию адекватных математических моделей реальных информационных потоков, функционирующих в цифровых сетях интегрального обслуживания, так называемых дважды стохастических потоков событий.

Данная статья является непосредственным развитием исследований, проведенных в [1, 2, 3, 4, 5, 6]. При исследовании потоков событий можно сформулировать два класса задач: 1) оценивание состояний потока событий [1, 2, 3]; 2) оценивание параметров потока [4, 5, 6].

Одним из искажающих факторов при оценке состояний и параметров потока выступает мертвое время регистрирующих приборов [7], которое порождается зарегистрированным событием. Другие события, которые наступают в течение периода мертвого времени, недоступны наблюдению (теряются). Этот период продолжается некоторое фиксированное время (непродлевающееся мертвое время). В качестве примера приведем CSMA|CD – протокол случайного множественного доступа с обнаружением конфликта, широко используемый в компьютерных сетях. В момент регистрации (обнаружения) конфликта на входе некоторого узла сети рассылается сигнал "заглушки"("пробки"); в течение времени рассылки сигнала "заглушки" заявки, поступившие в данный узел сети, получают отказ в обслуживании и направляются в источник повторных вызовов. Здесь время, в течение которого узел сети закрыт для обслуживания заявок, поступающих в него после обнаружения конфликта, можно трактовать как мертвое время прибора, регистрирующего конфликт в узле сети.

В данной статье находятся явные виды плотности вероятностей значений длительности интервала между моментами наступления соседних событий в модулированном MAP-потоке с непродлевающимся мертвым временем и совместной плотности вероятностей значений длительности двух соседних интервалов. Явный вид плотности вероятностей необходим для решения задачи оценивания параметров потока и длительности мертвого времени. Устанавливаются условия рекуррентности потока.

2. Постановка задачи

Рассматривается модулированный поток событий с интенсивностью, представляющей собой кусочно-постоянный стационарный случайный процесс $\lambda(t)$ с двумя состояниями: $\lambda(t) = \lambda_1$ либо $\lambda(t) = \lambda_2 \quad (\lambda_1 > \lambda_2 \geq 0)$. Длительность пребывания процесса $\lambda(t)$ в i-м состоянии, $i = 1, 2$, опреде-

ляется двумя случайными величинами: первая случайная величина распределена по экспоненциальному закону $F_i^{(1)}(t) = 1 - e^{-\alpha_i t}$, $i = 1, 2$; в момент окончания i-го состояния процесс $\lambda(t)$ переходит с вероятностью единица из i-го состояния в j-е, $i, j = 1, 2$ $(i \neq j)$; вторая случайная величина распределена по экспоненциальному закону $F_i^{(2)}(t) = 1 - e^{-\lambda_i t}$, $i = 1, 2$; в момент окончания i-го состояния процесс $\lambda(t)$ переходит с вероятностью $P_1(\lambda_j|\lambda_i)$ в j-е состояние $(i \neq j)$ с наступлением события либо с вероятностью $P_0(\lambda_j|\lambda_i)$ переходит в j-е состояние $(i \neq j)$ без наступления события, либо с вероятностью $P_1(\lambda_i|\lambda_i)$ переходит в i-е состояние с наступлением события $P_0(\lambda_j|\lambda_i) + P_1(\lambda_j|\lambda_i) + P_1(\lambda_i|\lambda_i) = 1$, $i, j = 1, 2$, $(i \neq j)$. Первая и вторая случайные величины являются независимыми друг от друга. В сделанных предположениях $\lambda(t)$ – марковский процесс.

Матрицы инфинитезимальных характеристик процесса $\lambda(t)$ при этом примут вид:

$$\mathbf{D}_1 = \begin{Vmatrix} -(\alpha_1 + \lambda_1) & \alpha_1 + \lambda_1 P_0(\lambda_2|\lambda_1) \\ \alpha_2 + \lambda_2 P_0(\lambda_1|\lambda_2) & -(\alpha_2 + \lambda_2) \end{Vmatrix},$$

$$\mathbf{D}_1 = \begin{Vmatrix} \lambda_1 P_1(\lambda_1|\lambda_1) & \lambda_1 P_1(\lambda_2|\lambda_1) \\ \lambda_2 P_1(\lambda_1|\lambda_2) & \lambda_2 P_1(\lambda_2|\lambda_2) \end{Vmatrix}.$$

Элементами матрицы $\mathbf{D}_1$ являются интенсивности перехода процесса $\lambda(t)$ из состояния в состояние с наступлением события. Диагональные элементы матрицы $\mathbf{D}_0$ – интенсивности выхода процесса $\lambda(t)$ из своих состояний, взятые с противоположным знаком. Недиагональные элементы матрицы $\mathbf{D}_0$ – интенсивности переходов процесса $\lambda(t)$ из состояния в состояние без наступления события. Следует заметить, что если $\alpha_i = 0$, $i = 1, 2$, имеет место обычный MAP-поток событий [2].

После каждого зарегистрированного в момент времени t_k события наступает время фиксированной длительности T (мертвое время), в течение которого другие события исходного модулированного MAP-потока недоступны наблюдению. По окончании мертвого времени первое наступившее событие снова создает период мертвого времени длительности T и т.д. Вариант возникающей ситуации показан на рис. 1, где $t_1, t_2, \ldots$ – моменты наступления событий в наблюдаемом потоке; 1 и 2 – состояния случайного процесса $\lambda(t)$; черными кружками обозначены события модулированного MAP-потока, недоступные наблюдению; штриховкой – длительность мертвого времени.

Процесс $\lambda(t)$ является принципиально ненаблюдаемым (скрытый марковский процесс); наблюдаемыми являются только временные моменты наступления событий потока $t_1, t_2, \ldots$. Рассматривается стационарный режим функционирования потока. В силу предпосылок в моменты времени $t_1, t_2, \ldots, t_k$ наступления событий потока последовательность $\{\lambda(t_k)\}$ представляет собой вложенную цепь Маркова, т.е. наблюдаемый поток обла-

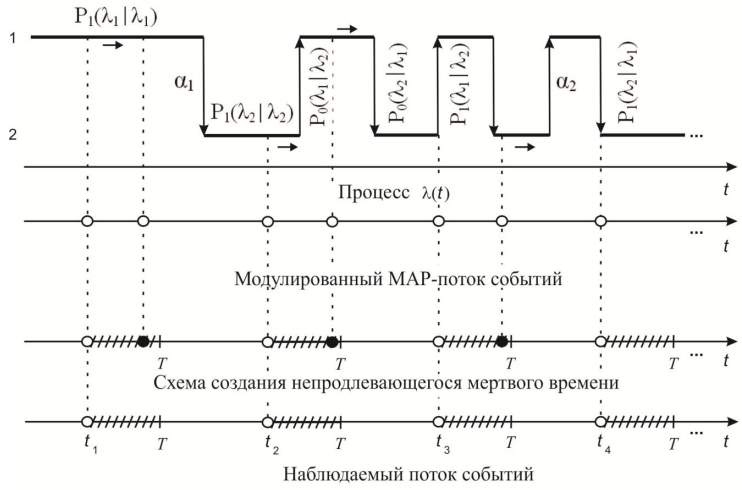


Рис. 1: Формирование наблюдаемого потока событий.

дает марковским свойством, если его эволюцию рассматривать с момента времени t_k – момента наступления события потока, $k = 1, 2, \ldots$.

Обозначим $\tau_k = t_{k+1} - t_k$, $k = 1, 2, \ldots$, – значение длительности k-го интервала между соседними событиями в наблюдаемом потоке. Так как рассматривается стационарный режим, то плотность вероятностей значений длительности k-го интервала $p_T(\tau_k) = p_T(\tau)$, $\tau \geq 0$, для любого k. В силу этого момент времени t_k наступления события без ограничения общности можно положить равным нулю или, что то же самое, момент наступления события есть $\tau = 0$. С другой стороны, поскольку зарегистрированное в момент времени $\tau = 0$ событие создает период мертвого времени длительности T, то $\tau = T + t$, где t – значение длительности интервала между моментом окончания мертвого времени $\tau = T$ и моментом времени наступления следующего события в наблюдаемом потоке ($t > 0$). Предполагается, что значение T точно известно.

Пусть (t_k, t_{k+1}), (t_{k+1}, t_{k+2}) – два смежных интервала, значения длительностей которых есть $\tau_k = t_{k+1} - t_k$ и $\tau_{k+1} = t_{k+2} - t_{k+1}$ соответственно; их расположение на временной оси, в силу стационарности потока, произвольно. Тогда, полагая $k = 1$, будем рассматривать два соседних интервала (t_1, t_2), (t_2, t_3) с соответствующими значениями длительностей $\tau_1 = t_2 - t_1$ и $\tau_2 = t_3 - t_2$, $\tau_1 \geq 0$, $\tau_2 \geq 0$. При этом $\tau_1 = 0$ соответствует моменту t_1 наступления события потока; $\tau_2 = 0$ соответствует моменту t_2 наступления

следующего события наблюдаемого потока. Соответствующая совместная плотность вероятностей есть $p_T(\tau_1, \tau_2)$, $\tau_1 \geq 0$, $\tau_2 \geq 0$.

Задача заключается в нахождении явного вида $p_T(\tau)$ и явного вида $p_T(\tau_1, \tau_2)$, а также в установлении условий рекуррентности модулированного MAP-потока событий, функционирующего в условиях непродлевающегося мертвого времени.

3. Плотность вероятностей $p_T(\tau)$

Функция плотности вероятностей длительности интервала между соседними событиями потока определяется формулой:

$$\begin{cases} p_T(\tau) = 0, \quad 0 \leq \tau < T, \\ p_T(\tau - T) = \sum\limits_{i=1}^{2} \pi_i(0|T) \sum\limits_{j=1}^{2} q_{ij}(T) \sum\limits_{k=1}^{2} \tilde{p}_{jk}(\tau - T), \quad \tau \geq T, \end{cases} \tag{1}$$

где $\tilde{p}_{jk}(\tau - T)$ – плотность вероятностей того, что процесс $\lambda(\tau)$ изменяет свое состояние с j на k без наступления события на интервале $(T, T+t)$ и с наступлением события в момент времени $\tau = T+t$, $j, k = 1, 2$; $q_{ij}(T)$ – переходная вероятность того, что процесс $\lambda(\tau) = \lambda_j$ в момент окончания мертвого времени $\tau = T$ при условии, что процесс $\lambda(0) = \lambda_i$ в момент времени $\tau = 0$, $i, j = 1, 2$; $\pi_i(0|T)$ – условная стационарная вероятность того, что процесс $\lambda(\tau)$ находится в состоянии i в момент времени $\tau = 0$ при условии, что в этот же момент времени $\tau = 0$ наступило событие потока и наступил период мертвого времени длительности T ($\pi_1(0|T) + \pi_2(0|T) = 1$).

Можно показать, что

$$\tilde{p}_{11}(t) = \lambda_1 P_1(\lambda_1|\lambda_1)\, p_{11}(t) + \lambda_2 P_1(\lambda_1|\lambda_2)\, p_{12}(t),$$

$$\tilde{p}_{12}(t) = \lambda_1 P_1(\lambda_2|\lambda_1)\, p_{11}(t) + \lambda_2 P_1(\lambda_2|\lambda_2)\, p_{12}(t),$$

$$\tilde{p}_{21}(t) = \lambda_1 P_1(\lambda_1|\lambda_1)\, p_{21}(t) + \lambda_2 P_1(\lambda_1|\lambda_2)\, p_{22}(t),$$

$$\tilde{p}_{22}(t) = \lambda_1 P_1(\lambda_2|\lambda_1)\, p_{21}(t) + \lambda_2 P_1(\lambda_2|\lambda_2)\, p_{22}(t),$$

$$p_{11}(t) = \frac{1}{z_2 - z_1}\left[(\lambda_2 + \alpha_2 - z_1)\, e^{-z_1 t} - (\lambda_2 + \alpha_2 - z_2)\, e^{-z_2 t}\right],$$

$$p_{12}(t) = \frac{\alpha_1 + \lambda_1 P_0(\lambda_2|\lambda_1)}{z_2 - z_1}\left[e^{-z_1 t} - e^{-z_2 t}\right],$$

$$p_{21}(t) = \frac{\alpha_2 + \lambda_2 P_0(\lambda_1|\lambda_2)}{z_2 - z_1}\left[e^{-z_1 t} - e^{-z_2 t}\right],$$

$$p_{22}(t) = \frac{1}{z_2 - z_1}\left[(\lambda_1 + \alpha_1 - z_1)\, e^{-z_1 t} - (\lambda_1 + \alpha_1 - z_2)\, e^{-z_2 t}\right],$$

$$z_1, z_2 = \frac{1}{2}\left[(\lambda_1+\lambda_2+\alpha_1+\alpha_2) \mp \right.$$
$$\left. \mp \sqrt{(\lambda_1-\lambda_2+\alpha_1-\alpha_2)^2 + 4\left(\alpha_1+\lambda_1 P_0\left(\lambda_2|\lambda_1\right)\right)\left(\alpha_2+\lambda_2 P_0\left(\lambda_1|\lambda_2\right)\right)}\right],$$
$$0 < z_1 < z_2; \tag{2}$$

$$q_{11}(T) = \pi_1 + \pi_2 e^{-aT}, \quad q_{12}(T) = \pi_2 - \pi_2 e^{-aT},$$
$$q_{21}(T) = \pi_1 - \pi_1 e^{-aT}, \quad q_{22}(T) = \pi_2 + \pi_1 e^{-aT},$$
$$a = \alpha_1 + \alpha_2 + \lambda_1\left[1 - P_1(\lambda_1|\lambda_1)\right] + \lambda_2\left[1 - P_1(\lambda_2|\lambda_2)\right],$$

$$\begin{array}{c} \pi_1 = \left(\alpha_2 + \lambda_2\left[1 - P_1(\lambda_2|\lambda_2)\right]\right)/a, \\ \pi_2 = \left(\alpha_1 + \lambda_1\left[1 - P_1(\lambda_1|\lambda_1)\right]\right)/a; \end{array} \tag{3}$$

$$\pi_1(0|T) = \frac{p_{21} + \pi_1(p_{11} - p_{21})\left[1 - e^{-aT}\right]}{p_{12} + p_{21} - (1 - p_{11} - p_{22})\left[1 - e^{-aT}\right]},$$
$$\pi_2(0|T) = \frac{p_{12} + \pi_2(p_{22} - p_{12})\left[1 - e^{-aT}\right]}{p_{12} + p_{21} - (1 - p_{11} - p_{22})\left[1 - e^{-aT}\right]},$$
$$p_{11} = \left[\lambda_1 P_1(\lambda_1|\lambda_1)(\lambda_2+\alpha_2) + \lambda_2 P_1(\lambda_1|\lambda_2)(\alpha_1 + \lambda_1 P_0(\lambda_2|\lambda_1))\right]/z_1 z_2,$$
$$p_{12} = \left[\lambda_1 P_1(\lambda_2|\lambda_1)(\lambda_2+\alpha_2) + \lambda_2 P_1(\lambda_2|\lambda_2)(\alpha_1 + \lambda_1 P_0(\lambda_2|\lambda_1))\right]/z_1 z_2,$$
$$p_{21} = \left[\lambda_2 P_1(\lambda_1|\lambda_2)(\lambda_1+\alpha_1) + \lambda_1 P_1(\lambda_1|\lambda_1)(\alpha_2 + \lambda_2 P_0(\lambda_1|\lambda_2))\right]/z_1 z_2,$$
$$p_{22} = \left[\lambda_2 P_1(\lambda_2|\lambda_2)(\lambda_1+\alpha_1) + \lambda_1 P_1(\lambda_2|\lambda_1)(\alpha_2 + \lambda_2 P_0(\lambda_1|\lambda_2))\right]/z_1 z_2,$$
$$\begin{array}{c} z_1 z_2 = \lambda_1\lambda_2\left[1 - P_0\left(\lambda_1|\lambda_2\right)P_0\left(\lambda_2|\lambda_1\right)\right] + \\ +\lambda_1\alpha_2\left[1 - P_0\left(\lambda_2|\lambda_1\right)\right] + \lambda_2\alpha_1\left[1 - P_0\left(\lambda_1|\lambda_2\right)\right]. \end{array} \tag{4}$$

С учетом (2)–(4) в результате достаточно трудоемких преобразований находим явный вид $p_T(\tau)$

$$p_T(\tau) = \begin{cases} 0, & 0 \le \tau < T, \\ \gamma(T) z_1 e^{-z_1(\tau - T)} + \left[1 - \gamma(T)\right] z_2 e^{-z_2(\tau - T)}, & \tau \ge T, \end{cases}$$

$$\gamma(T) = \frac{1}{z_2 - z_1}\left\{z_2 - \lambda_1\pi_1(T)\left[1 - P_0(\lambda_2|\lambda_1)\right] - \lambda_2\pi_2(T)\left[1 - P_0(\lambda_1|\lambda_2)\right]\right\},$$

$$\pi_1(T) = \pi_1 + \left[\pi_2 - \pi_2(0|T)\right]e^{-aT}, \quad \pi_2(T) = \pi_2 - \left[\pi_2 - \pi_2(0|T)\right]e^{-aT}, \tag{5}$$

z_1 и z_2 определены в (2), π_1 и π_2 определены в (3), $\pi_1(0|T)$ и $\pi_2(0|T)$ – в (4).

4. Совместная плотность вероятностей $p_T(\tau_1, \tau_2)$

В моменты времени $t_1, t_2, ..., t_k$ наступления событий последовательность $\{\lambda(t_k)\}$ представляет собой вложенную цепь Маркова, поэтому совместная плотность вероятностей значений длительности двух соседних интервалов $p_T(\tau_1, \tau_2)$ примет вид

$$p_T(\tau_1, \tau_2) = \sum_{i=1}^{2} \pi_i(0|T) \sum_{j=1}^{2} q_{ij}(T) \sum_{k=1}^{2} \tilde{p}_{jk}(\tau_1) \sum_{s=1}^{2} q_{ks}(T) \sum_{n=1}^{2} \tilde{p}_{sn}(\tau_2), \quad (6)$$

где $\tilde{p}_{jk}(\tau_1)$, $\tilde{p}_{sn}(\tau_2)$ – плотности вероятностей, соответствующие переходным вероятностям $p_{jk}(\tau_1)$, $p_{sn}(\tau_2)$ и вычисленные по формулам (2) при $t = \tau_1$ и $t = \tau_2$. Подставляя в (6) сначала $\tilde{p}_{jk}(\tau_1)$, $\tilde{p}_{sn}(\tau_2)$, затем $p_{jk}(\tau_1)$, $p_{sn}(\tau_2)$, определенные формулами (2) при $t = \tau_1$ и $t = \tau_2$, и $q_{ij}(T)$, $q_{ks}(T)$, определенные формулами (3), и, наконец, $\pi_i(0|T)$, $i = 1, 2$, определенные в (4), и, проделывая необходимые достаточно трудоемкие преобразования, находим

$$\begin{gathered} p_T(\tau_1, \tau_2) = 0; \quad 0 \le \tau_1 < T, \quad 0 \le \tau_2 < T; \\ p_T(\tau_1, \tau_2) = p_T(\tau_1)\, p_T(\tau_2) + \gamma(T)\,[1 - \gamma(T)]\, e^{-aT} \times \\ \times \frac{\lambda_1 \lambda_2}{z_1 z_2} \left[P_1(\lambda_1|\lambda_1) P_1(\lambda_2|\lambda_2) - P_1(\lambda_1|\lambda_2) P_1(\lambda_2|\lambda_1)\right] \times \\ \times \left(z_1 e^{-z_1(\tau_1 - T)} - z_2 e^{-z_2(\tau_1 - T)}\right) \left(z_1 e^{-z_1(\tau_2 - T)} - z_2 e^{-z_2(\tau_2 - T)}\right), \end{gathered} \quad (7)$$

где $\tau_1 \ge T$, $\tau_2 \ge T$, а $\gamma(T)$, $p_T(\tau_1)$, $p_T(\tau_2)$ определены в (5) для $\tau = \tau_1$ и $\tau = \tau_1$; z_1 и z_2 определены в (2), a – в (3).

5. Условия рекуррентности

Рассмотрим случаи, при которых модулированный MAP-поток событий при его неполной наблюдаемости становится рекуррентным. С учетом выражения (5) для $\gamma(T)$ и выражений (4) для $\pi_i(0|T)$, $i = 1, 2$, находим

$$\begin{gathered} \gamma(T)\,[1 - \gamma(T)] = \frac{\lambda_1 [1 - P_0(\lambda_2|\lambda_1)] - \lambda_2 [1 - P_0(\lambda_1|\lambda_2)]}{a^2 (z_2 - z_1)^2} z_1 z_2 \times \\ \times \{(\alpha_1 + \lambda_1 [1 - P_1(\lambda_1|\lambda_1)]) [\lambda_2 P_1(\lambda_1|\lambda_2)(\alpha_1 + \lambda_1) + \\ + \lambda_1 P_1(\lambda_1|\lambda_1)(\alpha_2 + \lambda_2 P_0(\lambda_1|\lambda_2))] - (\alpha_2 + \lambda_2 [1 - P_1(\lambda_2|\lambda_2)]) \times \\ \times [\lambda_1 P_1(\lambda_2|\lambda_1)(\alpha_2 + \lambda_2) + \lambda_2 P_1(\lambda_2|\lambda_2)(\alpha_1 + \lambda_1 P_0(\lambda_2|\lambda_1))]\} \times \\ \times \left\{z_1 z_2 - \lambda_1 \lambda_2 \left[P_1(\lambda_1|\lambda_1) P_1(\lambda_2|\lambda_2) - P_1(\lambda_1|\lambda_2) P_1(\lambda_2|\lambda_1)\right] e^{-aT}\right\}^{-2} \times \\ \times \left\{z_1 z_2 + e^{-aT} [a(z_1 + z_2) - 2 z_1 z_2] + \right. \\ \left. + e^{-2aT} [z_1 z_2 - a(\lambda_1 P_1(\lambda_1|\lambda_1) + \lambda_2 P_1(\lambda_2|\lambda_2))]\right\}. \end{gathered} \quad (8)$$

Анализируя выражение (8), замечаем, что если выполняется одно из условий:

1) $\lambda_1 [1 - P_0 (\lambda_2 |\lambda_1)] - \lambda_2 [1 - P_0 (\lambda_1 |\lambda_2)] = 0;$

2) $\{(\alpha_1 + \lambda_1 [1 - P_1 (\lambda_1|\lambda_1)]) [\lambda_2 P_1 (\lambda_1|\lambda_2) (\alpha_1 + \lambda_1) + \lambda_1 P_1 (\lambda_1|\lambda_1) (\alpha_2 + \lambda_2 P_0 (\lambda_1|\lambda_2))] - (\alpha_2 + \lambda_2 [1 - P_1 (\lambda_2|\lambda_2)]) \times [\lambda_1 P_1 (\lambda_2 |\lambda_1) (\alpha_2 + \lambda_2) + \lambda_2 P_1 (\lambda_2 |\lambda_2) (\alpha_1 + \lambda_1 P_0 (\lambda_2|\lambda_1))]\} = 0,$

то совместная плотность (7) факторизуется, то есть наблюдаемый поток становится рекуррентным.

Из выражения (7) следует третье условие факторизации плотности $p_T(\tau_1, \tau_2)$: $[P_1 (\lambda_1|\lambda_1) P_1 (\lambda_2|\lambda_2) - P_1 (\lambda_1|\lambda_2) P_1 (\lambda_2|\lambda_1)] = 0$, при выполнении которого наблюдаемый поток также становится рекуррентным.

6. Заключение

Полученные формулы для плотности вероятностей $p_T(\tau)$ и совместной плотности вероятностей $p_T(\tau_1, \tau_2)$ позволяют решить задачу оценивания неизвестных параметров и длительности мертвого времени в модулированном MAP-потоке событий с непродлевающимся мертвым временем методом моментов или методом максимального правдоподобия. В первом случае строится система уравнений моментов относительно неизвестных параметров, а во втором выписывается функция правдоподобия и решается задача ее максимизации относительно неизвестных параметров.

ЛИТЕРАТУРА

1. Nezhel'skaya L. A. Optimal State Estimation in Modulated Map Event Flows with Unextendable Dead Time // Communications in Computer and Information Sciences: proceeding of the 13th International Scientific Conference ITMM 2014 named after A.F. Terpugov "Information Technologies and Mathematical Modelling"(November 20-22, 2014). Cham Heidelberg New York Dordrecht London: Springer. 2014. P. 342-350.
2. Gortsev A. M., Nezhel'skaya L. A., Solov'ev A. A. Optimal State Estimation in MAP Event Flows with Unextendable Dead Time // Automation and Remote Control. 2012. V. 73(8). P. 1316-1326.
3. Gortsev A. M., Nezhel'skaya L. A., Shevchenko T. I. Estimation of the states of an MC-stream of events in the presence of measurement errors // Russian Physics Journal. 1993. V. 36(12). P. 1153-1167.
4. Gortsev A. M., Nezhel'skaya L. A. An asynchronous double stochastic flow with initiation of superfluous events // Discrete Mathematics and Applications. 2011. V. 21(3). P. 283-290.
5. Bushlanov I. V., Gortsev A. M., Nezhel'skaya L. A. Estimating parameters of the synchronous twofold-stochastic flow of events // Automation and Remote Control. 2008. V. 69(9). P. 1517-1533.
6. Gortsev A. M., Nezhel'skaya L. A. Estimation of the dead time period and intensities of the synchronous double stochastic event flow // Radiotekhnika. 2004. V. 10. P. 8-16.
7. Apanasovich V. V., Kolyada A. A., Chernyavskii A. F. Statistical Analysis of Random Flows in a Physical Experiment // Universitetskoe, Minsk. 1988.

STATISTICAL PROPERTIES OF PERFORMANCE MEASURES OF SIP SERVER MODEL WITH BATCH ARRIVALS[1]

Y. Orlov[1], *Yu. Gaidamaka*[2], *E. Zaripova*[2]

[1] Keldysh Institute of Applied Mathematics Russian Academy of Sciences, Moscow, Russia,

[2] Peoples' Friendship University of Russia, Moscow, Russia,

ov31509f@yandex.ru, ygaidamaka@sci.pfu.edu.ru, ezarip@sci.pfu.edu.ru

Abstract

In this paper an approach to analysis of dependence of Session Initiation Protocol server model with batch arrivals performance measures on batch size distribution is considered. Proposed approach employs non-parametric methods of statistical analysis. It is shown that there is statistical reliable dependence of performance measures, taken for signaling traffic analysis, on distance between distributions in definite norm. On the basis of proposed analysis elasticity coefficients were evaluated depending on distance between batch size distributions. This approach enables to get correction factors for estimation of these parameters in case distribution functions differ from uniform.

Keywords: Optimization, SIP mathematical model, distribution function, norm, performance measure, queuing system, parameter sensibility, sample, batch arrivals.

1. Introduction

Developing telecommunication services are successfully provided via IP-based Multimedia Subsystem (IMS), where Session Initiation Protocol (SIP) is the main signaling protocol. Signaling traffic load is auxiliary for communication nodes and is used for providing communications services to users. Signaling messages have so-called *life time*. When life time is over, information becomes not actual, signaling messages are retransmitted and overload the SIP server that processes them. Providing telecommunication services of required quality results in necessity of detailed research of communication system structure, statistic data analysis, implementation of processing algorithm for different types of signaling messages in order to increase number of successfully initiated sessions and decrease of service and sojourn time at a SIP server [1].

The following parameters were chosen for the investigation: average queue length and average waiting time. The first one enables estimation of SIP server's buffer capacity during the busy hour, the second when summed with

[1]The reported study was partially supported by RFBR, research project No. 15-07-03608.

service time enables estimation of SIP server sojourn time and is comparable to the message life time.

Process of signaling messages arrival and processing at a SIP server is performed via a single-server queue with batch arrivals and vacations. The vacation models the time interval when server processes messages that differ from signaling ones. In a model the batch arrival of customers corresponds to the simultaneous requests of the group of online subscribers.

In [2, 3, 4] there is an estimation of average value of queue length and average waiting time for general distribution of batch size. In [4] the analysis for the four batch size distribution functions has been done: Zipf, geometric, logarithmic and uniform distribution. In [4] a statistical dependence of investigating parameters on batch size distribution in case of Poisson process has been analyzed. In contrast with the papers where an investigation was performed only for geometrical distribution, in [4] it is recommended to use a uniform distribution of batch size that essentially simplifies formulas for calculations. In this paper an approach to performance measures analysis is proposed, sensitivity of model parameters to batch size distribution was estimated.

Distribution function variation was evaluated in [4] under different norms. Those with highest reliability of estimated statistical dependence were chosen. It turned out that analysis of numeric evaluation of SIP server performance parameters sensitivity to batch size distribution variation can be carried out with high reliability (0.99).

Since only sampling distribution functions of parameters are available for preliminary analysis, so we want to evaluate model parameters with statistical fluctuation of sampling distribution function when the distribution is not converge to any general population because of non-stationary behavior. We denote, that the batch size is a measured on practice, whereas average waiting time and average queue length depend on messages service time and processor vacation time. Distribution functions of the last two parameters are assumed to be known, because they depend on the SIP server hardware implementation. However, confidence interval for average queue length may not be got from corresponding empirical distribution, because we do not know this distribution function. We can evaluate these parameters by means of the simplified models only. In [4] the sensitivity analysis for one of models was carried out. Analysis was later applied to empirical distribution function of batch size.

2. SIP server model as a queue with batch arrivals and vacations

A mathematical model of signaling messages processing at a SIP server is investigated as a queuing system with batch arrivals and vacations. According to Basharin-Kendall notation this system is denoted as $M^{[X]} \mid G \mid 1 \mid \infty$. Let suppose that a batch of customers arrives according to Poisson process with rate λ. Customer service time is a random variable with a distribution function $B(t)$, where b_1 is the mean value and b_2 is the finite second moment.

If there are no customers in the queue, the server goes for a vacation. The vacation time is a random variable with a distribution function $V(t)$ with finite first and second moments v_1 and v_2.

Let $f(k)$ be probability that batch size is equal to $k \geq 1$. We denote the corresponding distribution function $F(k)$. For the queuing system $M^{[X]} \mid G \mid 1 \mid \infty$ with vacations we can find average queue length and average waiting time depending on offered load ρ which is estimated as follows:

$$\rho = \lambda b_1 l^{(1)}, \tag{1}$$

where $l^{(1)}$ is average batch size with distribution function $F(k)$.

In [2] a generating function was obtained for the queue length distribution for a single server model with vacations. The result is expressed as follows:

$$P(z) = \frac{1-\rho}{\lambda v_1} \cdot \frac{1-z}{1-L(z)} \cdot \frac{1-\phi(\lambda, z)}{\beta(\lambda, z) - z}, \tag{2}$$

where $L(z)$ is a generating function for batch size with the distribution function $F(k)$: $L(z) = \sum\limits_{k=0}^{\infty} f(k) z^k$. Other functions mentioned in (2) are expressed with the following equations: $v_1 = \int\limits_0^{\infty} t\, dV(t)$, $\phi(\lambda, z) = \int\limits_0^{\infty} e^{-\lambda t(1-L(z))}\, dV(t)$, and $\beta(\lambda, z) = \int\limits_0^{\infty} e^{-\lambda t(1-L(z))}\, dB(t)$. According to [4] average queue length depending on load ρ is obtained by the generating function (2) as follows:

$$N = \lim_{z \to 1} P'(z) = \frac{v_2}{2 v_1 b_1} \rho + \frac{l^{(2)} - l^{(1)}}{2 l^{(1)}} \frac{\rho}{1-\rho} + \frac{b_2}{2 b_1^2} \frac{\rho^2}{1-\rho}. \tag{3}$$

Taking into account the notation $v_s = \int\limits_0^{\infty} t^s dV(t)$, $b_s = \int\limits_0^{\infty} t^s dB(t)$ and $l^{(s)} = \sum\limits_k k^s f(k)$, $s = 1; 2$, the average waiting time is obtained as follows:

$$\tau = \frac{N b_1}{\rho} = \frac{v_2}{2 v_1} + \frac{l^{(2)} - l^{(1)}}{2 l^{(1)}} \frac{b_1}{1-\rho} + \frac{b_2}{2 b_1} \frac{\rho}{1-\rho}. \tag{4}$$

3. Sensitivity of the model parameters to probability variation

Let find out how to change the values N and τ in formulas (3)-(4), if probability variation of $f(k)$ is low. Let a new distribution can be expressed as $\tilde{f}(k) = f(k) + \varepsilon(k) f(k)$, moreover, under the same norming the equality $\sum\limits_k \varepsilon(k) f(k) = 0$ is fulfilled. Let introduce the variable $E = \sup\limits_k |\varepsilon(k)|$. Then we get the following estimation of distance between distributions in L1 norm:

$$\varepsilon = \left\| f - \tilde{f} \right\|_{L1} = \sum_k \left| f(k) - \tilde{f}(k) \right| = \sum_k |\varepsilon(k)|\, f(k) \leq E. \tag{5}$$

Let use (5) to get estimation of some differentiable function variation depending on average batch size in case $E << 1$:

$$\left|\delta l^{(1)}\right| = \left|\sum_k k\left(f(k) - \tilde{f}(k)\right)\right| < \sum_k k\left|f(k) - \tilde{f}(k)\right| = \sum_k k\left|\varepsilon(k)\right| f(k) \le El^{(1)};$$

$$\left|\delta u\left(l^{(1)}\right)\right| = \left|\frac{\partial u\left(l^{(1)}\right)}{\partial\left(l^{(1)}\right)}\right| \cdot \left|\delta l^{(1)}\right| \le Eu\left|\frac{\partial \ln u}{\partial \ln l}\right|_{l=l^{(1)}} + o(E). \tag{6}$$

Logarithmic derivative of function with respect to parameter's logarithm is called function sensitivity to parameter variation. Then, from (6) comes that variation of some function from the mean value, obtained due to distribution variation, at linear approximation on E does not exceed this function multiplied by the supremum of distribution density variation and by modulus of specified sensitivity. In our case estimations of average queue length (3) and average waiting time (4), that are linear for E, are expressed as follows:

$$|\delta N| \le \frac{E\rho}{2} \cdot \left|\frac{v_2}{v_1 b_1} + \frac{l^{(2)} - l^{(1)}}{l^{(1)}} \frac{b_1}{(1-\rho)^2} + \frac{b_2}{b_1^2}(2-\rho)\frac{\rho}{(1-\rho)^2}\right|;$$

$$|\delta\tau| \le \frac{E\rho}{2(1-\rho)^2} \cdot \left|\frac{l^{(2)} - l^{(1)}}{l^{(1)}} b_1 + \frac{b_2}{b_1}\right|. \tag{7}$$

However, theoretical estimations (7) do not possess adequate accuracy as they appear to be too excessive. Despite this fact they cannot be improved within of functions to be chosen for empirical distribution functions. Unimprovability comes out of existence of a variation when $\varepsilon(k) = \{0; \pm E\}$. Inadequacy of accuracy comes out of condition (5): as far as $\varepsilon \le E$, then for heavily nonuniform distributions the variation norm in the form of $\sup_k |\varepsilon(k)|$ is too crude estimate, since for such kind of distributions the distance between distributions can be significantly less than E. That is why in [4] an analysis of sensitivity of these parameters based on the numerical results for various functions $f(k)$ was carried out. It was found out that four types of norms – L1 and C for $F(k)$, L1 for $f(k)$ and similar to them the forth norm that is a supplement for total area S to the unity of two densities – determine rather exactly the variation of values $N$and τunder variation of densities.

Let us denote variation of $f(k)$ by δf, which is denominated in norm C for distribution function. We denote by $\delta N(\rho)$ variation of average queue length and by $\delta\tau(\rho)$ variation of average waiting time for a given value of load ρ. Data analysis showed that under variation of uniform distribution, which determined on the interval from 1 to the maximum batch size (in our case maximum is equal to 8), there is a relationship between $\delta N(\rho)$ and δf, also $\delta\tau(\rho)$ and δf with

determination 0.99:

$$\left|\frac{\delta N(\rho)}{N(\rho)}\right| = 0,2\,(1-\ln\rho)\,(\delta f)^{0,27};$$

$$\left|\frac{\delta \tau(\rho)}{\tau(\rho)}\right| = \frac{0,106+0,041\ln\rho}{\rho}\,(\delta f)^{0,27}; \tag{8}$$

$$\rho \in [0,1;\ 0,9].$$

Practical application of the described method is following. Let take a uniform distribution on the interval $1 \le k \le M$ as a basic predicted distribution of $f(k)$. For this distribution function equation (3) of average queue length is converted to

$$N_0 = \frac{v_2}{2v_1 b_1}\rho + \frac{M-1}{3}\frac{\rho}{1-\rho} + \frac{b_2}{2b_1^2}\frac{\rho^2}{1-\rho}, \tag{9}$$

that is considered as null approximation.

Let consider that empirical distribution, being investigated, is stationary, $F(k)$ is its distribution function. The distance between this and uniform distribution may be calculated according to the following equation:

$$\delta f = \sup_k |F(k) - k/M|\,. \tag{10}$$

Substituting the result of (10) in (8) we get estimation of average queue length that corresponds to the following empirical distribution:

$$N/N_0 \approx 1 + 0,2\,(1-\ln\rho)\,(\delta f)^{0,27},\ \rho \in [0,1;\ 0,9] \tag{11}$$

Estimation for average waiting time variation is expressed in the same way.

Equation (11) is computationally much simpler than calculation of the generating function in accordance with (2), where $L(z)$ is calculated through empirical distribution of $F(k)$. That is, firstly, rather difficult and, secondly, leads to calculation errors that may exceed approximation inaccuracy for (11).

For example, let consider that an empirical distribution of $f_n(k)$ is taken from a general population $f(k)$ that has geometric distribution with a parameter q on the interval $1 \le k \le M$. That means $f(k) = \frac{1-q}{1-q^M}q^{k-1}$.

In [4] estimations were performed for the case of $q = 0,67;\quad M = 8$. According to (11) we get $\delta f = 0,35$ for these parameters. Consequently, we get that average queue length for this distribution will exceed same value for a uniform distribution approximately by $0,14\,(1-\ln\rho)\,N_0(\rho)$.

4. Conclusion

This paper presents the approach to estimation of the performance measures for SIP server model with batch arrivals and vacations depending on the batch

size distribution. Investigation of this particular dependence was motivated by the fact that the batch size distribution is not known as a general population and, moreover, cannot be recognized as far as empirical evaluations of this population are non-stationary. That is why approximate evaluation methods, that are not associated with a specified functional class of mentioned distributions, are of great significance and actuality. Therefore, the method that considers coefficients of the model parameters sensitivity to adjustment of the distance between distribution functions seems to be efficient among nonparametric techniques. This method may be used for non-stationery distributions when non-stationary behavior is interpreted as definite variation of some basic distribution (for example, uniform). This approach enables to circumvent technical difficulty coming from absence of convergence theorem both for probability and the norm for random variables being investigated.

Proposed approach to evaluation of performance measures of a Session Initiation Protocol server model and given analysis of parameters sensitivity leads to recommendations for engineers to use simple formulas for preliminary evaluation of presence signaling messages service.

Acknowledgment

The reported study was partially supported by RFBR, research project No. 15-07-03608.

We thank Professor Konstantin Samouylov from Peoples' Friendship University of Russia for comments that greatly improved the paper.

REFERENCES

1. Abhayawardhana V.S., Babbage R. A traffic model for the IP multimedia subsystem (IMS), Proceedings of 65^{th} vehicular technology conference. 2007. P. 783-787.
2. Samouylov K.E., Sopin E.S. On Analysis of M[X]|G|1|r Queuing System. Bulletin of Peoples' Friendship University of Russia. Series Mathematics. Information Sciences. Physics. 2011. No. 1. P. 91-97.
3. Gaidamaka Yu., Pechinkin A., Razumchik R., Samouylov K., Sopin E. Analysis of M|G|1|R queue with batch arrivals and two hysteretic overload control policies. International Journal of Applied Mathematics and Computer Science. 2014. Vol. 24. No. 3. P. 519–534.
4. Gaidamaka Yu.V., Zaripova E.R., Orlov Y. N. Analysis of the impact the batch size distribution on parameters of the SIP-server queueing model with batch arrivals. KIAM Preprint No. 27. Moscow, 2015. P. 1-16 / URL: http://library.keldysh.ru/preprint.asp?id=2015-27

THE RESEARCH OF THE QUEUEING SYSTEM $MAP|M|\infty$ WITH HETEROGENEOUS SERVERS BY THE METHOD OF ASYMPTOTIC ANALYSIS PROVIDED EXTREMELY RARE STATE CHANGES OF MAP ARRIVALS

E. Pankratova[1]
[1] National Research Tomsk State University, Tomsk, Russia

The research of the queueing system with MAP arrivals, n types of customers, infinite number of servers and exponential service time is proposed. There are expressions for the characteristic function of the number of busy servers for different types of customers in the system $MAP|M|\infty$ under the asymptotic condition of extremely rare changes of states of MAP arrivals.

Keywords: queueing system, Markovian arrival process, different types of customers, asymptotic analysis provided extremely rare changes of states of MAP arrivals

ИССЛЕДОВАНИЕ СИСТЕМЫ МАССОВОГО ОБСЛУЖИВАНИЯ MAP|M|∞ С РАЗНОТИПНЫМ ОБСЛУЖИВАНИЕМ МЕТОДОМ АСИМПТОТИЧЕСКОГО АНАЛИЗА В УСЛОВИИ ПРЕДЕЛЬНО РЕДКИХ ИЗМЕНЕНИЙ СОСТОЯНИЙ ВХОДЯЩЕГО MAP-ПОТОКА

Е. Панкратова[1]
[1] Национальный исследовательский Томский государственный университет, Томск, Россия,
pankate@sibmail.com

Аннотация

Предлагается исследование системы массового обслуживания с входящим MAP-потоком разнотипных заявок, неограниченным числом обслуживающих приборов и экспоненциальным временем обслуживания. Получены выражения для характеристических функций числа занятых приборов каждого типа заявок в системе MAP|M|∞ с разнотипным обслуживанием в асимптотическом условии предельно редких изменений состояний входящего MAP-потока.

Ключевые слова: MAP-поток разнотипных заявок, метод асимптотического анализа в условии ПРИС.

1. Введение

Исследования систем массового обслуживания(СМО) с неограниченным числом приборов можно встретить в статьях П.П. Бочарова, А.В. Печинкина[1], А.А. Назарова, P. Abaev, R. Razumchik [2], B. D'Auria [3], D. Baum и L. Breuer [4, 5], E.A. van Doorn и A.A Jagers [6], N.G. Duffield [7], C. Fricker и M. R. Jaïbi [8] и многих других. В то же время многочисленные исследования реальных потоков в различных предметных областях, в частности, телекоммуникационных потоков, а также потоков в экономических системах, позволили сделать вывод о существенной неадекватности классических моделей потоков (пуассоновских и рекуррентных) реальным данным.Поэтому разработка новых математических моделей СМО, а именно систем с марковизируемыми входящими потоками и различными вариантами обслуживания, в том числе с использованием в рамках одной системы разных типов обслуживающих приборов (имеющих различные интенсивности обслуживания), является актуальной задачей. Для исследования таких СМО, как правило, применяются численные методы либо имитационное моделирование. Альтернативным подходом является применение метода асимптотического анализа для исследования таких систем [9, 10].

2. Постановка задачи

Рассмотрим СМО MAP|M|∞, на вход которой поступает MAP-поток разнотипных заявок, заданный набором неотрицательных чисел λ_k, матрицей инфинитезимальных характеристик Q для управляющей цепи Маркова $k(t)$ и вероятностями $d_{\nu k}$. В момент наступления события в рассматриваемом потоке в систему поступает только одна заявка, которая определяется как заявка i-ого типа ($i = 1, \ldots, n$), и выполняется ее обслуживание в течение случайного времени, распределенного по экспоненциальному закону с параметром $\mu_i\,(i = 1, \ldots, n)$, соответсвующим типу заявки. Поставим задачу исследования n-мерного случайного процесса $\{i_1(t), \ldots, i_n(t)\}$, характеризующего число занятых приборов i-ого типа в момент времени t.

Для марковизируемого MAP-потока $(n+1)$-мерный случайный процесс $\{k(t), i_1(t), \ldots, i_n(t)\}$ является цепью Маркова.

Рассмотрим совместное распределение вероятностей $P(k, i_1, \ldots, i_n, t) = P\{k(t) = k, i_1(t) = i_1, \ldots, i_n(t) = i_n\}$ и запишем для

него систему дифференциальных уравнений Колмогорова

$$\frac{\partial P(k,i_1,\ldots,i_n,t)}{\partial t} = \left(-\lambda_k - \sum_{l=1}^{n} i_l\mu_l\right) P(k,i_1,\ldots,i_n,t) +$$
$$+P(k,i_1-1,\ldots,i_n,t)\lambda_k p_1 + \cdots + P(k,i_1,\ldots,i_n-1,t)\lambda_k p_n +$$
$$+P(k,i_1+1,\ldots,i_n,t)(i_1+1)\mu_1 + \cdots + P(k,i_1,\ldots,i_n+1,t)(i_n+1)\mu_n +$$
$$+\sum_{\nu=1}^{K}\{(1-d_{\nu k})P(\nu,i_1,\ldots,i_n,t) + d_{\nu k}\,(p_1 P(\nu,i_1-1,\ldots,i_n,t) + \cdots +$$
$$+p_n P(\nu,i_1,i_2,\ldots,i_n-1,t))\}\,q_{\nu k},\ k=1,\ldots,K. \qquad (1)$$

Начальные условия определим в виде

$$P(k,i_1,\ldots,i_n,0) = P(k,0,\ldots,0,t) = R(k), \qquad (2)$$

где $R(k)$ —стационарное распределение вероятностей цепи Маркова $k(t)$.

Решение системы (1) будем искать при стационарном функционировании рассматриваемой системы.

Для частичных характеристических функций вида

$$H(k,u_1,\ldots,u_n) = \sum_{i_1=0}^{\infty}\ldots\sum_{i_n=0}^{\infty} e^{ju_1i_1}\times\cdots\times e^{ju_ni_n}P(k,i_1,\ldots,i_n),$$
$$k=1\ldots,K,\ j=\sqrt{-1},$$

перепишем систему дифференциальных уравнений Колмогорова (1) в виде

$$\sum_{l=1}^{n}\mu_l j(e^{-ju_l}-1)\frac{\partial H(k,u_1,\ldots,u_n)}{\partial u_l} = \lambda_k\left(\sum_{l=1}^{n}p_l e^{ju_l}-1\right)H(k,u_1,\ldots,u_n) +$$
$$+\sum_{\nu=1}^{K}H(\nu,u_1,\ldots,u_n)\left[1+\left(\sum_{l=1}^{n}p_l e^{ju_l}-1\right)d_{\nu k}\right]q_{\nu k}. \qquad (3)$$

Начальные условия (2) примут вид

$$H(k,0,\ldots,0) = R(k),\ k=1,\ldots,K. \qquad (4)$$

3. Асимптотический анализ в условии предельно редких изменений состояний входящего MAP-потока

Значения инфинитезимальных характеристик q_{kk} определяют времена пребывания MAP-потока в k-х состояниях $k=1,\ldots,K$.

Пусть ε —некоторый малый положительный параметр.

Условием предельно редких изменений состояний входящего MAP-потока будем называть равенства

$$q_{\nu k}^{(1)} = \varepsilon q_{\nu k}, \nu = 1, \ldots, K,\ k = 1, \ldots, K, \tag{5}$$

определяющие достаточно малые значения инфинитезимальных характеристик, что влечет достаточно редкие изменения состояний потока.

Учитывая (5), систему (3) перепишем в виде

$$\sum_{l=1}^{n} \mu_l j(e^{-ju_l} - 1) \frac{\partial H(k, u_1, \ldots, u_n)}{\partial u_l} = \lambda_k \left(\sum_{l=1}^{n} p_l e^{ju_l} - 1 \right) H(k, u_1, \ldots, u_n) +$$

$$+ \varepsilon \sum_{\nu=1}^{K} H(\nu, u_1, \ldots, u_n) \left[1 + \left(\sum_{l=1}^{n} p_l e^{ju_l} - 1 \right) d_{\nu k} \right] q_{\nu k}. \tag{6}$$

Решение этой системы $H(k, u_1, \ldots, u_n)$, зависящее от параметра ε и удовлетворяющее начальному условию (4), обозначим

$$H(k, u_1, \ldots, u_n) = F(k, u_1, \ldots, u_n, \varepsilon), \tag{7}$$

$$F(k, 0, \ldots, 0, \varepsilon) = R(k),\ k = 1, \ldots, K.$$

Сформулируем и докажем следующее утверждение.

Теорема 1. *Предельное значение функции $F(k, u_1, \ldots, u_n, \varepsilon)$ при $\varepsilon \to 0$ имеет вид*

$$\lim_{\varepsilon \to 0} F(k, u_1, \ldots, u_n, \varepsilon) = F_1(k, u_1, \ldots, u_n) =$$

$$= R(k) \exp \left\{ \lambda_k \sum_{l=1}^{n} \frac{p_l}{\mu_l} \left(e^{ju_l} - 1 \right) \right\},\ k = 1, \ldots, K. \tag{8}$$

Доказательство. Выполним в системе (6) предельный переход при $\varepsilon \to 0$ и для $F_1(k, u_1, \ldots, u_n)$ получим систему уравнений в частных производных

$$\sum_{l=1}^{n} j\mu_l (e^{-ju_l} - 1) \frac{\partial F_1(k, u_1, \ldots, u_n)}{\partial u_l} = \lambda_k \left(\sum_{l=1}^{n} p_l e^{ju_l} - 1 \right) F_1(k, u_1, \ldots, u_n),$$

$$k = 1, \ldots, K. \tag{9}$$

Общее решение (9)имеет вид

$$C_1 \exp \left\{ \lambda_k \sum_{l=1}^{n} \frac{p_l}{\mu_l} \left(e^{ju_l} - 1 \right) \right\} = F_1(k, u_1, \ldots, u_n).$$

Для того чтобы найти C_1, воспользуемся начальными условиями (7) и получим выражение для $F_1(k, u_1, \ldots, u_n)$

$$F_1(k, u_1, \ldots, u_n) = R(k) \exp\left\{\lambda_k \sum_{l=1}^{n} \frac{p_l}{\mu_l}\left(e^{ju_l} - 1\right)\right\},\ k = 1, \ldots, K,$$

которое совпадает с выражением (8). ■

Таким образом, для частичных характеристических функций можно записать асимптотическое равенство

$$H(k, u_1, \ldots, u_n) = F(k, u_1, \ldots, u_n, \varepsilon) \approx F_1(k, u_1, \ldots, u_n).$$

Тогда для характеристической функции процесса $\{i_1(t), i_2(t), ..., i_n(t)\}$ получаем выражение

$$h_1(u_1, ..., u_n) = Me^{j\sum\limits_{l=1}^{n} u_l i_l(t)} = \sum_{k=1}^{K} H(k, u_1, ..., u_n) =$$
$$= \sum_{k=1}^{K} R(k) \exp\left\{\lambda_k \sum_{l=1}^{n} \frac{p_l}{\mu_l}\left(e^{ju_l} - 1\right)\right\}.$$

Определение 1. Функцию $h_1(u_1, ..., .u_n)$ будем называть асимптотической характеристической функцией первого порядка числа занятых приборов разного типа в системе $MAP|M|\infty$ в условии предельно редких изменений состояний входящего MAP-потока.

Определение 2. Функцию

$$h(u) = M\{e^{ju\sum_{l=1}^{n} i_l}\} = \sum_{k=1}^{K} R(k) \exp\left\{\lambda_k \sum_{l=1}^{n} \frac{p_l}{\mu_l}\right\}$$

будем называть асимптотической характеристической функцией первого порядка общего числа занятых приборов системы $MAP|M|\infty$ с разнотипными заявками в условии предельно редких изменений состояний входящего MAP-потока.

Определение 3. Функции

$$h_1{}^{(l)}(u_l) = Me^{ju_l i_l(t)} = h_1(0, ..., u_l, ..., 0) = \tag{10}$$
$$= \sum_{k=1}^{K} R(k) \exp\left\{\lambda_k \frac{p_l}{\mu_l}\left(e^{ju_l} - 1\right)\right\},\ l = \overline{1, n},$$

будем называть асимптотической характеристической функцией первого порядка числа занятых приборов каждого типа в системе $MAP|M|\infty$ в условии предельно редких изменений состояний входящего MAP-потока.

4. Численный анализ

Для сравнения асимптотических и допредельных значений основных характеристик исследуемой системы найдем асимптотическое значение математического ожидания числа занятых приборов

$$fm_l^{as}(k) = \frac{\partial H(k,u_1,\ldots,u_n)}{j\partial u_l}\bigg|_{\substack{u_s=0,\\ s=\overline{1,n}}},\ k=1,\ldots,K,\ l=1,\ldots,n.$$

Откуда

$$fm_l^{as} = \frac{p_l}{\mu_l}\mathbf{r}\mathbf{\Lambda}\mathbf{e},\ l=1,\ldots,n, \tag{11}$$

где $\mathbf{e}$ —единичный вектор-столбец, $\mathbf{r}$ —вектор-строка $[R(1),R(2),\ldots,R(K)]$.

Допредельное значение математического ожидания числа занятых приборов в системе MAP|M|∞ с разнотипным обслуживанием определяется следующим образом [10]:

$$fm_l = \frac{p_l}{\mu_l}\mathbf{r}\mathbf{B}\mathbf{e},\ l=1,\ldots,n. \tag{12}$$

Асимптотическое выражение для моментов второго порядка числа занятых приборов l-ого типа может быть вычисленно как

$$sm_l^{as}(k) = \frac{\partial^2 H(k,u_1,\ldots,u_n)}{j^2\partial u_l^2}\bigg|_{\substack{u_s=0,\\ s=\overline{1,n}}},\ k=1,\ldots,K,\ l=1,\ldots,n.$$

Тогда имеем следующее выражение для вторых моментов

$$sm_l^{as} = \frac{p_l}{\mu_l}\sum_{k=1}^{K} R(k)\lambda_k(\lambda_k\frac{p_l}{\mu_l}+1),\ l=1,\ldots,n. \tag{13}$$

Допредельное значение второго моента имеет вид [10]

$$sm_l = \mathbf{r}\mathbf{B}p_l\{\mathbf{I}+[\mu_l\mathbf{I}-\mathbf{Q}]^{-1}[\mu_l\mathbf{I}+2\mathbf{B}p_l]\}\{2\mu_l\mathbf{I}-\mathbf{Q}\}^{-1}\mathbf{e},\ l=1,\ldots,n. \tag{14}$$

Аналогично, запишем выражение для асимптотического и допредельного значений корреляционного моментов

$$\begin{gathered} cm_{lg}^{as} = \frac{p_l p_g}{\mu_l\mu_g}\sum_{k=1}^{K}\lambda_k^2 R(k), \\ cm_{lg} = \frac{1}{\mu_l+\mu_g}(p_l\mathbf{fm_g}+p_g\mathbf{fm_l})\mathbf{B}\mathbf{e}, \\ l=1,\ldots,n,\ g=1,\ldots,n,\ l\neq g. \end{gathered} \tag{15}$$

Для определения области применимости данного метода рассмотрим пример при $n = 3$, $K = 3$.

Определим MAP-поток следующими параметрами:

$$\mathbf{Q} = \begin{pmatrix} -11 & 5 & 6 \\ 0,5 & -1 & 0,5 \\ 2,5 & 2,5 & -5 \end{pmatrix}, \qquad \mathbf{\Lambda} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 12 & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

$$\mathbf{D} = \begin{pmatrix} 0 & 0,2 & 0,5 \\ 0,1 & 0 & 0 \\ 0,9 & 0,6 & 0 \end{pmatrix},$$

$p_1 = 0,2,\ p_2 = 0,5,\ p_3 = 0,3$ —вероятности поступления в систему заявок соответствующего типа.

Определим относительную погрешность вида

$$\Delta_l = \left| \frac{sm_l - sm_l^{as}}{sm_l} \right|, l = \{1, 2, 3\}.$$

ε	sm_1	sm_2	sm_3
0,1	106,765	2411,406	30,429
0,01	116,213	2651,467	31,992
0,001	119,449	2786,991	32,293
Асимптотика	$sm_1^{as} = 119,895$	$sm_2^{as} = 2808,947$	$sm_3^{as} = 32,329$

Таблица 1: Сравнение допредельных и асимптотических значений характеристик системы.

ε	0,1	0,01	0,001
Δ_1	0,123	0,032	0,004
Δ_2	0,165	0,059	0,008
Δ_3	0,062	0,011	0,001

Таблица 2: Сравнение значений относительной погрешности.

Из полученных результатов видно, что с уменьшением ε относительная погрешность стремится к нулю.

5. Заключение

В работе проведено исследование СМО MAP|M|∞ методом асимптотического анализа в условии предельно редких изменений состояний входящего MAP-потока. Получены выражения для асимптотических характеристических функций первого порядка общего числа занятых приборов и числа приборов каждого типа в рассматриваемой системе. Проведено численное сравнение допредельных и асимптотических значений основных характеристик системы.

ЛИТЕРАТУРА

1. Pechinkin A. V., Sokolov I. A., Chaplygin V. V. Stationary characteristics of multi-line queuing system with simultaneous failures of devices // Computer Science and Applications. 2007. V. 1(2). P. 28-38 (in Russian)
2. Abaev P. On Mean Return Time in Queueing System with Constant Service Time and Bi-level Hysteric Policy // Modern Probabilistic Methods for Analysis and Optimization of Information and Telecommunication Networks. Proc. of the International Conference, Minsk, 2013
3. Auri, B. D. $M|M|\infty$ queues in semi-Markovian random environment // Queueing Systems. 2008. V.58(3). P.221-237
4. Baum D. The infinite server queue with Markov additive arrivals in space. Probabilistic analysis of rare events // Proc. of the International Conference. Riga, 1999. P. 136-142
5. Baum D., Breuer L.: The Inhomogeneous $BMAP|G|\infty$ queue // Proc. of the 11th GI/ITG Conference on measuring, modelling and evaluation of computer and communication systems (MMB 2001). Aachen, P. 209-223
6. Doorn E. A., Jagers A. A. Note on the $GI|GI|\infty$ system with identical service and interarrival-time distributions // Journal of queueing systems. 2004. V. 47. P. 45-52
7. Duffield N. G. Queueing at large resources driven by long-tailed $M|G|\infty$-modulated processes // Queueing Systems. 1998. V. 28(1-3). P. 245-266
8. Fricker C., Jaïbi M. R. On the fluid limit of the $M|G|\infty$ queue // Queueing Systems. 2007. V. 56(3-4). P. 255-265
9. Parulekar M., Makowski A. M. Tail probabilities for $M|G|\infty$ input processes(I): Preliminary asymptotics // Queueing Systems. 1997. V.27(3-4). P. 271-296
10. Pankratova E., Moiseeva S. Queueing System MAP|M|∞ with n Types of Customers // 13th Intern.Scien.Conf. ITMM 2014 named after A.F.Terpugov. Anzhero-Sudzhensk, 2014. P. 356-366.

JOINT STATIONARY DISTRIBUTION OF QUEUES IN MULTI-SERVER RESEQUENCING QUEUE

A. Pechinkin, *R. Razumchik*[1]

[1] Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, Moscow, Russia
rrazumchik@ipiran.ru

Abstract

Consideration is given to a classical problem which is important in data networks where data is divided into packets that are sent using different routes. At the destination packets should be re-ordered so as to recover the original data. We model data network as N-server queueing system ($N > 3$) with single infinite capacity buffer, where each server represents a route. Customers arrive at the system according to Poisson flow and receive service which is exponentially distributed with the same parameter. The order of customers upon arrival has to be preserved upon departure. Customers which violated the order are kept in resequencing buffer which also has infinite capacity. It is assumed that resequencing buffer may be partitioned into n, $1 \leq n \leq N-1$, queues, depending on the number of busy servers, and i-th queue contains customers which have to wait for i service completions before they can leave the system. Equations for computation of joint stationary distribution of number of customers in buffer, servers and each queue in resequencing buffer, which admit recursive solution, are being obtained. Numerical example is given.

Keywords: resequencing, queueing system, joint distribution, infinite capacity

1. Introduction

It is well-known that performance of multiserver simultaneous processing systems can suffer from resequencing issue, i.e. when the order of arriving customers (jobs, items etc.) is violated due to disordering which may be introduced by service process or other external/internal factors. As consequence of disordering some customers have to wait for other customers before they leaving the system. Various analytical methods and models have been proposed to study the impacts of resequencing. Survey on the resequencing problem that covers period up to 1997 and review of queueing theoretic methods and early models for the modelling and analysis of parallel and distributed systems with resequencing can be found in [6] and [7]. Queueing-theoretic approach to resequencing problem implies that the system under consideration is represented as interconnected queueing systems/networks where disordering of customers takes place. The system is followed with resequencing buffer where the order of customers is recovered. In [13] convenient suggestion was made to group existing papers on

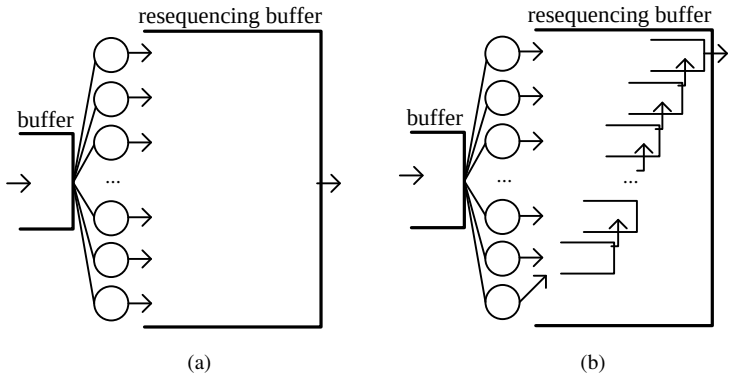


Figure 1: Sketch of multiserver resequencing queue

resequencing into two categories: papers that characterize the disordering process using single queueing system with several servers sharing a single queue (see, e.g. [1]) and papers where disordering is modelled by a queueing system with several parallel servers and queues, and each server has its own dedicated queue (see, e.g. [19]). Paper [13] contains the survey of papers belonging to these two categories. Up to now various problems setting have been considered and solved including distribution of number of packets in resequencing buffer and in system under different assumptions about arrival and service process, distribution of the resequencing delay, and optimal allocation of customers (see, e.g. [2], [4]–[10], [12], [14], [17]–[19]).

In this paper we study the system belonging to the first category, specifically $M|M|N|\infty$ resequencing queue with $N > 3$. The sketch of the system can be seen in Fig.1 (a). There is one Poisson incoming flow, one queue of infinite capacity and several homogeneous servers which serve customers during exponentially distributed times according with FCFS or LCFS or Random choice discipline. Customers which violated arrival order are kept in resequencing buffer (RB) of infinite capacity before they can leave the system. As it was noticed in [16], in such $M|M|N|\infty$ resequencing queue with $N > 2$ servers, resequencing buffer can be thought of either as a single queue where all customers which violated arrival order reside together (see Fig.1 (a)) or as collection of several separate interconnected queues (see Fig. 1 (b)). In the latter case i-th queue contains those customers which have to wait for i service completions before they can leave the system. Notice that number of service completions needed by a customer in RB to leave the system cannot be greater than $(N-1)$.

The motivation to introduce such partition of resequencing buffer into separate queues can be best described by example. Consider $M|M|4|\infty$ resequencing queue. Without loss of generality we suppose that customers upon entering the system obtain sequential number; the sequence starts from 1 and coincides with the row of natural numbers. If we started observing the system when it was initially empty then at some time instant we may see the system in the state as depicted in Fig.2 (a) or Fig.2 (b).

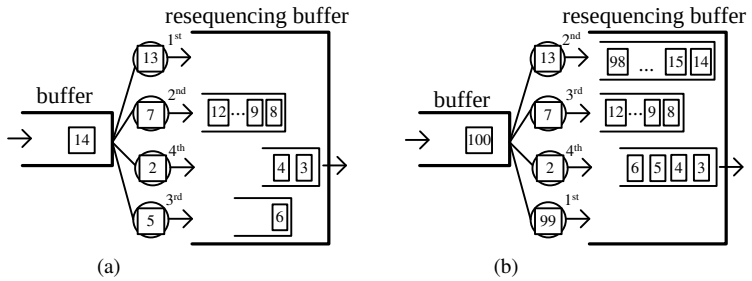


Figure 2: Examples of resequencing system's contents at two different time instants

Each square represents one customer and number in the square is its sequential number. Though it is easy to calculate the total number of customers in RB, there is one interesting observation. For example, in Fig.2 (b) if the next service completion is of customer 2, then customers 2, 3, 4, 5, 6 will leave the system (in batch), whereas other customers in RB will keep on waiting. But if the next service completion is of customer 7, then 7, 8, 9, 10, 11, 12 will not leave the system, but will wait together with customers 3, 4, 5, 6 until the service of customer 2 is completed. Due to such evolution of RB contents it is natural to group customers in RB in different queues. They are marked out in Fig.2 (a) or Fig.2 (b). Partitioning of RB into several queues gives more detailed view of its dynamics and leads to number of interesting questions: what is the joint stationary distribution of all queues in the system; are there any dependencies between queues' sizes; what happens with queues in RB if N grows without bound; what the influence of service distribution on queues' sizes in RB. In this paper we focus on first two questions.

In the system with $N \geq 2$ servers, if all of them are busy, then the resequencing buffer can be partitioned into $(N-1)$ queues (see Fig. 2(a) and Fig.2(b) as example for $N = 4$). If the number of busy servers is less than N, then number of queues in resequencing buffer is equal to the number of busy servers. The analysis of joint stationary distribution of number of customers even in simple cases with Poisson flow and homogeneous exponential servers turns out to be a challenging task. In [16] for $M/M/3/\infty$ queue followed with infinite resequencing buffer one obtains expressions for joint stationary distribution of number of customers in buffer and servers, and number of customers in *each* of two queues in resequencing buffer both in explicit form and in terms of generating functions. In [3] for $M/M/N/\infty$ queue followed with infinite resequencing buffer there was obtained algorithm for recursive computation joint stationary distribution of number of customers in buffer and servers, and *sum of number of customers* in two, three, ..., and $(N-1)$ queues in resequencing buffer. In this paper consideration is given again to $M/M/N/\infty$ queue followed with infinite resequencing buffer. The main contribution is the methodology for computation of joint stationary distribution of number of customers in buffer and servers, and number of customers in *each* queue in resequencing buffer. It is shown that joint stationary

distribution can be computed recursively. We note that joint distribution can be also obtained in terms of generating functions but this analysis is left for further research.

Next section is devoted to the description of the system. In Section 3 the system of equations which admits recursive computation of joint stationary distribution of number of customers in buffer and servers, and number of customers in *each* queue in resequencing buffer is given. Short numerical example concludes the paper.

2. System description and notation

Consider a queueing system with $3 < N < \infty$ servers, infinite capacity buffer, incoming Poisson flow of customers of intensity λ, exponential distribution of service in each server with parameter μ and resequencing buffer of infinite capacity. Customers upon entering the system obtain sequential number and join buffer. Without loss of generality we suppose that the sequence starts from 1 and coincides with the row of natural numbers, i.e. customer upon entering the empty system receives number 1, the next one — number 2 and so on and so forth. Customers leave the system strictly in order of their arrival. Thus after customer's arrival it enters server (if there are any idle) or remains in the buffer for some time and then receives service on one of the servers. If at the moment of its service completion there are no customers in the system or all other customers present at that moment in the buffer and in all other servers have greater sequential numbers it leaves the system. Otherwise it occupies a place in the RB. Customer from RB leaves it if and only if its sequential number is less than sequential numbers of all other customers present in system. It may be noticed that customers may leave RB in groups. For example, in Fig.2(a) if customer with sequential number 2 is the next to finish service then it leaves the system at one together with customer with sequential number 3 and 4.

In order to correctly define partitioning of RB into several queues we use the following approach. Assume there are n, $n = \overline{1, N}$, busy servers in the system. Each time any server becomes free or busy we label customers in servers according to the order in which they occupied servers. Let us refer to customer which was the last to enter server as "1^{st} level" customer. Customer which entered server right before "1^{st} level" customer is referred to as "2^{nd} level" level customer. The "3^{rd} level" customer is the one which entered server before "2^{nd} level" customer. Proceeding in similar manner customer which was the first (among n) to enter server is referred to as "n^{th} level" the customer. Customers which reside in resequencing buffer form $(n-1)$ separate queues in the following way. Customers which entered RB between "1^{st} level" and "2^{nd} level" customer form queue #1; customers which entered RB between "2^{nd} level" and "3^{rd} level" customer form queue #2 and so on. Customers which entered RB between "$(n-1)$ level" and "n^{th} level" customer form queue #$(n-1)$. Example of such partitioning of resequencing buffer into separate queues in case when $N = 4$ is given in Fig. 2a and Fig. 2b. Let us denote by p_n, $n \geq 0$, stationary probabilities of the fact, that there are n customer in buffer and servers (customers in RB are not taken into account). One can notice that p_n, $n \geq 0$, are determined by the same equations as in simple $M/M/N/\infty$ queue. For stationary probabilities of the considered system

with resequencing to exist it is necessary and sufficient that necessary and sufficient condition for existence of probabilities p_n is fulfilled i.e. $\rho/N < 1$ must hold.

Let us denote by $p_{n;i_1,\ldots,i_m},\ m = \overline{1, N-1},\ i_1, \ldots, i_m \geq 0$, stationary probability of the fact that there are $n \geq N$ customers in buffer and servers, and in RB there are i_1 customers in queue #1, i_2 customers in queue #2, $\ldots$, i_m customers in queue #m. If number of busy servers is $n < N$, then we denote by $p_{n;i_1,\ldots,i_m},\ m = \overline{1, n},\ i_1, \ldots, i_m \geq 0$, stationary probability of same fact. The only difference between cases $n \geq N$ and $n < N$ is that in the former case number of queues in RB may vary from 1 to $(N-1)$ and in the latter case it may vary only from 1 to n.

3. System of equilibrium equations

Due to the space limitation we are unable to give detailed description of how balance equations for all $p_{n;i_1,\ldots,i_m}$ can be written out. We give the idea instead. In order to write out equilibrium equations one has to consider step-by-step different partitions of the state space and use rate-in-rate-out principle (local balance). Notice that if one sums up say probability $p_{N;i_1,\ldots,i_{N-1}}$ over all possible values of $i_2, \ldots, i_{N-1}$ (the result of summation we will denote by $p_{N;i_1}$, i.e. we simply omit the indexes over which the summation is performed), then one obtains the probability of the fact that there are N customers in buffer and servers, and queue #1 contains i_1 customers (irrespective of the number of customer in queue #2, #3 ... #$(N-1)$ in RB). For probabilities of such state sets it is possible to analyse one-step transitions and write out balance equations that eventually lead to determination of the whole joint distribution.

For probabilities $p_{n;i_1},\ n \geq N,\ i_1 \geq 0$, the following equations hold

$$p_{n;0}(\lambda + N\mu) = p_{n-1;0}\lambda + p_{n+1}(N-1)\mu,\ \ n \geq N, \tag{1}$$

$$p_{n;i_1}(\lambda + N\mu) = p_{n-1;i_1}\lambda + p_{n+1;i_1-1}\mu,\ \ n \geq N,\ \ i_1 \geq 1. \tag{2}$$

Probabilities $p_{N-1;i_1},\ i_1 \geq 0$, are governed by the following equations

$$p_{N-1;0}[\lambda + (N-1)\mu] = p_{N-2}\lambda + p_N(N-1)\mu, \tag{3}$$

$$p_{N-1;i_1}[\lambda + (N-1)\mu] = p_{N;i_1-1}\mu,\ \ i_1 \geq 1. \tag{4}$$

Probabilities $p_{n;i_1},\ n = \overline{1, N-2},\ i_1 \geq 0$, are given by

$$p_{n;0}(\lambda + n\mu) = p_{n-1}\lambda + p_{n+1;0}n\mu,\ \ n = \overline{1, N-2}, \tag{5}$$

$$p_{n;i_1}(\lambda + n\mu) = p_{n+1;i_1}n\mu + \sum_{j=0}^{i_1-1} p_{n+1;i_1-j-1,j}\mu,\ \ n = \overline{1, N-2},\ \ i_1 \geq 1. \tag{6}$$

For probabilities $p_{n;i_1,\ldots,i_m},\ m = \overline{2, N-1},\ n \geq m,\ i_1, \ldots, i_{N-1} \geq 0$, one can write out the system of balance equations in general form. It holds

$$p_{n;0,i_2,\ldots,i_m}(\lambda + N\mu) = p_{n-1;0,i_2,\ldots,i_m}\lambda + p_{n+1;i_2,\ldots,i_m}(N-m)\mu +$$
$$+ \sum_{j=0}^{i_2-1} p_{n+1;j,i_2-j-1,i_3,\ldots,i_m}\mu + \ldots + \sum_{j=0}^{i_m-1} p_{n+1;i_2,\ldots,i_{m-1},j,i_m-j-1}\mu,\ n \geq N,\ \ i_2, \ldots, i_m \geq 0, \tag{7}$$

$$p_{n;i_1,\ldots,i_m}(\lambda + N\mu) = p_{n-1;i_1,\ldots,i_m}\lambda + p_{n+1;i_1-1,i_2,\ldots,i_m}\mu,\ n \geq N,\ \ i_1 \geq 1,\ \ i_2,\ldots,i_m \geq 0, \quad (8)$$

$$p_{N-1;0,i_2,\ldots,i_m}[\lambda + (N-1)\mu] = p_{N-2;i_2,\ldots,i_m}\lambda + p_{N;i_2,\ldots,i_m}(N-m)\mu +$$
$$+ \sum_{j=0}^{i_2-1} p_{N;j,i_2-j-1,i_3,\ldots,i_m}\mu + \ldots + \sum_{j=0}^{i_m-1} p_{N;i_2,\ldots,i_{m-1},j,i_m-j-1}\mu,\ \ i_2,\ldots,i_m \geq 0, \quad (9)$$

$$p_{N-1;i_1,\ldots,i_m}[\lambda + (N-1)\mu] = p_{N;i_1-1,i_2,\ldots,i_m}\mu,\ \ i_1 \geq 1,\ \ i_2,\ldots,i_m \geq 0, \quad (10)$$

$$p_{n;0,i_2,\ldots,i_m}(\lambda+n\mu) = p_{n-1;i_2,\ldots,i_m}\lambda + p_{n+1;0,i_2,\ldots,i_m}(n-m+1)\mu + \sum_{j=0}^{i_2-1} p_{n+1;0,j,i_2-j-1,i_3,\ldots,i_m}\mu +$$
$$+ \ldots + \sum_{j=0}^{i_m-1} p_{n+1;0,i_2,\ldots,i_{m-1},j,i_m-j-1}\mu,\ m \neq N-1,\ n = \overline{m, N-2},\ i_2,\ldots,i_m \geq 0, \quad (11)$$

$$p_{n;i_1,\ldots,i_m}(\lambda + n\mu) = p_{n+1;i_1,\ldots,i_m}(n-m+1)\mu + \sum_{j=0}^{i_1-1} p_{n+1;j,i_1-j-1,i_2,\ldots,i_m}\mu + \ldots +$$
$$+ \sum_{j=0}^{i_m-1} p_{n+1;i_1,\ldots,i_{m-1},j,i_m-j-1,}\mu,\ m \neq N-1, n = \overline{m, N-2},\ \ i_1 \geq 1,\ \ i_2,\ldots,i_m \geq 0. \quad (12)$$

In equations (7)–(12) for the sake of brevity we used agreement that $\sum_{i=0}^{-1} a_i = 0$. For fixed value of N system (1)–(12) can be solved recursively. Due to space limitation we were unable to give here the procedure and just show the final computational results in the next section.

4. Numerical examples

We have carried out extensive numerical experiments computing joint stationary distribution of number of customers in buffer and servers, and number of customers in queues in RB, as well as several important performance characteristics. We note that the complexity of the solution algorithm grows very fast as number of servers increases and already when $N = 10$ computation of the whole joint stationary distribution becomes very slow. Below we give several results of numerical experiments. It is assumed that number of servers is $N = 4$ and service intensity is $\mu = 1$. For graphic presentation of joint stationary distribution $p_{n;i_1,i_2,i_3}$ one can choose only two coordinates at once. In Fig. 3 for system load $\rho/N = 0.75$ one can see the behaviour of joint stationary distribution of number of customers in buffer and in queue #1 in RB (Fig. 3(a)), joint stationary distribution of number of customers in buffer and in queue #2 in RB (Fig. 3(b)), joint stationary distribution of number of customers in buffer and in queue #3 in RB (Fig. 3(c)).

In Fig.4 one can see the behaviour of the same distribution for system's load $\rho/N = 0.9$.

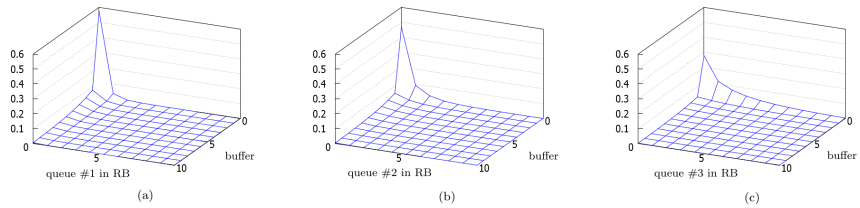


Figure 3: System's load $\rho/N = 0.75$. Joint stationary distribution of number of customers in buffer and in (a) queue #1 in RB, (b) queue #2 in RB, (c) queue #3 in RB

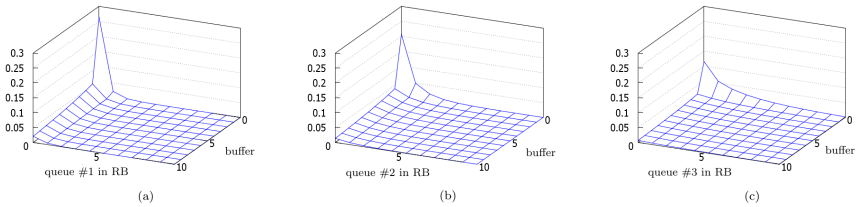


Figure 4: System's load $\rho/N = 0.9$. Joint stationary distribution of number of customers in buffer and in (a) queue #1 in RB, (b) queue #2 in RB, (c) queue #3 in RB

Acknowledgments.

This work was supported by the Russian Foundation for Basic Research (grant 15-07-03007, 13-07-00223).

REFERENCES

1. Agrawal S., Ramaswamy R. Analysis of the resequencing delay for M/M/m systems // Proceedings of the ACM SIGMETRICS conference on Measurement and modeling of computer systems, 1987. Pp. 27–35.
2. Baccelli F., Gelenbe E., Plateau B. An end-to-end approach to the sequencing problem // Rapports de Recherche, INRIA, 1981.
3. Pechinkin A. V., Caraccio I., Razumchik R. V. On joint stationary distribution in exponential multiserver reordering queue // Proceedings of the 12th International Conference on Numerical Analysis and Applied Mathematics, 2015. Vol. 1648. pp. 250003.
4. Chowdhury S. Distribution of the total delay of packets in virtual circuits // Proceedings of the Tenth Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM 91), 1991. Vol. 2. Pp. 911–918.
5. Chakravarthy S., Chukova S., Dimitrov B. Analysis of MAP/M/2/K queueing model with infinite resequencing buffer // Journal of Performance Evaluation, 1998. Vol. 31. Issue 3-4. Pp. 211–228.

6. Boxma O., Koole G., Liu Z. Queueing-theoretic solution methods for models of parallel and distributed systems // Performance Evaluation of Parallel and Distributed Systems Solution Methods. CWI Tract 105 & 106, 1994. pp. 1–24.
7. Dimitrov B., Green Jr. D., Rykov V., Stanchev P. On performance evaluation and optimization problems in queues with resequencing // Advances in Stochastic Modelling, edt. J. Artalejo, A. Krishnamoorty. Notable publications, 2002. Pp. 55–72.
8. De Nicola C., Pechinkin A., Razumchik R. Stationary Characteristics of Homogenous Geo/Geo/2 Queue with Resequencing in Discrete Time // Proceedings of the 27th European Conference on Modelling and Simulation, 2013. Pp. 594–600.
9. Gogate R., Panwar S. Assigning customers to two parallel servers with resequencing // IEEE Communications Letters, 1999. Vol. 3. No. 4. Pp. 119–121.
10. Huisman T., Boucherie R.J. The sojourn time distribution in an infinite server resequencing queue with dependent interarrival and service times // Journal of Applied Probability, 2002. Vol. 39. No. 3. Pp. 590–603.
11. Huisman T., Boucherie R.J. Running times on railway sections with heterogeneous train traffic // Transportation Research Part B: Methodological, 2001. Vol. 35. No. 3.
12. Jain M., Sharma G.C. No passing multiserver queue with additional heterogeneous servers and inter-dependent rates // 5-th Canadian Conference in Applied Statistics. 20-th conference of the Forum for Interdisciplinary Mathematics - Interdisciplinary Mathematical Statistical Techniques, 2011.
13. Leung K., Li V.O.K. A resequencing model for high-speed packet-switching networks // Journal Computer Communications, 010. Vol. 33. Issue 4. Pp. 443–453.
14. Lelarge M. Packet reordering in networks with heavy-tailed delays // Mathematical Methods on Operation Research, 2008. Vol. 67. Pp.341–371.
15. Matyushenko S. I. Stationary characteristics of the two-channel queueing system with reordering customers and distributions of phase type // Informatics and its Applications, 2010. Vol. 4. Issue 4. Pp. 67–71. (in Russian).
16. Pechinkin A.V., Caraccio I, Razumchik R.V. Joint Stationary Distribution Of Queues In Homogenous M/M/3 Queue With Resequencing // Proceedings of the 28th European Conference on Modelling and Simulation, 2014. Pp. 558–564.
17. Caraccio I., Pechinkin A.V., Razumchik R.V. Stationary characteristics of MAP—PH—2 resequencing queue // Proceedings of The First European Conference on Queueing Theory, 2014. P. 46
18. Takine T., Ren J., Hasegawa T. Analysis of the Resequencing Buffer in a Homogeneous M/M/2 Queue // Performance Evaluation, 1994. Vol. 19. Issue 4. Pp. 353–366.
19. Ye Xia, Tse, D.N.C. On the large deviations of resequencing queue size: 2-M/M/1 Case // IEEE Transactions on information theory, 2008. Vol. 54. No. 9. Pp. 4107–4118.

MODELS OF SYSTEMS WITH VARIABLE STRUCTURE IN QUEUING THEORY AND OUTPUT FLOWS

E. Proidakova

Lobachevsky State University of Nizhny Novgorod, Nizhny Novgorod, Russia
pev_1@mail.ru

Abstract

This paper is a continuation of a publication series, devoted to the study of probabilistic properties of non-classical queuing systems output flows. Here we consider a priority system controlling conflict requirement flow. With the help of a cybernetic approach a mathematical model of the output system flows in the form of a five-dimensional random vector sequence is constructed and investigated. In particular the conditions of the stationary distribution existence for this random vector sequence are obtained.

Keywords: output flow, queuing system with variable structure, priority service, cybernetic approach, nonlocal description, homogeneous vector Markov chain

1. Introduction

The study of output flow properties is most promising and pressing in modern queuing theory. This is primarily due to the widespread use of objectives and methods of queuing theory in production organization and information system development. Typically, systems like that have a branched structure and are composed of two or more subsystems combined in some way. Frequently the output flow of one subsystem is the input to the other, and in this case, the question naturally arises about the description of a subsystem output flow. To be more exact this is actually the problem of investigating a subsystem output flow. The first results in the field of research output flows were obtained in the 60-ies of XX century by P.J. Burke, J.W. Cohen, E. Reich and P.D. Finch. These results are related to the simplest systems. In our country the output flows at different times were investigated by many famous mathematicians. In his writings B.V. Gnedenko, I.N. Kovalenko, G.S. Tsitsiashvili, I.I. Yezhov, N.V. Markov V.F. Kadankov considered single-channel systems with some complications in the type of input, queuing mechanism discipline and service. As a rule, the output flow is designated as the input one is. For this purpose one of the following equivalent ways is used:

- the way determined by a random process $\{\bar{\xi}(t); t \geqslant 0\}$, where the random variable $\bar{\xi}(t)$, $t > 0$ characterizes the number of requests served by the system during the time interval $[0, t)$ and $\bar{\xi}(0) = 0$;

- the way specified by the random sequence $\{(\tau'_i, \bar{\xi}'_i); i \geqslant 1\}$, which is denoted by τ'_i and $\bar{\xi}'_i$ respectively i-th moment of the occurrence of output requirements and the number of customers served by the system at this time.

If one of the proposed methods is used for the description of the output flow the finite-dimensional distributions of the constructed object cannot be found.

2. The cybernetic approach

While constructing and analyzing the mathematical model of the control system service and its output flows a so-called **cybernetic** approach is used which was methodologically developed by A.A. Lyapunov and S.V. Yablonsky [1]. Cybernetic approach is based on the assumptions of:

- discreteness of the system acts in time;
- non-locality in the description of the structure of the control system;
- joint consideration of the block structure of the control system and its functioning over time.

These principles allow to distinguish scheme, information, coordinates and function of the control system. For the scheme defined the following blocks: the external environment, the input and output terminals, external memory, internal memory, information processing unit of external and internal memory. Particular scheme may not include some of the blocks.

The scheme of the control system of service reflects its skeletal structure. The information represents the status of memory cells in the current time. The coordinates describe the blocks location in the scheme of the control system. The function specifies the action that the system can do, moving to the next discrete points in time. Also introduces the concept of the algorithm controlling of scheme blocks states.

3. Application of cybernetic approach to construct a mathematical model of a priority system

In this paper the system control m independent and conflict traffic flows $\Pi_1, \Pi_2, \ldots, \Pi_m$ using traffic light at the intersection of m highways is explored. Conflict flows means that their service takes place in disjoint intervals. Incoming flows are divided into three types: Π_1 — low-intensity flow; $\Pi_2, \ldots, \Pi_{m-1}$ — medium intensity flows and Π_m — high flow. Only flow Π_1 has priority. It means that any flow request Π_1 must be serviced as soon as possible, but without interrupting other service requirements. Here machine service means its moving through the intersection. Each of the m flows has the main stage of the service and stage readjustment. During the readjustment continues to be served the same flow as in the previous main stage, but with greater intensity. For the high flow Π_m is additionally introduced another time period in which it continues to service. So servicing device must have $2m + 1$ states, and besides: $\Gamma^{(2j-1)}$ is a state in which the flow Π_j served only with intensity $\mu_j > 0$; $\Gamma^{(2j)}$

is a state in which the flow Π_j served only with intensity $\mu'_j > \mu_j$; $\Gamma^{(2m+1)}$ is a state in which the flow Π_m served only with intensity $\mu''_m > \mu_m$. Here μ_j (μ'_j) determines the average number requests serviced per unit time in the $\Gamma^{(2j-1)}$ ($\Gamma^{(2j)}$), respectively, and μ''_m determines the average number of claims processed per unit of time in the state $\Gamma^{(2m+1)}$.

Cybernetic approach allowed to identify the following blocks of the control system scheme:

- the input flows Π_1, Π_2, ..., Π_m are the first type of input poles;
- the flows of saturation Π'_1, Π'_2, ..., Π'_m are the second type of input poles;
- the queues O_1, O_2, ..., O_m are the external memory;
- the service mechanism strategies δ_1, δ_2, ..., δ_m are the information processing unit of external memory;
- the servicing device with states $\Gamma^{(1)}$, $\Gamma^{(2)}$, ..., $\Gamma^{(2m+1)}$ is the internal memory, the algorithm of variation these states is the information processing unit of external memory;
- the system output flows $\overline{\Pi}_1$, $\overline{\Pi}_2$, ..., $\overline{\Pi}_m$ are the output poles.

Listed above blocks of the control system scheme are shown in figure 1.

Set of: the states of queues, the states of servicing device, the states of input flows, the states of output flows and the states of flows of saturation determines the information of the priority queueing system. The numbers of: the input flows, the flows saturation, the output streams, the queues, and the servicing device states define the coordinates of the priority control system. The function of the system is flow control (permit or deny the start of service for each of them), and direct service heterogeneous requirements. According to cybernetic approach provisions the system was observed in discrete timepoints τ_i, $i = 0, 1, \ldots$ or on an interval $[\tau_i, \tau_{i+1})$. Here τ_i is a moment of switching of a servicing device phase. Actually dot stochastic process $\{\tau_i; i \geqslant 0\}$ defines (sets) the scale cycles operating time control system.

We define with $j = \overline{1, m}$ and $i = 0, 1, \ldots$ the following random elements:

1) $\eta_{j,i}$ is number of the flow Π_j demands which arrive over the time interval $[\tau_i, \tau_{i+1})$, $\eta_{j,i} \in X = \{0, 1, \ldots\}$;

2) Γ_i is a state of the servicing device on a time interval $[\tau_i, \tau_{i+1})$, and $\Gamma_i \in \Gamma = = \{\Gamma^{(1)}, \ldots, \Gamma^{(2m)}\}$;

3) $\xi_{j,i}$ is the maximum possible number of demands which can be served for a time interval $[\tau_i, \tau_{i+1})$ on a flow Π_j, $\xi_{j,i} \in \{0, l'_j, l_j\}$, $j = \overline{1, m-1}$, $\xi_{m,i} \in \{0, l''_m, l'_m, l_m\}$. Here l_j with $j = \overline{1, m}$ determines the maximum number of the flow Π_j customers, which can be served for state $\Gamma^{(2j-1)}$ operating time, and $l_j = [\mu_j T_{2j-1}]$, $l'_j = [\mu'_j T_{2j}]$ and $l''_m = [\mu''_m T_{2m+1}]$, $l_j \geqslant l'_j$, $l_m \geqslant l''_m$;

4) $\varkappa_{j,i}$ is a flow Π_j queue length at a time τ_i, $\varkappa_{j,i} \in X = \{0, 1, \ldots\}$;

5) $\bar{\xi}_{j,i}$ is a actual number of the flow Π_j serviced requests for a time interval $[\tau_i, \tau_{i+1})$, $\bar{\xi}_{j,i} \in Y_j = \{0, 1, \ldots, l_j\}$;

6) $\bar{\xi}_{j,-1}$ is a actual number of the flow Π_j serviced requests for a time interval $[0, \tau_0)$, $\bar{\xi}_{j,-1} \in Y_j$.

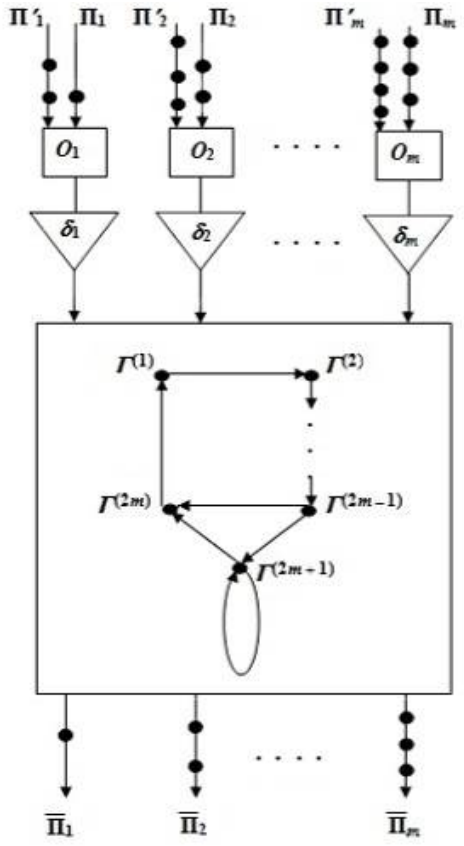


Fig. 1: Functional scheme of a priority control system.

The family $\{\bar{\xi}_{j,i}; j = \overline{1,m}, i = 0,1,\ldots\}$ defines the output flows nonlocal description. Applications are selected out of flow queue according to an extreme strategy service mechanism, that is $\bar{\xi}_{j,i} = \min\{\varkappa_{j,i} + \eta_{j,i}; \xi_{j,i}\}$. Input flows are Poisson, so we have a relation $P(\eta_{j,i} = u_j | \Gamma_i = \Gamma^{(r)}) = (\lambda_j T_r)^{u_j}(u_j!)^{-1}\exp\{-\lambda_j T_r\} = \varphi_j(u_j, T_r)$, $u_j \in X$, $r = \overline{1, 2m+1}$. To describe the servicing device states changing priority algorithm we introduce the following function $U(\Gamma^{(r)}, w_1, u_1)$, $\Gamma^{(r)} \in \Gamma$, $w_1 \in X$ and $u_1 \in X$:

$$U(\Gamma^{(r)}, w_1, u_1) = \begin{cases} \Gamma^{(1)}, r = 2m; \\ \Gamma^{(r+1)}, r = \overline{1, 2m-2}; \\ \Gamma^{(2m)}, r \in \{2m-1, 2m+1\}, w_1 = 0, u_1 > 0; \\ \Gamma^{(2m)}, r \in \{2m-1, 2m+1\}, w_1 > 0; \\ \Gamma^{(2m+1)}, r \in \{2m-1, 2m+1\}, w_1 = u_1 = 0. \end{cases}$$

Then $\Gamma_{i+1} = U(\Gamma_i, \varkappa_{1,i}, \eta_{1,i})$.

4. Analytical studies results

In view of the independence of input flows and flows saturation studied only vector sequence $\{(\Gamma_i, \varkappa_{1,i}, \varkappa_{m,i}, \bar{\xi}_{1,i-1}, \bar{\xi}_{m,i-1}); i \geqslant 0\}$, that determines the behavior of the system on the priority Π_1 and intensity Π_m of the flow. This sequence sets and nonlocal description of output flows on these directions. Here for the output flow components meet the $\bar{\xi}_{1,i}$ and $\bar{\xi}_{m,i}$, and $\Gamma_i, \varkappa_{1,i}, \varkappa_{m,i}$ into play tag.

We consider that an initial distribution is defined, that is known probabilities $P(\Gamma_0 = \Gamma^{(s)}, \varkappa_{1,0} = x_1, \varkappa_{m,0} = x_m, \bar{\xi}_{1,-1} = y_1, \bar{\xi}_{m,-1} = y_m)$. Well as for the vector sequence $\{(\Gamma_i, \varkappa_{1,i}, \varkappa_{m,i}, \bar{\xi}_{1,i-1}, \bar{\xi}_{m,i-1}); i \geqslant 0\}$ were proved several assertions.

Theorem 1. *For the random vector sequence* $\{(\Gamma_i, \varkappa_{1,i}, \varkappa_{m,i}, \bar{\xi}_{1,i-1}, \bar{\xi}_{m,i-1}); i \geqslant 0\}$ *the following recurrence relation satisfies:* $(\Gamma_{i+1}, \varkappa_{1,i+1}, \varkappa_{m,i+1}, \bar{\xi}_{1,i}, \bar{\xi}_{m,i}) = (U(\Gamma_i, \varkappa_{1,i}, \eta_{1,i}),$
$\max\{0, \varkappa_{1,i}+\eta_{1,i}-\xi_{1,i}\}, \max\{0, \varkappa_{m,i}+\eta_{m,i}-\xi_{m,i}\}, \min\{\varkappa_{1,i}+\eta_{1,i}, \xi_{1,i}\}, \min\{\varkappa_{m,i}+\eta_{m,i}, \xi_{m,i}\})$.

Theorem 2. *The random vector sequence* $\{(\Gamma_i, \varkappa_{1,i}, \varkappa_{m,i}, \bar{\xi}_{1,i-1}, \bar{\xi}_{m,i-1}); i \geq 0\}$ *is a homogeneous Markov chain with a countable number of states with a known distribution of the initial vector* $(\Gamma_0, \varkappa_{1,0}, \varkappa_{m,0}, \bar{\xi}_{1,-1}, \bar{\xi}_{m,-1})$.

Theorem 3. *The space of all possible states of a homogeneous vector Markov chain* $\{(\Gamma_i, \varkappa_{1,i}, \varkappa_{m,i}, \bar{\xi}_{1,i-1}, \bar{\xi}_{m,i-1}); i \geq 0\}$ *splits into non-closed sets of minor states and the minimal closed set of essential communicating aperiodic states.*

Theorem 4. *For the existence of a stationary distribution of a homogeneous vector Markov chain* $\{(\Gamma_i, \varkappa_{1,i}, \varkappa_{m,i}, \bar{\xi}_{1,i-1}, \bar{\xi}_{m,i-1}); i \geq 0\}$ *are sufficient that the two inequalities:* $\lambda_1 T - l_1 - l'_1 < 0$, $\lambda_m T - l_m - l'_m < 0$.

Partially proof of these theorems are presented in [2, 3].

5. Conclusion

In this paper the probability properties of output flows arising in non-classical queuing systems are studied. Here we consider a priority system controlling conflict requirement flow. With the help of a cybernetic approach a mathematical model of the output system flows in the form of a five-dimensional random vector sequence is constructed and investigated. In particular the conditions of the stationary distribution existence for this sequence are obtained.

The work was supported by the state budget theme N 01201456585 "Mathematical modeling and analysis of stochastic evolution and decision-making processes" and state program "Support to the leading universities of the Russian Federation in order to enhance their competitiveness among the world's leading research and education centers".

REFERENCES

1. Lyapunov A. A., Yablonsky S. V. Theoretical Problems of Cybernetics // Problems of Cybernetics. Moscow. 1963. P. 5-22.
2. Proydakova E. V. Probability properties of output flows in a priority queuing system // Vestnik of Lobachevsky State University of Nizhni Novgorod. 2012. V. 5(2). P. 190-196.
3. Proidakova E. Control of output flows of priority system service with feedforward // Materials of XI All-Russian school-conference of young scientists "Managing large systems", 9-12 September 2014, Arzamas. P. 330-349.

METHOD FOR APPROXIMATING JOINT STATIONARY DISTRIBUTION IN FINITE CAPACITY QUEUE WITH NEGATIVE CUSTOMERS AND BUNKER FOR OUSTED CUSTOMERS

R. Razumchik[1]

[1] Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, Moscow, Russia
rrazumchik@ipiran.ru

Abstract

Queueing systems and networks with negative customers have been a subject of extensive research for two decades providing means to describe work-removal phenomena in different communication and information processing systems. This paper is the continuation of research devoted to elaboration of mathematical techniques for performance evaluation of queueing systems with different types of negative customers. Specifically we consider the Markovian single server queueing system with two finite queues in which the system time of a tagged customer may depend on both the customers arrived to the system earlier and later than the tagged one. New regular customers arrive at the system according to Poisson flow, occupy one place in buffer (if it is not full) and receive service in FIFO order. External negative signals also arrive at the system according to Poisson flow with different parameter. Each negative signal transforms one regular customer into delayed one by moving it to another finite-capacity queue (bunker), wherefrom, if it was accepted, it is served with lower priority than the regular ones. We propose new method based on Chebyshev and Gegenbauer polynomials for approximate calculation of joint stationary probability distribution of queues in buffer and bunker. Numerical example is provided.

Keywords: negative arrivals, finite capacity, queueing system, stationary distribution, approximation

1. Introduction

Queueing systems and networks with negative arrivals or customers have received significant attention from researches since they were introduced in early 1990-s (see [6], [7], [10]). From practical point of view negative arrivals may represent commands to delete some jobs, as in distributed processing systems, or transactions in database systems when certain operations become impossible. Such arrivals typically happen due to to external causes and thus present different approach to work-removal phenomena in comparison with classic queues with impatient customers or queues with reneging. This topic has collected a big body of knowledge and it is hardly possible

to give the comprehensive review here. For review of the publications up to 2011 one can refer to [1], [3] and [19]. Detailed bibliography up to 2011 can be found in [20]. Among the latest research results one can mention, for example, [11]–[21] and novel approach to the analysis of networks with negative customers given in [22], [23].

The concept of negative customer has turned out to be quite versatile. Quite often it is assumed that the arrival of negative customer leads to permanent removal of other (ordinary) customers residing in the system. It can also be the case that negative arrivals move ordinary customers between queues or nodes in the network. One of the problems' statements introduced in [13] implies that negative customers do not remove ordinary customers waiting for service in the buffer from the system, but delay their service by displacing them into another queue (bunker) wherefrom they are served according to a certain discipline (say, for example, with relative priority). This may also be view as a type of re-sequencing as stated in [24]. Queues with such negative arrivals can be used in the modelling of, for example, fault-related processes in distributed computing system, in databases with two-phase commit strategy and in the design of traffic generators, when one needs to introduce (controlled) stochastic disordering in the flow.

It should be mentioned that we are primarily interested in the study of stationary characteristics. The latest results of the analysis of queueing systems with this type of negative customers are presented in [14]–[17], [24]. Particular interest presents the case when queues' capacities (of buffer and bunker) are finite. It is shown in [16] that apart from matrix analytic approach, Gaussian elimination method and other numerical methods it is possible to obtain joint stationary distribution utilizing properties of generating functions and special functions (Chebyshev and Gegenbauer polynomials). Note that such approach is not new as it was already used in [2], but for the analysis of another type of queueing system. The method presented in [16], which allows one to do it, is algebraic and suitable for exact arithmetics implementation though still suffers from one computational problem. The problem is the need to efficiently perform basic operations with arbitrary big numbers. In this paper we present the modification of method which makes it from the one hand approximate and from the other hand allows one to carry out computations of stationary performance characteristics for higher values of queues' capacities almost without loosing accuracy. One interesting feature of the proposed method revealed during numerical experiments (see section 4) indicates that, depending on initial values, one can obtain exact values for several performance characteristics using worst-case approximation (which substantially cuts computation time). Due to space limitation we will omit tedious derivations and just present main ideas and key results. Comparisons with other known solution methods (matrix-geometric, Gaussian elimination) and discussion of computational complexity and accuracy is also left aside.

The rest of the paper is organized as follows. Section 2 is devoted to the detailed description of the system, equilibrium equations for joint stationary probabilities and several useful relations that follow from rate-in-rate-out principle. In section 3 the idea of approximation is briefly described. Some computational results with short discussion are provided in section 4.

2. Description of the system

Consideration is given to the queueing system (QS) with incoming Poisson flow of the ordinary customers of intensity λ, Poisson flow of negative customers of intensity λ^- and one server. For ordinary customers there is a buffer of finite capacity $k > 0$. If upon arrival of the ordinary customer the buffer is full, ordinary customer is lost. A negative customer upon arrival moves one ordinary customer from buffer, if it is not empty at the instant of arrival, to another queue (bunker) of finite capacity $r > 0$. If upon arrival of a negative customer the queue in the buffer is not empty but bunker is full, the customer displaced from buffer is considered to be lost. Service times of customers from both buffer and bunker are exponential with the same parameter μ. Upon service completion customer from buffer enters server and, if buffer is empty, customer from bunker goes to service (i.e. customers from bunker are served with relative priority).

Denote by p_{ij} stationary probability of the fact that there are i customers in buffer, j customers in bunker and server is busy. By p_0 we denote stationary probability of empty system. Stationary distribution satisfies the following system of equilibrium equations:

$$\lambda p_0 = \mu p_{00}, \tag{1}$$

$$(\lambda + \mu)p_{00} = \lambda p_0 + \mu p_{10} + \mu p_{01}, \;\; i = 0, \;\; j = 0, \tag{2}$$

$$(\lambda + \mu + \lambda^-)p_{i0} = \lambda p_{i-1,0} + \mu p_{i+1,0}, \;\; i = 1, \ldots, k-1, \;\; j = 0, \tag{3}$$

$$(\mu + \lambda^-)p_{kj} = \lambda p_{k-1,j}, \;\; i = k, \;\; j = 0, \ldots, r, \tag{4}$$

$$(\lambda + \mu)p_{0j} = \mu p_{1j} + \lambda^- p_{1,j-1} + \mu p_{0,j+1}, \;\; i = 0, \;\; j = 1, \ldots, r-1, \tag{5}$$

$$(\lambda + \mu)p_{0r} = (\mu + \lambda^-)p_{1r} + \lambda^- p_{1,r-1}, \;\; i = 0, \;\; j = r, \tag{6}$$

$$(\lambda + \mu + \lambda^-)p_{ij} = \lambda p_{i-1,j} + \mu p_{i+1,j} + \lambda^- p_{i+1,j-1}, \;\; i = 1, \ldots, k-1, \;\; j = 1, \ldots, r-1, \tag{7}$$

$$(\lambda + \mu + \lambda^-)p_{ir} = \lambda p_{i-1,r} + \lambda^- p_{i+1,r-1} + (\mu + \lambda^-)p_{i+1,r}, \;\; i = 1, \ldots, k-1, \;\; j = r, \tag{8}$$

with the normalization condition

$$p_0 + \sum_{i=0}^{k} \sum_{j=0}^{r} p_{ij} = 1. \tag{9}$$

Although for the sake of clarity we will consider the case $k = r$ and $r \geq 6$, the proposed analysis is valid also for the general case $k > 0$, $r > 0$.

One of the key relations that allows the computation of the joint stationary probability distribution p_{ij} is the one which shows how probability of empty system p_0 depends on boundary probabilities $\{p_{0j}, \; j = 1, \ldots, r\}$, $\{p_{i0}, \; i = 1, \ldots, r-1\}$, $\{p_{ir}, \; i = 1, \ldots, r-1\}$ and $\{p_{rj}, \; j = 0, \ldots, r\}$. It is shown in [16], from rate-in-rate-out principle it follows that for probability p_0 equals

$$p_0 = 1 - \frac{(1 - q^{r+1})\rho}{(1 - \rho + \frac{\lambda^-}{\mu})q^{r+1}} \sum_{j=0}^{r} p_{rj}. \tag{10}$$

where, here and henceforth, $\rho = \lambda/\mu$, $q = \lambda/(\mu + \lambda^-)$. Thus for the considered queueing system p_0 depends only on $\{p_{rj},\ j = 0, \dots, r\}$. In [16] it is shown that if one finds X_{rj} such that $p_{rj} = X_{rj}p_0$, then from (10) one can compute probability p_0 and consequently joint stationary probability distribution. In the next section we show how the approximation to the joint stationary probability distribution can be obtained by introducing several modifications into the method of [16].

3. Approximation of joint stationary probability distribution

Let us introduce double probability generating function (PGF)

$$P(u, v) = \sum_{i=0}^{r} \sum_{j=0}^{r} p_{ij} u^i v^j, \quad 0 \le u \le 1, \quad 0 \le v \le 1. \tag{11}$$

Multiplying (1)–(8) by $u^i v^j$ and summing over all values of i and j, having collected common terms, we obtain

$$-[\lambda u^2 - (\lambda + \mu + \lambda^-)u + \mu + \lambda^- v]P(u, v) = \frac{\mu u(v-1)}{v} p_{00} + \lambda u^{r+1}(1-u) \sum_{j=0}^{r} p_{rj} v^j + \\ + \frac{(\mu + \lambda^- v)(u - v)}{v} \sum_{j=0}^{r} p_{0j} v^j + \lambda^- v^r (1 - v) \sum_{i=1}^{r} p_{ir} u^i. \tag{12}$$

The expression in the square brackets in the left part of the previous equation is a polynomial of a second degree in u. Its roots have the form

$$u_{1,2} = u_{1,2}(v) = \frac{\lambda + \mu + \lambda^- \mp \sqrt{(\lambda + \mu + \lambda^-)^2 - 4\lambda(\mu + \lambda^- v)}}{2\lambda}. \tag{13}$$

It can be seen, that $u_2(v) > 1$ and $0 < u_1(v) \le 1$ for $0 \le v \le 1$. Generating function $P(u, v)$ is the ratio of two polynomial functions. For each value of v PGF $P(u, v)$ is continuous function of u on the whole set $\mathbf{R}$ of real numbers. Then, as left part in (12) vanishes at points $(u_1(v), v)$ and $(u_2(v), v)$ then the right part must vanish at these point too. Therefore we obtain two equations:

$$\frac{\mu u_1(v-1)}{v} p_{00} + \lambda u_1^{r+1}(1-u_1) \sum_{j=0}^{r} p_{rj} v^j + \\ + \frac{(\mu + \lambda^- v)(u_1 - v)}{v} \sum_{j=0}^{r} p_{0j} v^j + \lambda^- v^r (1 - v) \sum_{i=1}^{r} p_{ir} u_1^i = 0, \tag{14}$$

$$\frac{\mu u_2(v-1)}{v} p_{00} + \lambda u_2^{r+1}(1-u_2) \sum_{j=0}^{r} p_{rj} v^j + \\ + \frac{(\mu + \lambda^- v)(u_2 - v)}{v} \sum_{j=0}^{r} p_{0j} v^j + \lambda^- v^r (1 - v) \sum_{i=1}^{r} p_{ir} u_2^i = 0. \tag{15}$$

If we now express term with p_{00} from (14) and put it in (15), after collecting common terms, we get the following equation

$$\Big(\frac{u_2^r - u_1^r}{u_2 - u_1} - \frac{u_2^{r+1} - u_1^{r+1}}{u_2 - u_1}\Big)\sum_{j=0}^{r} p_{rj}v^j + \frac{\lambda^- v^r(1-v)}{\lambda}\sum_{i=1}^{r} p_{ir}\frac{u_2^{i-1} - u_1^{i-1}}{u_2 - u_1} + \sum_{j=0}^{r} p_{0j}v^j = 0. \quad (16)$$

Now if one expresses term with $\sum_{j=0}^{r} p_{0j}v^j$ from (14) and puts it in (15), one obtains

$$\begin{aligned}
&\mu(1-v)p_{00} + \mu\Big(\frac{u_2^r - u_1^r}{u_2 - u_1} - \frac{u_2^{r+1} - u_1^{r+1}}{u_2 - u_1}\Big)\sum_{j=0}^{r} p_{rj}v^j + \\
&+\Big(\lambda^- \frac{u_2^r - u_1^r}{u_2 - u_1} - (\lambda + \lambda^-)\frac{u_2^{r+1} - u_1^{r+1}}{u_2 - u_1} + \lambda\frac{u_2^{r+2} - u_1^{r+2}}{u_2 - u_1}\Big)\sum_{j=0}^{r} p_{rj}v^{j+1} + \\
&+\frac{\mu\lambda^- v^r(1-v)}{\lambda}\sum_{i=1}^{r} p_{ir}\frac{u_2^{i-1} - u_1^{i-1}}{u_2 - u_1} - \lambda^- v^{r+1}(1-v)\sum_{i=1}^{r} p_{ir}\frac{u_2^i - u_1^i}{u_2 - u_1} + \\
&+\frac{(\lambda^-)^2 v^{r+1}(1-v)}{\lambda}\sum_{i=1}^{r} p_{ir}\frac{u_2^{i-1} - u_1^{i-1}}{u_2 - u_1} = 0. \quad (17)
\end{aligned}$$

In [16] it is shown that $(u_2^i - u_1^i)/(u_2 - u_1)$, $i \geq 1$, are polynomials of integer degree $\lfloor \frac{i}{2} \rfloor$ in v with real coefficients and thus the left part of (16) and (17) are polynomials of integer degree in v with real coefficients depending on λ, λ^-, μ and certain p_{ij}. Specifically, the degree of the polynomial in (16) is $r + 1 + \lfloor \frac{r-1}{2} \rfloor$ and of the polynomial in (17) is $r + 2 + \lfloor \frac{r-1}{2} \rfloor$. From the fact that these both polynomial are equal to zero $\forall v \in [0, 1]$, their coefficients are all equal to zero. From this observation in [16] one obtains two systems of algebraic equations, whose solution eventually allows computation of the whole joint stationary distribution p_{ij}. The complexity of terms involved in equations is very high which complicates the computation for high values of r though the procedure itself is simple. This complexity can be reduced if one considers approximations to roots u_1 and u_1 instead of their exact form (13). Specifically one can consider Lagrange interpolating polynomials $L_n^1(v)$ and $L_n^2(v)$, which coincide with $u_1(v)$ are $u_2(v)$, respectively, at $n + 1$ different points. As interpolation nodes it is better to take zeros of Chebyshev polynomials $T_{n+1}(v)$ of the first kind of degree $n + 1$. Now expressions $(u_2^i - u_1^i)/(u_2 - u_1)$, $i \geq 1$, that enter (16) and (17), can be rewritten in the form

$$\frac{u_2^i - u_1^i}{u_2 - u_1} = \frac{u_2(v)^i - u_1(v)^i}{u_2(v) - u_1(v)} = \frac{L_n^2(v)^i - L_n^1(v)^i}{L_n^2(v) - L_n^1(v)} = f_{in}(v),\ i \geq 1, v \in [0, 1].$$

It is known (see e.g. [9]) that function $f_{in}(v)$ that interpolates $(u_2^i - u_1^i)/(u_2 - u_1)$ at $n+1$ zeros of Chebyshev polynomials of the first kind, can be written as their combination, that is $f_{in}(v) \approx \sum_{k=0}^{n} w_{ik}T_k(2v - 1)$, where w_k are properly defined numbers. If in this expansion for $f_{in}(v)$ one takes the *same* number of interpolation nodes n for each $i > 1$ in such a way that $n < \lfloor \frac{r}{2} \rfloor$, this will reduce the degrees of polynomials in (16) and

Table 1: Performance characteristics for $r = 70$, $\lambda = 7$, $\lambda^- = 5$, $\mu = 10$

n	π_1	π_2	M_{buff}	M_{bunk}	ω_{buff}	ω_{bunk}
3	0.000000	-0.428571	0.875000	2352588.31	0.125000	-40.982058
4	0.000000	-0.428571	0.875000	-1423864.72	0.125000	-41.253602
5	0.000000	-0.428571	0.875000	654765.49	0.125000	-41.263460
10	0.000000	-0.428571	0.875000	-656.41	0.125000	-33.860188
15	0.000000	-0.154337	0.707032	23.859852	0.101005	8.045378
20	0.000000	-0.000011	0.612507	1.021670	0.087501	0.625477
25	0.000000	-0.000008	0.612505	1.021108	0.087501	0.625142
	0.000000	0.000000	0.612507	1.021670	0.087501	0.625477

(17). Specifically the degree of the polynomial in (16) will be reduced to $r + 1 + n$ and of the polynomial in (17) to $r + 2 + n$. No by solving the system of equations, which follows from (16) and (17), one obtains more simple but approximate algorithm for computation of p_{ij}, which we omit here due to space limitation, and proceed to numerical example.

4. Numerical example

Let us have a look on how performance characteristics of the considered queueing system depend on the quality of approximation. We will be interested in probability that the arriving customer is lost due to the full buffer (π_1), probability that customer displaced from buffer is lost due to the full bunker (π_2), mean number of customers in buffer (M_{buff}) and bunker (M_{bunk}), mean time customer spends in buffer (ω_{buff}) and bunker (ω_{bunk}).

We consider two different combinations of initial parameters, that is

1) $r = 70$, $\lambda = 7$, $\lambda^- = 5$, $\mu = 10$ (Table 1);
2) $r = 70$, $\lambda = 12$, $\lambda^- = 5$, $\mu = 10$ (Table 2);

Numerical computations were carried out using different number of interpolation nodes n, specifically $3 \leq n < \lfloor \frac{r}{2} \rfloor$. The results are presented in Table 1 and Table 2. Last line in each table states exact values of performance characteristics, obtained using algorithm from [16].

From the results in the tables one can see that approximate results can be quite accurate even when number of interpolation points n is very low. Numerical experiments show that computational time decreases substantially with slow decrease of n. Nevertheless it remains an open questing how to choose appropriate number of nodes n so as not to loose much in accuracy but noticeably gain in computation time and whether the value of n is the same for all performance characteristics independently of combination of initial parameters. It should also be stated that the proposed method is suitable not only for the problem considered but is also applicable to a range of problems, which borders are to be determined.

Table 2: Performance characteristics for $r = 70$, $\lambda = 12$, $\lambda^{-} = 5$, $\mu = 10$

n	π_1	π_2	M_{buff}	M_{bunk}	ω_{buff}	ω_{bunk}
3	0.000000	0.166667	3.999991	5.709E+10	0.333333	-40.996245
4	0.000000	0.166667	3.999991	-4.311E+10	0.333333	-41.310644
5	0.000000	0.166667	3.999991	2.553E+10	0.333333	-41.386324
10	0.000000	0.166667	3.999991	-1.044E+08	0.333333	-41.789411
15	0.000000	0.166665	3.999999	7987.995960	0.333333	-64.828352
20	0.000000	0.166667	3.999990	69.952997	0.333332	36.561334
25	0.000000	0.166667	3.999990	65.092835	0.333333	31.908287
	0.000000	0.166667	3.999990	69.952997	0.333333	36.561334

Acknowledgments.

This work was supported by the Russian Foundation for Basic Research (grant 15-07-03007, 13-07-00223).

REFERENCES

1. Artalejo J.R. G-networks: a versatile approach for work removal in queueing systems // European journal of operation research, 2000. Vol. 126. pp. 233–249.
2. Avrachenkov K.E., Vilchevsky N.O., Shevljakov G.L. Priority queueing with finite buffer size and randomized push-out mechanism // Proceedings of the ACM international conference on measurement and modelling of computer, San Diego, 2003. pp. 324–335.
3. Bocharov P.P., Vishnevskii V.M.: G Networks: Development of the Theory of Multiplicative Networks // Automation and remote control, 2003. no. 5, pp. 46–74.
4. Dao-Thi T., Fourneau J., Tran M.: Networks of Order Independent Queues with Signals // 21st International Symposium on Modeling, Analysis and Simulation of Computer and Telecommunication Systems, 2013. pp. 131–140.
5. Erdelyi A., Bateman H.: Higher transcendental functions. Volume II. Robert E. Krieger Publishing Company (1985)
6. Fourneau J., Gelenbe E., Suros R. G-Networks with Multiple Classes of Negative and Positive Customers // Theor. Comput. Sci., 1996. Vol. 155. No. 1. pp. 141–156
7. Gelenbe E., Glynn P., Sigman K.: Queues with Negative Arrivals // Journal of Applied Probability, 1991. vol. 28, pp. 245–250.
8. Gelenbe E. G-networks with instantaneous customer movement // Journal of Applied Probability, 1993. Vol. 30. No. 3. pp. 742–748.
9. Gil A., Seguram J., Temme N.M. Numerical Methods for Special Functions (1 ed.). Soc. for Industrial and Applied Math., Philadelphia, PA, USA, 2007.

10. Harrison P.G., Pitel E.: Sojourn Times In Single-Server Queues With Negative Customers. Journal of Applied Probability, 1993. vol. 30, pp. 943–963.
11. Klimenok V., Dudin A.: A BMAP/PH/N Queue with Negative Customers and Partial Protection of Service // Communications in Statistics - Simulation and Computation, 2012. vol. 41, Issue 7, pp. 1062–1082.
12. Krishna Kumar B., Pavai Madheswari S., Anantha Lakshmi S. R.: An M/G/1 Bernoulli feedback retrial queueing system with negative customers // Operational Research, 2013. vol. 13, Issue 2, pp. 187–210.
13. Manzo R., Cascone N., Razumchik R.V.: Exponential queuing system with negative customers and bunker for ousted customers. Automation and Remote Control, 2008. vol. 69. pp. 1542–1551.
14. Pechinkin A.V., Razumchik R. V.: Stationary waiting time distribution in queueing system with negative customers and bunker for ousted customers under LAST-LIFO-LIFO service discipline // Journal of Communications Technology and Electronics, 2012. vol. 57. no. 12. pp. 1331–1339.
15. 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. vol. 57, no. 8, pp. 882–891.
16. Razumchik R.V. Analysis of Finite Capacity Queue with Negative Customers and Bunker for Ousted Customers Using Chebyshev and Gegenbauer Polynomials // Asia-Pacific Journal of Operational Research, 2014. Vol. 31. No. 4.
17. Razumchik R. V.: Stationary Waiting Time Distribution In Queueing System With Negative Customers And Bunker For Ousted Customers Under First-Fifo-Fifo Service Discipline // Informatica and Appications Scientific journal, 2013. vol. 7, Issue 2, pp. 34–39. (in Russian)
18. Suetin P.K.: Classical orthogonal polynomials. Nauka, Moscow, 1978. (In Russian)
19. Van Dijk N. Queueing Networks - A Fundamental Approach. International Series in Operations Research and Management Science (Vol. 154). Ed. R.J. Boucherie and N.M. van Dijk: Springer (2011)
20. Van Do T. An initiative for a classified bibliography on G-networks // Performance Evaluation, 2011. vol. 68, no. 4, pp. 385–394.
21. Van Do T., Papp D., Chakka R., Sztrik J., Wang J.: M/M/1 retrial queue with working vacations and negative customer arrivals // Int. J. of Advanced Intelligence Paradigms, 2014. Vol. 6, no.1, pp.52–65.
22. Balsamo S., Harrison P.G., Marin A. A unifying approach to product-forms in networks with finite capacity constraints // SIGMETRICS, 2010. pp. 25–36.
23. Harrison P.G., Marin A. Product-forms in Multi-way synchronisations // Computer Journal, 2014. pp.16–93.
24. Razumchik R., Telek M. Delay analysis of a queue with re-sequencing buffer and Markov environment // Queueing Systems, 2015. DOI 10.1007/s11134-015-9444-z.

TRACE-DRIVEN WORKLOAD MODELING IN CLUSTER SYSTEMS

A. Rumyantcev[1]*, R. Razumchik*[2]

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

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

ar0@krc.karelia.ru, rrazumchik@ipiran.ru

Abstract

The problem of energy saving in high performance clusters is considered. An extension of dynamic resizing method is proposed based on the prediction of future workload, that aims at provisioning of service level control.

Keywords: cluster, HPC, energy consumption, workload modeling

1. Introduction

Energy saving in big data centres (DC), as well as in high performance clusters (HPC) is the subject of extensive research nowadays. The overall power consumption of DCs is about 1.5% of total electricity usage [3]. A number of commercial energy saving software for DCs is available (e.g. HP Power Management, Eaton Intelligent Power Software Suite), and energy saving functions are built in a number of HPC queue managers (SLURM, Moab), naturally supported by energy saving modes of hardware and operation system. However, most this software either exploits the so-called power capping mode (when the maximum power consumption is limited above), or user-defined rules, regardless of the possible degradation of service level. Still there is a number of papers in which the performance-energy tradeoff is investigated in terms of the server management policies, see e.g. [3, 4].

Here we consider the problem of energy saving in a moderately loaded HPC, realizing that energy saving in a highly utilized HPC is hardly possible, as time needed for switching to the energy saving mode is usually greater than jobs interarrival times. We utilize the natural assumption that, given a future workload, the energy saving mode may be effectively utilized with limited (if any) loss of service level. Note that, in contrast to DCs, the service level of HPCs usually does not get fixed. Our basic idea is to forecast the future workload and, depending on the forecast, apply dynamic resizing, i.e. switch idle servers to the energy saving mode. Although we explicitly investigate only a coarse management method, in concluding remarks we discuss a finer-grained method. We also provide some numeric results that illustrate the applicability of the method.

2. Session-based prediction

Users generally access the resources of an HPC via `ssh` access sessions. During each session the user may submit tasks, check the state of the queue, edit the sources, upload/download files. The session activity is basically logged, e.g. linux-type operating systems have the so-called `utmp`, `wtmp` files.

We suggest the following raw algorithm of energy saving mode management. Pick a user u from the set of all HPC users U. The work of user u with the HPC is a sequence of consecutive ON and OFF periods of lengths $\{t_i^{ON}(u), t_i^{OFF}(u)\}$, $i \geqslant 1$, assumed to be i.i.d. Once all users of the HPC enter OFF period, say at time t and HPC becomes idle (is said to be in the global OFF period), then one.

1) estimates the probability $P_{sub}(t, T)$ of (one or more) task submission before $t+T$. If $P_{sub}(t, T) < \alpha$ for some tolerance level α, then enter (or keep) the energy saving mode in all free servers until time $t + T$;
2) if at time $t + T$ HPC is still idle (global OFF has not finished), return to step 1.

We suggest to forecast every time a session/task ends in a global OFF. The probability $P_{sub}(t, T)$ can be estimated as follows. Denote by $F_{OFF}^{(u)}$ the (unknown) cumulative distribution function of $\{t_i^{OFF}(u)\}$. Then the probability $P_1^{(u)}(t, T)$ that the single OFF session of user u will end (and thus user will log in) in $[t, t + T)$ is

$$P_1^{(u)}(t, T) = \left[F_{OFF}^{(u)}(t(u) + T) - F_{OFF}^{(u)}(t(u))\right] / \left[1 - F_{OFF}^{(u)}(t(u))\right], \tag{1}$$

where $t(u)$ is the length of the unfinished OFF period. Assume also that a user decides to submit (one or more) tasks during each ON session independently of his previous decisions with some probability $P_2^{(u)}$. Since users are assumed to act independently, the event that the global OFF period ends before instant $t + T$ and there will be at least one submission, provided global OFF at time t, has probability

$$P_{sub}(t, T) = 1 - \prod_{u \in U}\left(1 - P_{sub}^{(u)}(t, T)\right) = 1 - \prod_{u \in U}\left(1 - P_1^{(u)}(t, T)P_2^{(u)}\right). \tag{2}$$

Note also that in real life the choice of T should obey the hardware limitation inequality $T > T_0$, where T_0 is the minimum time required to go into standby mode and return back.

Below we give some experimental results. We assumed exponential durations of the OFF periods (with rate λ_u) and used the log files of the `SLURM` queue manager (task submissions), and the `wtmp` log of the HPC of the Karelian Research Centre [2]. The data covers the range from September 26, 2014 till July 13, 2015, and consists of 976 sessions with 2311 tasks submitted by 33 users.

For each user u we estimated the value $\lambda_u = n(u) / \left(\sum_{i=1}^{n(u)} t_i^{OFF}(u)\right)$, where $n(u)$ is the total number of session records available for user u, and values $\{t_i^{OFF}(u)\}, i \geqslant 1$ are taken from `wtmp` log. Value of $P_2^{(u)}$ was estimated from the `SLURM` log. Next, we selected a sequence of values for $T = 300, 600, \ldots, 30000$ seconds and calculated the probability $P_{sub}^{(u)}(t, T) = \left(1 - e^{-\lambda_u T}\right) P_2^{(u)}$. Then for user u we calculated the estimate

$\hat{P}^{(u)}_{sub}(t,T)$ = #{ON sessions with submissions}/#{periods of length T in OFF sessions}. The value $\hat{P}^{(u)}_{sub}(t,T)$ may be considered a service level degradation for given T.

In Fig. 1 (left) we depict the behaviour of both these probabilities for one user u.

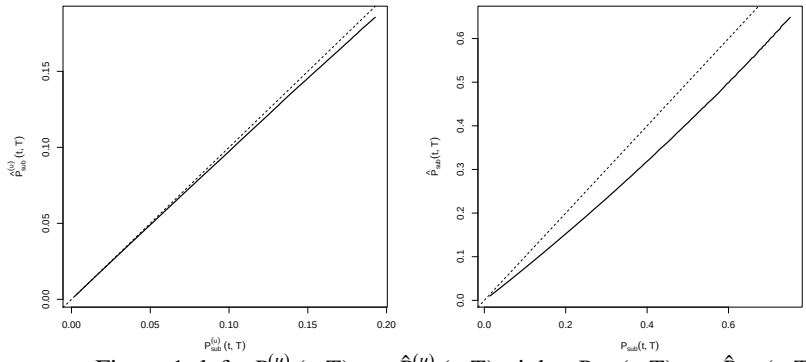


Figure 1: left: $P^{(u)}_{sub}(t,T)$ vs. $\hat{P}^{(u)}_{sub}(t,T)$, right: $P_{sub}(t,T)$ vs. $\hat{P}_{sub}(t,T)$

We performed the same procedure for the probability $P_{sub}(t,T)$. As one can see from Fig.1 (right), the probability $\hat{P}_{sub}(t,T)$ tends to be overestimated, that is, a selection of T by using (2), given $P^{(u)}_{sub}(t,T) = \left(1 - e^{-\lambda_u T}\right) P^{(u)}_2$, makes the system more conservative and should guarantee the given tolerance level α.

A more fine-grained method may concern prediction of the workload at busy times of the HPC. It seems natural to fit job submission process of each user into Markovian arrival process. It is suggested to track the required number of servers of submitted jobs in the Markovian manner. Given these two ingredients, each time a task requiring some n servers depart from HPC, one can calculate probability that users will submit a job within time T, which requires not more that n processors. Comparing the results with given threshold α decision is made whether to enter energy-saving mode.

Acknowledgments.

This work was supported by the Russian Foundation for Basic Research (grants 15-07-03007, 13-07-00008, 14-07-31007, 15-07-02341,15-07-02354, 15-29-07974).

REFERENCES

1. Kai Wang, Minghong Lin, Florin Ciucu, Adam Wierman, and Chuang Lin. 2012. Characterizing the impact of the workload on the value of dynamic resizing in data centers. SIGMETRICS Perform. Eval. Rev. 40, 1 (June 2012), 405-406.
2. Centre for collective use of Karelian Research Centre of Russian Academy of Sciences, cluster.krc.karelia.ru
3. A. Gandhi, V. Gupta, M. Harchol-Balter, and M. A. Kozuch. Optimality analysis of energy-performance trade-off for server farm management. Performance Evaluation, vol. 67, no. 11. pp. 1155–1171, 2010.
4. E. Hyytiä, R. Righter, and S. Aalto, Task assignment in a heterogeneous server farm with switching delays and general energy-aware cost structure, Performance Evaluation, vol. 75–76, pp. 17–35, May 2014.

OPERATIONAL MANAGEMENT OF WIRELESS NETWORKS RESOURCES

V.Shirokov
LANIT, Moscow, Russia

Abstract

Analyzed the traffic models, the wireless network parameters, the resources, the loads. Considered the questions of the network resources evaluations, resource allocation, resource management in real time. It is proposed to evaluate the use of network resources and to allocate resources according to one of the criteria: increased productivity (network performance, service efficiency) and energy saving. The Erlang B formula probability of denial of service and performance of trunked radio (standard DMR, TETRA) is used as an example of the calculation.

ОПЕРАТИВНОЕ УПРАВЛЕНИЕ РЕСУРСАМИ БЕСПРОВОДНОЙ СЕТИ

В.Широков
ЛАНИТ, Москва, Россия
vshirokov@lanit.ru

Аннотация

Анализируются модели трафика, параметры беспроводной сети, ресурсы, нагрузка. Рассматриваются вопросы оценки ресурсов сети, распределения ресурсов, управления ресурсами в реальном времени. Предлагается оценивать использование ресурсов сети и распределять ресурсы по одному из критериев: повышению продуктивности (производительности сети, эффективности обслуживания) и энергосбережению. Приводится пример расчета по B-формуле Эрланга вероятности отказа в обслуживании и производительности транкинговой радиосети (стандарта DMR, TETRA).

Ключевые слова: беспроводная сеть, ресурс, радиотехнология, реальное время, пакет, запрос, сообщение, тайм-слот, пропускная способность, количество каналов, базовая станция, интенсивность обслуживания, нагрузка, эрланг, час наивысшей нагрузки.

1. Введение

Управление и качественное обслуживание пользователей в коммуникационных сетях (КС) в настоящее время всҷ более актуально. Особенно управление ресурсами необходимо в беспроводных сетях, использующих радиотехнологии Wi-Fi, WiMAX, LTE, DMR, TETRA. Беспроводные

КС имеют более ограниченные ресурсы по сравнению с проводными сетями. Однако число пользователей и услуг в беспроводных сетях постоянно растчт, и всч больше мультимедийного трафика передачтся беспроводным способом. Таким образом, современная беспроводная КС является мультисервисной системой обмена информацией (МСОИ) [1]. Поскольку нагрузка на сеть определяется текущим состоянием системы, необходимо управлять передачей данных и мультимедийным трафиком в реальном времени.

2. Методология анализа и моделирования сети

2.1. Модели и параметры трафика. Большинство исследователей при оценке систем исходят из типовых моделей трафика. Некоторые исследователи полагают, что трафик является экспоненциальным, другие исходят из того, что трафик имеет фрактальные свойства [2].

Фрактальность означает, что трафик обладает последействием (памятью) и имеет структуру, не зависимую от масштаба (час, день, неделя и т.д.).

Однако целесообразнее учитывать ограничения трафика по определчнным параметрам передачи:

- временны́м задержкам и
- используемым ресурсам.

Это важно для беспроводных КС, т.к. их ресурсы более ограничены, чем у проводных сетей.

Учитывая фрактальность, можно при анализе трафика абстрагироваться от масштаба и оценивать ресурсные затраты двумя параметрами:

- длиной запроса (например, в битах или байтах);
- числом тайм-слотов на обработку запроса.

Под запросами понимаются пакеты, сообщения или транзакции в зависимости от уровня абстрагирования. Длина запроса определяет задержку при передаче запроса, а число тайм-слотов – используемый временно́й ресурс.

2.2. Ресурсы беспроводной сети. Для эффективного управления беспроводной КС требуется оперативное распределение ресурсов.

Ресурсами беспроводной КС являются, наряду с временны́м ресурсом, следующие:

- пропускная способность C каналов (линий связи) сети;
- интенсивность μ обслуживания запросов базовыми станциями (узлами).

Таким образом, основной задачей для управления беспроводной КС является использование ресурсов (C и μ) в рабочем режиме функционирования сети.

Пропускная способность C в рабочем режиме монополизируется запросом пропорционально длине b пакетов и нагрузке (активности) a, создаваемой этими запросами и оцениваемой в часы наивысшей нагрузки (ЧНН).

Нагрузка (активность) a выражается в Эрлангах.

1 Эрланг = количеству минут занятости / 60 (т.е. в течение ЧНН).

Количество минут занятости зависит от вида трафика (данные, видео или голос). Например, согласно техническим требованиям Министерства связи РФ от 29.06.1995, нагрузка a телефонных сообщений от одного абонента в ЧНН принимается равной 0,025. Соответственно число абонентов N, которые могут обслуживаться базовой станцией (БС) в ЧНН, рассчитывается исходя из этой нагрузки.

Соответственно общая нагрузка A от всех абонентов N в ЧНН составляет:

$A = a * N$,

где a - нагрузка от одного абонента, N - количество абонентов.

Интенсивность μ обслуживания оценивается общим количеством c тайм-слотов (трафиковых каналов), которые выделяет БС в единицу времени:

$\mu \sim c = 1/a$.

Например, вероятность P отказа в обслуживании для системы транкинговой связи оценивается по "B-формуле Эрланга а именно:

$$P(K) = \frac{A^K}{K! \left[\sum_{k=0}^{K} \left(\frac{A^k}{k!}\right)\right]}, \quad (1)$$

где A=a*N – нагрузка на систему в ЧНН в Эрлангах, a – средняя нагрузка в ЧНН, создаваемая одним абонентом, N – количество абонентов в зоне обслуживания БС, K – общее количество тайм слотов БС.

На рис.1 показана зависимость вероятности P отказа в обслуживании (вертикальная ось) от количества K трафиковых каналов (горизонтальная ось) для 180 абонентов.

Критериями управления ресурсами как отдельной БС или кластера, так и КС в целом являются:

- максимизация производительности;
- минимизация энергопотребления.

Максимизация производительности достигается благодаря динамической балансировке нагрузки, то есть управления распределением ресурсов: пропускной способности C каналов и интенсивностей обслуживания μ узлов (БС) в реальном времени.

Поскольку беспроводная КС имеет множество узлов (БС), необходимо принимать во внимание, что основным эксплуатационным расходом OPEX таких КС являются затраты на потребление электроэнергии. Поэтому другим критерием служит минимизация расходов на электроэнергию. Это достигается большей загрузкой ресурсов с помощью меньшего количества БС.

Таким образом, это достигаются следующими основными подходами:

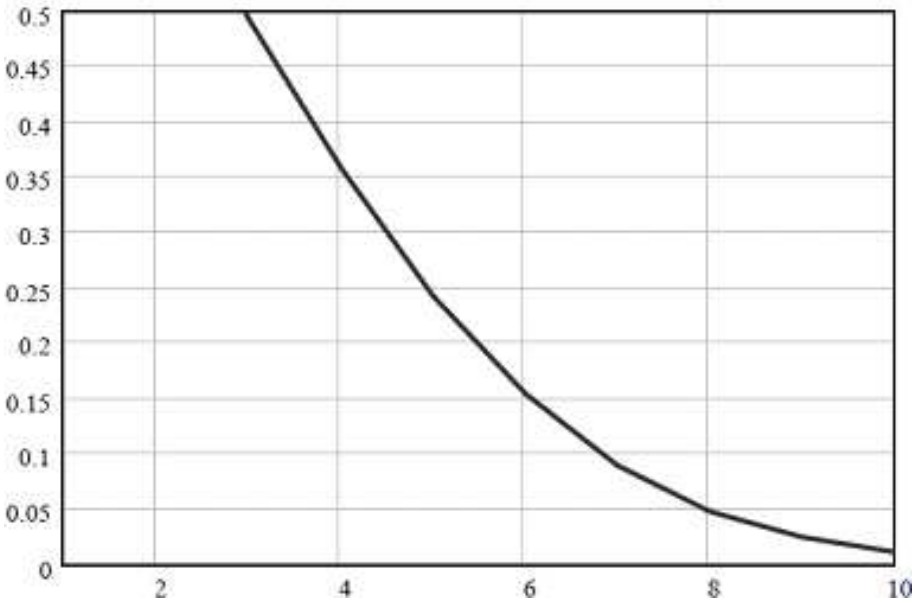


Рис. 1: Зависимость вероятности P отказа в обслуживании от числа K каналов

- балансировкой нагрузки, т.е. распределением ресурсов в реальном времени;
- энергосбережение, т.е. минимизация потребления электроэнергии.

2.3. Модели, методы и методики. . Для решения перечисленных задач используются следующие модели и методики:

- методика сбора исходных данных;
- декомпозиция КС на узлы, каналы, кластеры, виды услуг;
- модели элементов КС (узлов, каналов);
- модели трафика (данные, аудио, видео);
- методика расчета и оценка производительности элементов КС (узлов, каналов, кластеров);
- балансировка нагрузки (с целевой функцией энергосбережение);
- модель рабочего режима функционирования сети (БС, кластера);
- модель переходного режима (например, при включении БС).

В модели переходного режима работы сети используется гиперэкспоненциальное распределение, поскольку это единственная модель с дискретным временем, которая позволяет учитывать случайный характер возникновения запросов к БС [1].

Для рабочего режима функционирования сети используются марковские модели M/M/1 или M/M/m с одним или m обслуживающими приборами. Так же используются известные соотношения для систем массового

обслуживания с конечной очередью и ограниченным числом активных абонентов (пользователей). Тогда максимальное количество пользователей:

$N \approx \mu/\lambda$

с точностью до 1 запроса, или

$N \approx 1/\rho$.

Для расчета и оценки производительности базовых станций и кластеров, как моделей систем с конечным числом источников нагрузки N на входе, используются соотношения Шерра [2, 3]:

$T_0 = (N/\mu)/(1 - p_0) - 1/\lambda$,

где T_0 – среднее время ответа, или время, проведенное пакетом (запросом) в узле (или системе); N – среднее число запросов в системе; μ, λ - интенсивность обслуживания и трафика нагрузки соответственно; p_0 – вероятность того, что в системе нет требований:

$p_0 = 1/\sum_{i=0}^{N} [N!/(N-i)!] (\lambda/\mu)^i$.

Тогда максимальное число запросов в системе (в очереди и на обслуживании):

$N_{max} = \mu / \lambda + 1$,

где N_{max} – максимальное количество источников нагрузки, т.е. пользователей, находящихся в очереди и на обслуживании.

$N_{max} \approx 1/\rho$,

поскольку $\rho = \lambda / \mu$.

3. Заключение

Можно сделать следующие выводы.

- Учитывая закон сохранения времени ожидания, можно оперативно перераспределять запросы, обслуживая менее приоритетные запросы после более приоритетных.
- Основным параметром, который определяет производительность КС в реальном времени, является временя задержка τ, которая определяется как интервал между моментом готовности запроса (или началом передачи) и временем полного завершения передачи.
- Временная задержка оценивается числом тайм-слотов, затрачиваемых на обработку и передачу запроса.
- Нагрузка выражается общим количеством тайм-слотов, затрачиваемых на обслуживание активных запросов в ЧНН.

ЛИТЕРАТУРА

1. Широков В.Л. Разработка моделей и методов для оценки и выбора параметров мультисервисных систем обмена информацией // Диссертация на соискание ученой степени к.т.н.– М.: МЭИ(ТУ) - 2006.225 с.
2. В.Столлингс. Компьютерные системы передачи данных. – Издательская группа "Диалектика-Вильямс 2002. 928 с.
3. Клейнрок Л. Вычислительные системы с очередями. – М.: Мир, 1979. 588 с.

QUEUEING NETWORK WITH DIFFERENT TYPES CUSTOMERS AND DYNAMIC CHARACTERISTICS

A. Starovoitov
Belarussion State University of Transport, Gomel, Belarus

Abstract

We consider the queueing network with different types of customers and parameters depending on the state of the entire network. Establish a sufficient conditions for the submission of the stationary distribution in the generalized product-form.

СЕТЬ МАССОВОГО ОБСЛУЖИВАНИЯ С РАЗЛИЧНЫМИ ТИПАМИ ЗАЯВОК И ДИНАМИЧЕСКИМИ ХАРАКТЕРИСТИКАМИ

А. Старовойтов
Белорусский государственный университет транспорта, Гомель, Беларусь, astarovoytov@tut.by

Аннотация

Находятся достаточные условия представления стационарного распределения в обобщенной мультипликативной форме для сети массового обслуживания с различными типами заявок и параметрами, зависящими от состояния всей сети.

Ключевые слова: *сеть массового обслуживания, стационарное распределение, мультипликативная форма*

1. Введение

Сети массового обслуживания с матрицей маршрутизации, зависящей от состояния всей сети, рассматривались в [1]. В [2] исследовались сети массового обслуживания с параметрами, зависящими от состояния всей сети. Однако в этой работе интенсивности потока заявок и интенсивности обслуживания заявок в узлах имели специальный вид. В работе [3] рассматривались открытые экспоненциальные сети с параметрами, зависящими от состояния всей сети, и возможностью обхода заявками узла сети. В указанной работе получены условия, накладываемые на решение уравнений трафика, при выполнении которых удается найти стационарное распределение сети. В работе [4] исследованы сети с различными вариантами зависимости параметров от состояния сети, в том числе и сети с

различными классами заявок и параметрами, зависящими от состояния всей сети. Но в отличие от работы [3] в [4] используется другой подход: находятся условия, накладываемые на параметры сети.

В данной работе рассматриваются сети массового обслуживания с различными типами заявок и параметрами, зависящими от состояния всей сети. Для нахождения стационарного распределения используется подход, предложенный в работе [3]. Устанавливаются условия, при которых стационарное распределение сети имеет обобщенную мультипликативную форму.

2. Постановка задачи

Рассмотрим сеть массового обслуживания, которая состоит из N однолинейных узлов и обслуживает заявки R типов. Обозначим через $n_{ir}(t)$ – число заявок типа r в i-м узле в момент времени t, $i = \overline{1,N}$, $r = \overline{1,R}$. Состояние сети в момент времени $t \geq 0$ будем описывать случайным процессом $\mathbf{n}(t) = (n_1(t), n_2(t), \ldots, n_N(t))$, где $n_i(t) = (n_{i1}(t), n_{i2}(t), \ldots, n_{iR}(t))$ – состояние i-го узла.

В узлы сети поступают независимые пуассоновские потоки заявок. При этом, если в момент времени t состояние сети есть $\mathbf{n}$, то интенсивность потока заявок типа r в i-й узел равна $\lambda_{ir}(\mathbf{n})$, т.е. зависит от состояния сети, $i = \overline{1,N}$, $r = \overline{1,R}$.

Если в момент времени t состояние сети есть $\mathbf{n}$, то длительность обслуживания заявки r-го типа в i-м узле имеет показательное распределение с параметром $\mu_{ir}(\mathbf{n})$, т.е. зависит от состояния сети. Длительности обслуживания заявок в узлах независимы между собой, а также не зависят от процесса поступления. Предполагается, что $\mu_{ir}(\mathbf{n}) > 0$, если $n_{ir} \neq 0$. Заявка r-го типа, обслуженная i-м узлом, независимо от других заявок с вероятностью $\pi_{ij}^{rs}(\mathbf{n})$ направляется в j-й узел как заявка типа s, а с вероятностью $\pi_{i0}^{r}(\mathbf{n})$ покидает сеть. Для любого состояния $\mathbf{n}$ выполняется равенство $\sum\limits_{j=1}^{N}\sum\limits_{s=1}^{R}\pi_{ij}^{rs}(\mathbf{n}) + \pi_{i0}^{r}(\mathbf{n}) = 1$.

В силу вышесказанного процесс $\mathbf{n}(t)$ является марковским процессом. Обозначим через X пространство состояний процесса $\mathbf{n}(t)$, а $p(\mathbf{n})$ – его стационарное распределение. Цель работы – найти стационарное распределение $p(\mathbf{n})$ процесса $\mathbf{n}(t)$.

3. Основной результат

Будем предполагать, что матрица $(\pi_{ij}^{rs}(\mathbf{n}))$ при каждом фиксированном $\mathbf{n} \in X$ является неприводимой, $i,j = \overline{1,N}, r,s = \overline{1,R}$. Тогда система уравнений трафика

$$\varepsilon_{ir}(\mathbf{n}) = \lambda_{ir}(\mathbf{n}) + \sum_{j=1}^{N}\sum_{s=1}^{R}\varepsilon_{js}(\mathbf{n})\pi_{ji}^{sr}(\mathbf{n}+e_{js}),\ i = \overline{1,N}, r = \overline{1,R}, \qquad (1)$$

при каждом $\mathbf{n} \in X$ имеет единственное положительное решение $(\varepsilon_{ir}(\mathbf{n}))$.

Стационарное распределение $p(\mathbf{n})$ будем искать в предположении, что выполняется следующее равенство

$$\frac{p(\mathbf{n}+e_{ir})}{p(\mathbf{n})}=\frac{\varepsilon_{ir}(\mathbf{n})}{\mu_{ir}(\mathbf{n}+e_{ir})}, i=\overline{1,N}, r=\overline{1,R}, \mathbf{n}\in X. \tag{2}$$

Тогда для того, чтобы для любого состояния $\mathbf{n}\in X$ выражение $p(\mathbf{n})$ через $p(0)$ не зависело от формы пути, ведущего из состояния $\mathbf{n}$ в 0, должно выполняться следующее условие:

$$\frac{\varepsilon_{js}(\mathbf{n})\varepsilon_{ir}(\mathbf{n}+e_{js})}{\mu_{js}(\mathbf{n}+e_{js})\mu_{ir}(\mathbf{n}+e_{ir}+e_{js})}=\frac{\varepsilon_{ir}(\mathbf{n})\varepsilon_{js}(\mathbf{n}+e_{ir})}{\mu_{ir}(\mathbf{n}+e_{ir})\mu_{js}(\mathbf{n}+e_{ir}+e_{js})}, \tag{3}$$

$$i,j=\overline{1,N}, r,s=\overline{1,R}, \mathbf{n}\in X.$$

Обозначим

$$\rho_{ir}(\mathbf{n})=\frac{\varepsilon_{ir}(\mathbf{n})}{\mu_{ir}(\mathbf{n}+e_{ir})}, \; i=\overline{1,N}, r=\overline{1,R}. \tag{4}$$

Величины $\rho_{ir}(\mathbf{n})$ при выполнении условия (3) можно трактовать как загрузку i-го узла заявками r-го типа.

Используя (4), условие (3) можно переписать в виде

$$\rho_{ir}(\mathbf{n}+e_{js})\rho_{js}(\mathbf{n})=\rho_{ir}(\mathbf{n})\rho_{js}(\mathbf{n}+e_{ir}). \tag{5}$$

Введем функции

$$G_{i,r}(\mathbf{n})=\prod_{k=1}^{n_{ir}}\frac{\varepsilon_{ir}(\mathbf{n}-ke_{ir})}{\mu_{ir}(\mathbf{n}-(k-1)e_{ir})}=\prod_{k=1}^{n_{ir}}\rho_{ir}(\mathbf{n}-ke_{ir}),$$

$$Q_i(\mathbf{n})=\prod_{r=1}^{R}G_{i,r}(\mathbf{n}-n_{i1}e_{i1}-n_{i2}e_{i2}-\ldots-n_{i,r-1}e_{i,r-1}).$$

Справедлива следующая

Теорема 1. *Пусть процесс $\mathbf{n}(t)$ является эргодическим. Если выполняется условие (5), то стационарное распределение процесса $\mathbf{n}(t)$ имеет следующий вид*

$$p(\mathbf{n})=\prod_{i=1}^{N}Q_i(\mathbf{n}-T_1(\mathbf{n})-T_2(\mathbf{n})-\ldots-T_{i-1}(\mathbf{n}))p(0), \tag{6}$$

где

$$T_i(\mathbf{n})=(0,0,\ldots,0,n_{i1},n_{i2},\ldots,n_{iR},0,0,\ldots,0),$$

$$p(0)=\Big(\sum_{\mathbf{n}\in X}\prod_{i=1}^{N}Q_i(\mathbf{n}-T_1(\mathbf{n})-T_2(\mathbf{n})-\ldots-T_{i-1}(\mathbf{n}))\Big)^{-1}.$$

Доказательство. Пусть марковский процесс $\mathbf{n}(t)$ является эргодическим. Тогда у него существует единственное стационарное распределение $p(\mathbf{n})$, которое удовлетворяет следующим уравнениям глобального равновесия

$$\sum_{i=1}^{N}\sum_{r=1}^{R}(\lambda_{ir}(\mathbf{n})+\mu_{ir}(\mathbf{n})I_{\{n_{ir}\neq 0\}})p(\mathbf{n}) = \tag{7}$$

$$= \sum_{i=1}^{N}\sum_{r=1}^{R}\lambda_{ir}(\mathbf{n}-e_{ir})p(\mathbf{n}-e_{ir})I_{\{n_{ir}\neq 0\}}+\sum_{i=1}^{N}\sum_{r=1}^{R}\mu_{ir}(\mathbf{n}+e_{ir})\pi_{i0}^{r}(\mathbf{n}+e_{ir})p(\mathbf{n}+e_{ir})+$$

$$+\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{r=1}^{R}\sum_{s=1}^{R}\mu_{js}(\mathbf{n}+e_{js}-e_{ir})\pi_{ji}^{sr}(\mathbf{n}+e_{js}-e_{ir})p(\mathbf{n}+e_{js}-e_{ir})I_{\{n_{ir}\neq 0\}},$$

где I_A – индикатор события A, $e_{ir}\in X$ – единичный вектор, для которого все координаты n_{js} равны 0, а координата $n_{ir}=1$.

Уравнения глобального равновесия (7) разобъем на уравнения локального равновесия

$$\mu_{ir}(\mathbf{n})I_{\{n_{ir}\neq 0\}}p(\mathbf{n}) = \lambda_{ir}(\mathbf{n}-e_{ir})p(\mathbf{n}-e_{ir})I_{\{n_{ir}\neq 0\}}+ \tag{8}$$

$$+\sum_{j=1}^{N}\sum_{s=1}^{R}\mu_{js}(\mathbf{n}+e_{js}-e_{ir})\pi_{ji}^{sr}(\mathbf{n}+e_{js}-e_{ir})p(\mathbf{n}+e_{js}-e_{ir})I_{\{n_{ir}\neq 0\}}$$

и

$$\sum_{i=1}^{N}\sum_{r=1}^{R}\lambda_{ir}(\mathbf{n})p(\mathbf{n}) = \sum_{i=1}^{N}\sum_{r=1}^{R}\mu_{ir}(\mathbf{n}+e_{ir})\pi_{i0}^{r}(\mathbf{n}+e_{ir})p(\mathbf{n}+e_{ir}). \tag{9}$$

Подставляя (6) в уравнения (8) и учитывая (2), (5), получим уравнения трафика (1). Подставляя (6) в уравнения (9) и учитывая (2), получим следующее следствие уравнений трафика

$$\sum_{i=1}^{N}\sum_{r=1}^{R}\lambda_{ir}(\mathbf{n}) = \sum_{i=1}^{N}\sum_{r=1}^{R}\varepsilon_{ir}(\mathbf{n})\pi_{i0}^{r}(\mathbf{n}+e_{ir}).$$

Таким образом, мы нашли распределение $p(\mathbf{n})$, которое удовлетворяет (8) – (9), а, следовательно, оно будет удовлетворять и (7). Что и требовалось доказать. ■

4. Заключение

Для сети массового обслуживания с различными типами заявок и параметрами, зависящими от состояния всей сети, найдены условия представления стационарного распределения вероятностей состояний. При этом вероятность каждого состояния рекуррентно выражается через вероятности соседних состояний вплоть до состояния, когда все координаты равны

нулю. Указанные условия накладываются на решение системы уравнений трафика и гарантируют независимость вероятностей состояний от формы выбранного пути.

ЛИТЕРАТУРА

1. Абышкин В. А.,Самуйлов К. Е. Метод расчета характеристик сети массового обслуживания с матрицей переходных вероятностей, зависящей от состояния сети // Тр. XII Всесоюз. семинара. по вычислительным сетям. Тез. докл. Одесса, 1987. С. 227-231.
2. Башарин Г. П., Чумаев А. В. Условия частичного и детального баланса для модели гибкой производственной системы // АиТ. 1989(4). С. 109-115.
3. Евдокимович В. Е., Малинковский Ю. В., Сети массового обслуживания с динамической маршрутизацией и динамическими вероятностными обходами узлов заявками // Пробл. передачи информ. 2001(37:3). С. 55-66.
4. Ивницкий В. А. Теория сетей массвого обслуживания. Физматлит, 2004.

SYNERGETIC EFFECTS IN MULTISERVER QUEUING SYSTEMS WITH ALTERNATING INPUT FLOW

G.Sh. Tsitsiashvili[1], *M.A. Osipova*[2]
[1]IAM FEB RAS, FEFU, Vladivostok, Russia
[2]IAM FEB RAS, FEFU, Vladivostok, Russia
guram@iam.dvo.ru, mao1975@list.ru

Abstract

In this paper an aggregation of n oneserver queuing systems with widely used now model of ON-OFF input flows [1, 2] into n - server system is considered. A synergetic effect of a queue disappearance for $n \to \infty$ in the aggregated system based on a convergence of specially normalized ON-OFF input flow to partial Brownian motion is proved.

Keywords: a multiserver queueing system, a synergetic effect, a partial Brownian motion, an ON-OFF input flow.

1. Introduction

Consider n - server queuing system with $n = n(N)$, where N is large parameter: $N \to \infty$. This system has following input flow. Define continuous random flow with ON and OFF periods [1, 2]. Suppose that the sequence of independent and identically distributed random variables (i.i.d.r.v.'s) $X_0 \geq 0,\ X_1 \geq 0, \dots$ consists of ON - periods lengths and the sequence of i.i.d.r.v.'s $Y_0 \geq 0,\ Y_1 \geq 0, \dots$ consists of OFF - periods lengths and these random sequences are independent. Denote $F_1(t) = P(X_0 < t),\ F_2(t) = P(Y_0 < t),\ t \geq 0,\ \overline{F} = 1 - F$, and assume that $\overline{F}_1(t) = t^{-\alpha_1} l_1,\ \overline{F}_2(t) = t^{-\alpha_2} L_2(t),\ 1 < \alpha_1 < \alpha_2 < 2$, where for $t \to \infty$ the function $L_1(t) \to l_1 > 0$ and $L_2(t)$ is slowly varying function.

Introduce independent r.v.'s $B,\ X,\ Y$, which are independent with $X_n,\ Y_n, n \geq 1$, and B has the distribution $P(B{=}1) = \dfrac{\mu_1}{\mu},\ P(B{=}0) = \dfrac{\mu_2}{\mu},\ \mu{=}\mu_1 + \mu_2$, where

$$\mu_1 = EX_0,\ \mu_2 = EY_0,\ P(X{\leq}x) = \frac{1}{\mu_1}\int_0^x \overline{F}_1(s)ds,\ P(Y{\leq}x) = \frac{1}{\mu_2}\int_0^x \overline{F}_2(s)ds.$$

Then random sequence $\{T_n\}$:

$$T_0 = B(X + Y_0) + (1 - B)Y,\ T_n = T_0 + \sum_{i=1}^{n}(X_i + Y_i),\ n \geq 1,$$

creates random ON-OFF process

$$W(t) = BI_{[0,X)}(t) + \sum_{n=0}^{\infty} I_{[T_n, T_n + X_{n+1})}(t),\ t \geq 0.$$

(here $I_A(t)$ is the indicator function of the random event $t \in A$).

Random process $W(t)$ is binary and stationary: $W(t) = 1$ if t contains in ON-period, $W(t) = 0$ if t contains in OFF-period and stationary as $EW(t) = P(W(t) = 1) = \mu_1/\mu = a$. Denote $A(t) = \int_0^t W(s)ds$, then

$$EA(t) = at,\ t \geq 0. \tag{1}$$

Assume that $M = M(N) = [N^\gamma],\ \gamma > 0$, and random functions $A_m(t),\ m = 1, ..., M$, are independent copies of random function $A(t)$. Here $[c],\ \{c\}$ are integer and fractional parts of real number c. Designate

$$A_M(t) = \sum_{m=1}^{M} A_m(t),\ B(N) = [N^{3-\alpha_1} L_1(N) M]^{1/2},\ \sigma^2 = \frac{2\mu_2^2 \Gamma(2-\alpha_1)}{(\alpha_1 - 1)\mu^3 \Gamma(4-\alpha_1)},$$

$$n = n(N) = NM(N). \tag{2}$$

From (1), (2) we have that

$$EA_M(Nt) = aMNt = ant. \tag{3}$$

Quantizatize now continuous random flow described by random function $A_M(Nt)$. For this aim we take r.v. ψ independent with random process $A_M(Nt)$, denote

$$e_n(t) = [A_M(Nt) + \psi]. \tag{4}$$

Contrast to the nondecreasing function $e_n(t)$ the sequence of moments

$$\mathcal{T}_n = \{T_j = \inf(t :\ e_n(t) = j),\ j = 1, 2, \ldots\}$$

which describes arrival moments of customers to such aggregated n - server queuing system.

Using the representation of r.v. x mean Ex through conditional mean Ex/y by r.v. $y : Ex = E(Ex/y)$ we obtain

$$Ee_n(t) = E\left(E[A_M(Nt) + \psi]/A_M(Nt)\right) = E([A_M(Nt)](1 - \{A_M(Nt)\}) +$$

$$+([A_M(Nt)] + 1)\{A_M(Nt)\}) = EA_M(Nt) = ant,\ t \geq 0 \tag{5}$$

and so the input flow intensity to the n-server system equals na.

Assume that the sequence of i.i.d.r.v.'s $\tau_1, \tau_2, \ldots,\ P(\tau_j < t) = F(t)$ describes service times of customers with arrival moments $T_1, T_2, \ldots$ Denote $q_n(t)$ a number of busy servers in this n - server system at time moment $t \geq 0$.

Theorem 1. *Assume that 1)* $\gamma > \alpha_1 - 1$, *2) the following inequality* $\rho = aE\tau_j < 1$ *is true.*

Then in such aggregated n - server queuing system $P\left(\sup_{0 \leq t \leq T} q_n(t) = n\right) \to 0,\ N \to \infty.$

The statement of Theorem 1 means a convergence to zero of a virtual waiting time in such aggregation of n oneserver systems on time segment $[0, T]$ and characterizes a disappearance of a queue in this system for $N \to \infty$.

2. Some theorems on C - convergence of random processes

Denote by $\mathcal{F}_1$ the space of deterministic functions on the segment $[0, T]$ with uniform metric ρ. Designate by $\mathcal{F}$ the set of bounded by unit functionals f defined on $\mathcal{F}_1$ and continuous in the metric ρ. Say that the sequence of random processes $z_n = z_n(t),\ n \geq 1,\ 0 \leq t \leq T$, C - converges to the random process $z = z(t),\ 0 \leq t \leq T$, if for any functional $f \in \mathcal{F}$ we have that $Mf(z_n) \to Mf(z),\ n \to \infty$.

Denote by $\mathcal{D}$ the space of random functions on $[0, T]$, which almost surely (a.s.) have not breaks of second type and designate by C the space of random functions which a.s. are continuous on $[0, T]$. Produce some general and partial conditions of C - convergence [3, 4] (see also [5, Theorem 7B], [6, Theorem 2.3]).

Theorem 2. *If random functions $z(t) \in C,\ z_n(t) \in \mathcal{D},\ n \geq 1$, then for C - convergence $z_n(t) \to z(t),\ n \to \infty$, it is necessary and sufficient that:*
1) for $n \to \infty$ finite dimensional distributions of random functions $z_n(t),\ n \geq 1$, tend to finite dimensional distributions of random function $z(t)$ on some set S which is everywhere dense on $[0, T]$, 2) for any $\varepsilon > 0$

$$\lim_{\delta \to 0} \limsup_{n \to \infty} P(w_{z_n}(\delta) \geq \varepsilon) = 0,\ w_{z_n}(\delta) = \sup_{0 \leq t, s \leq T:\ |t-s| \leq \delta} |z_n(t) - z_n(s)|. \tag{6}$$

Corollary 1. Assume that random functions $z_n'(t) \in \mathcal{D},\ n \geq 1,\ z(t) \in C$ satisfy the following conditions. 1) For $n \to \infty$ the sequence $z_n(t),\ n \geq 1$, C - converges to $z(t)$. 2) There is the sequence of positive numbers $\beta(n),\ n \geq 1$, which converges to zero for $n \to \infty$ and a.s. $\rho(z_n', z_n) \leq \beta(n),\ n \geq 1$. Then the sequence $z_n'(t),\ n \geq 1$, also C - converges to $z(t)$.

Proof. From Theorem 2 for $n \to \infty$ finite dimensional distributions of $z_n(t),\ n \geq 1$, converges to finite dimensional distributions of $z(t)$ on some set S which is everywhere dense on $[0, T]$. Then from the condition 2) we have the convergence of finite dimensional distributions of $z_n'(t),\ n \geq 1$, to finite dimensional distributions of $z(t)$ on the set S. From Theorem 2 for any $\varepsilon > 0$ the condition (6) is true, so a.s. $w_{z_n'}(\delta) \leq w_{z_n}(\delta) + 2\beta(n)$, consequently

$$\lim_{\delta \to 0} \limsup_{n \to \infty} P(w_{z_n'}(\delta) \geq \varepsilon) = 0.$$

■

Assume that $z(t),\ 0 \leq t \leq T$, is Gaussian process, $Mz(t) = 0,\ Mz^2(t) < a < \infty$, then this process is a.s. continuous. Formulate a condition of the following limit relation for the process $z(t)$.

Theorem 3. *If there is positive number K satisfying the inequality*

$$\varepsilon^2(t, t+u) = M(z(t) - z(t+u))^2 \leq Ku,\ 0 \leq t \leq t+u \leq T,$$

then

$$P\left(\sup_{0 \leq t \leq T} z(t) > L\right) \to 0,\ L \to \infty. \tag{7}$$

Proof. In conditions of Theorem 3 minimal number $N(r)$ of balls with the radios r in metric space $([0,T],\varepsilon)$, (here $\varepsilon(t,t+u)$ is half metric on $[0,T]$) which cover the segment $[0,T]$, satisfies the inequality $N(r) \le TKr^{-2}$. Consequently the Dadly integral

$$\Psi(z) = \int_0^z (\ln N(r))^{1/2} dr \le \int_0^z (\ln \max(1, TK - 2\ln r))^{1/2} dr,$$

constructed by the relative entropy $\ln N(r)$ satisfies the conditions: $\Psi(T) < \infty$. From [7], [8, Theorem 1] we have (7). ∎

Theorem 4. *Assume that the sequence of random processes $z_n(t),\ 0 \le t \le T$, C - converges to Gaussian continuous process $z(t)$, the relation (7) is true and the sequence of positive numbers $L_n \to \infty,\ n \to \infty$, then*

$$P\left(\sup_{0\le t\le T} z_n(t) \ge L_n\right) \to 0,\ n \to \infty. \tag{8}$$

Proof. Take arbitrary number $\varepsilon > 0$ and choose such $L(\varepsilon)$, that $P(\sup\limits_{0\le t\le T} z(t) > L(\varepsilon)) < \varepsilon$. From C - convergence $z_n(t) \to z(t),\ n \to \infty$, it is possible to find $n(\varepsilon)$ that for $n > n(\varepsilon)$ the inequality

$$\left| P\left(\sup_{0\le t\le T} z(t) > L(\varepsilon)\right) - P\left(\sup_{0\le t\le T} z_n(t) > L(\varepsilon)\right)\right| < \varepsilon$$

is true and so

$$P\left(\sup_{0\le t\le T} z_n(t) > L(\varepsilon)\right) < 2\varepsilon,\ n > n(\varepsilon).$$

From the convergence $L_n \to \infty,\ n \to \infty$, we may take $n_1(\varepsilon) > n(\varepsilon)$, so that for $n > n_1(\varepsilon)$ the inequality $L_n > L(\varepsilon)$ is true and consequently

$$P\left(\sup_{0\le t\le T} z_n(t) \ge L_n\right) \le P\left(\sup_{0\le t\le T} z_n(t) > L(\varepsilon)\right) \le 2\varepsilon,\ n > n_1(\varepsilon).$$

The relation (8) is proved. ∎

Assume that random process $\zeta(t)$ is defined by the equality

$$\zeta(t) = \sigma \int_0^t \overline{F}(t-u) d\xi_H(u) + \Theta(t),\ 0 \le t \le T.$$

Here $\Theta(t)$ is centered Gaussian process independent with $\xi_H(t)$, which has covariance function $R(t, t+u) = \int_0^t \overline{F}(v+u)F(v)adv$.

Theorem 5. *For $0 \le t \le t+u \le T,\ K = a(2T\overline{f}+1) + \sigma^2 T^{2H-1}(\overline{f}T + 2H)$ the inequalities $E\zeta^2(t) \le \sigma^2 t^{2H} + at,\ E(\zeta(t) - \zeta(t+u))^2 \le Ku$ are true.*

Proof. From [10, Formula (1.1)], [11, Lemma 1] we have that for $0 \le t \le t+u \le T$

$$E\zeta^2(t) = H(2H-1)\sigma^2 \int_0^t \int_0^t |u-v|^{2H-2}\overline{F}(t-u)\overline{F}(t-v)dvdu + \int_0^t \overline{F}(v)F(v)adv \le$$

$$\le 2H(2H-1)\sigma^2 \int_0^t \int_0^u (u-v)^{2H-2}dvdu + at = \sigma^2 t^{2H} + at,$$

$$\varepsilon^2(t,t+u) = E(\zeta(t)-\zeta(t+u))^2 = [R(t,t) + R(t+u,t+u) - 2R(t,t+u)] +$$

$$+\sigma^2 E\left[\int_0^t \bar{F}(t-v)d\xi(v) - \int_0^{t+u} \bar{F}(t+u-v)d\xi(v)\right]^2 =$$

$$= a\left[\int_0^t \overline{F}(v)F(v) + \int_0^{t+u} \overline{F}(v)F(v) - 2\int_0^t \overline{F}(v+u)F(v)\right]dv +$$

$$+\sigma^2 H(2H-1)\left[\int_0^t \int_0^t \overline{F}(t-v)\overline{F}(t-w) + \int_0^{t+u}\int_0^{t+u} \overline{F}(t+u-v)\overline{F}(t+u-w) -\right.$$

$$\left. -2\int_0^t \int_0^{t+u} \overline{F}(t-v)\overline{F}(t+u-w)\right]|v-w|^{2H-2}dvdw \le a\left[2\int_0^t + \int_t^{t+u}\right]\left(\overline{F}(v) - \overline{F}(v+u)\right)dv +$$

$$+\sigma^2 H(2H-1)\left\{\int_0^t \int_0^t \left(\overline{F}(t-w) - \overline{F}(t+u-w)\right) + \int_t^{t+u}\int_0^t + \int_t^{t+u}\int_t^{t+u}\right\}|v-w|^{2H-2}dvdw \le$$

$$\le a(2t\overline{f}+1)u + \sigma^2(t^{2H}\overline{f}u + u^{2H} + uH(t+u)^{2H-1}) \le$$

$$\le a(2t\overline{f}+1)u + \sigma^2(T^{2H}\overline{f}u + uT^{2H-1} + uHT^{2H-1}) \le$$

$$\le a(2t\overline{f}+1)u + \sigma^2(T^{2H}\overline{f} + 2T^{2H-1})u = Ku.$$

■

3. Proof of Theorem 1

Denote

$$B(N) = [N^{3-\alpha_1}L_1(N)M]^{1/2},\ A(N) = \max(B(N), N^{1/2}),\ \sigma^2 = \frac{2\mu_2^2\Gamma(2-\alpha_1)}{(\alpha_1-1)\mu^3\Gamma(4-\alpha_1)},$$

$b(t)$ - the function converse to the function $1/\overline{F}_1(t)$. Prove that the sequence of random processes

$$x_{n(N)}(t) = \frac{e_{n(N)}(t) - Ee_{n(N)}(t)}{B(N)},\ N \ge 1,$$

for $N \to \infty$ C - converges to partial Brownian motion $\xi_H(t)$, $H = (3-\alpha_1)/2$ multiplied by σ.

It is obvious that

$$B(N) \sim l_1^{1/2}N^{(3+\gamma-\alpha_1)/2},\ b(N) \sim (lN)^{1/\alpha_1},\ N \to \infty,$$

and consequently from the condition 1) of Theorem 1

$$A(N) \sim B(N),\ b(MN)/N \to \infty,\ N \to \infty.$$

So [2, Theorem 4] leads to C - convergence of the sequence of random processes

$$\frac{A_M(Nt) - EA_M(Nt)}{B(N)},\ N \geq 1,$$

to the partial Brownian motion $\sigma\xi_H(t),\ N \to \infty$. From the inequalities

$$|e_n(t) - A_M(Nt)| = |[A_M(Nt) + \psi] - A_M(Nt)| \leq 1,\ t \geq 0,$$

we have that

$$\rho\left(\frac{e_n(t) - Ee_n(t)}{B(N)}, \frac{A_M(Nt)}{B(N)}\right) \to 0,\ N \to \infty.$$

Consequently C - convergence of the sequence $x_{n(N)}(t),\ N \geq 1$, to $\sigma\xi_H(t),\ N \to \infty$, is proved.

Denote $q_n^{\infty}(t)$ the number of busy servers at time moment t in queuing system with input flow characterized by the process $e_n(t)$ provided that the system has infinite number of servers. From the condition 2) and [9, Chapter II, § 1, Theorem 1] for $n \to \infty$ the sequence of random processes

$$z_n(t) = \frac{q_n^{\infty}(t) - nQ(t)}{A(N)},\ N \geq 1,$$

C - converges on the segment $[0, T]$ to the process $\zeta(t)$. From Theorem 5 we have that random process $\zeta(t)$ satisfies Theorem 3 conditions and so

$$P\left(\sup_{0 \leq t \leq T} \zeta(t) \geq L\right) \to 0,\ L \to \infty.$$

Consequently from the conditions 2) of Theorem 1

$$P\left(\sup_{0 \leq t \leq T} \zeta(t) \geq \frac{(1 - Q(T))n}{A(N)}\right) \to 0,\ N \to \infty.$$

As for $N \to \infty$ the sequence of random processes $z_n(t)$ C - converges to $\zeta(t)$ on the segment $[0, T]$ and from Theorem 4

$$P\left(\sup_{0 \leq t \leq T} z_n(t) \geq \frac{(1 - Q(T))n}{A(N)}\right) \to 0,\ n \to \infty$$

and so $P\left(\sup_{0 \leq t \leq T} q_n^{\infty}(t) \geq n\right) \to 0,\ N \to \infty$. Using [9, Chapter II, § 1, Theorem 1 end] remark that random events

$$\{q_n(t) < n,\ 0 \leq t \leq T\} = \{q_n^{\infty}(t) < n,\ 0 \leq t \leq T\}.$$

Consequently from the inequality $q_n(t) \le n$ the following relation is true:

$$P\left(\sup_{0\le t\le T} q_n(t) \ge n\right) = P\left(\sup_{0\le t\le T} q_n(t) = n\right) \to 0,\ N \to \infty.$$

Theorem 1 is proved.

4. Conclusion

Analogous methods allow to analyze synergetic effects in aggregated n - server queuing system with periodical arrivals of customers groups.

REFERENCES

1. Heath D., Resnick S., Samorodnitsky G. Heavy tails and long range dependence in on/off processes and associated fluid miodels // Math Oper. Res. 1998. V. 23 (1). P. 145-165.
2. Mikosch T., Resnick S., Rootzen H., Stegeman A. Is network traffic approximated by stable Levy motion or fractional Brownian motion? // Annals of Applied Probability. 2002. V. 12 (1). P. 23-68.
3. Prokhorov Yu.V. Convergence of random processes and limit theorems of probability theory// Probability theory and its applications. 1956. T. 1. Vol. 2. P. 177-238. (In Russian).
4. Skorokhod A.V. Limit theorems for random processes// Probability theory and its applications. 1956. T. 1. V. 1. P. 289-319. (In Russian).
5. Borovkov A.A. Convergence of functional distributions of random processes// UMN. 1972. T. 27. V. 1 (163). P. 3-41. (In Russian).
6. Afanasiev V.I. Random walks and branching processes// Lecture courses NOC. 2007. V. 6. Moscow: Mathematical Institute of RAS. (In Russian).
7. Dmitrovskiy V.A. Condition of boundedness and estimates of maximum distribution for random fields on arbitrary sets// Lectures Academy Sciences of USSR. 1980. T. 253. V. 2. P. 271-274. (In Russian).
8. Lifshits M.A. On distribution of Gaussian process maximum// Probability theory and its applications. 1986. T. 31. V. 1. P. 134-142. (In Russian).
9. Borovkov A.A. Asymptotic methods in queueing theory. Moscow: Nauka, 1980. (In Russian).
10. Unterberger J. Stochastic calculus for fractional Brownian motion with Hurst exponent $H > 1/4$: A rough path method by analytic extension// The Annals of Probability. 2009. V. 37 (2). P. 565-614.
11. Sinai Ya.G. On distribution of maximum of partial Brownian motion// UMN. 1997. Vol. 52 (2). P. 119 - 138. (In Russian). 1998. Vol. 23. P. 145-165.

DEVELOPEMENT OF DOCUMENT-FLOWS HANDLING MECHANISM FOR IMPLEMENTATION OF THE INTERACTION BETWEEN ENTERPRISE DMS

D. Volchkov
Keldysh Institute of Applied Mathematics (RAS), Moscow, Russia

Abstract

In the presented paper is considered the mechanism of automation of information processes of interorganizational interaction in document management systems. As part of this mechanism is designed module receiving the documents, an integrated software package interaction between organizational systems. The module provides automatic transfer details of received documents in the registration card of DMS, verification documet card format, compiling statistical and analytical summaries of documents between organizations. In the case of documents that are not subject to registration it forms corresponding notification, which allows for rapid feedback between the organization and the correspondent.

ПРОЕКТИРОВНИЕ МЕХАНИЗМА ОБРАБОТКИ ДОКУМЕНТНЫХ ПОТОКОВ ДЛЯ РЕАЛИЗАЦИИ ВЗАИМОДЕЙСТВИЯ МЕЖДУ КОРПОРАТИВНЫМИ СДОУ

Д. Волчков
Институт прикладной математики им. М.В. Келдыша РАН, Москва, Россия
vol@keldysh.ru

Аннотация

В данной работе рассмотрен механизм автоматизации информационных процессов межорганизационного взаимодействия в системах документального обеспечения управления. В рамках реализации данного механизма разработан модуль приема документов, интегрированный в программный комплекс взаимодействия между организационными системами. Модуль предусматривает функции автоматического переноса реквизитов поступивших документов в регистрационную карточку СДОУ, верификации формата карточек документов, формирования статистических и аналитических сводок по

документообороту между организациями. В случае выявления документов, не подлежащих регистрации, формируются соответствующие уведомления, что позволяет обеспечить оперативную обратную связь между организацией и корреспондентом.

Ключевые слова: информационные системы, организационное управление, межорганизационный электронный документооборот

1. Введение

В результате активного внедрения технологий безбумажного документооборота в крупных территориально-распределенных организациях все актуальнее становится задача обеспечения взаимодействия между системами документального обеспечения управления (СДОУ) отдельных подразделений. В качестве реализации взаимодействия в работе [1] была предложена общая структура программного комплекса обмена документами между действующими СДОУ организаций. В данной работе рассматривается разработка модуля автоматизированного приема документов, поступающих по системе межорганизационного документооборота, который необходим для повышения оперативности и эффективности обработки документных потоков в программном комплексе взаимодействия. Модуль упрощает процедуру регистрации поступивших документов, а также обеспечивает оперативную обратную связь между организацией и корреспондентом.

2. Постановка задачи

Основной задачей разрабатываемого модуля является прием документов, поступивших через систему взаимодействия, для их дальнейшей регистрации в СДОУ организации. Поступившие электронные документы загружаются в базу данных обмена и состоят из карточки документа в формате XML, содержащей все основные реквизиты для загрузки в СДОУ (номер, данные о корреспонденте, кратком содержании и т.д.) и графического образа оригинального документа в формате PDF или TIFF. В случае, если обнаружены ошибки в формате электронного документа или документ был ранее получен на бумажном носителе и уже зарегистрирован в СДОУ, модуль должен сформировать для корреспондента уведомление об отказе в регистрации. В ответ на уведомление, адресат отправляет в организацию квитанцию, подтверждающую его получение. Таким образом, модуль приема документов должен обеспечивать выполнение следующих функций:

- прием электронных документов и квитанций, поступивших из других организаций и их отображение в виде списка;
- поиск документа по основным реквизитам;
- просмотр всех реквизитов и образа поступившего документа;
- переход в режим регистрации входящего документа;
- формирование уведомления об отказе в регистрации.

Для выполнения описанных функций, в составе модуля должен быть реализован удобный пользовательский интерфейс.

3. Функциональная схема

На рис. 1 представлена функциональная схема модуля приема документов, интегрированного в программный комплекс взаимодействия между корпоративными СДОУ.

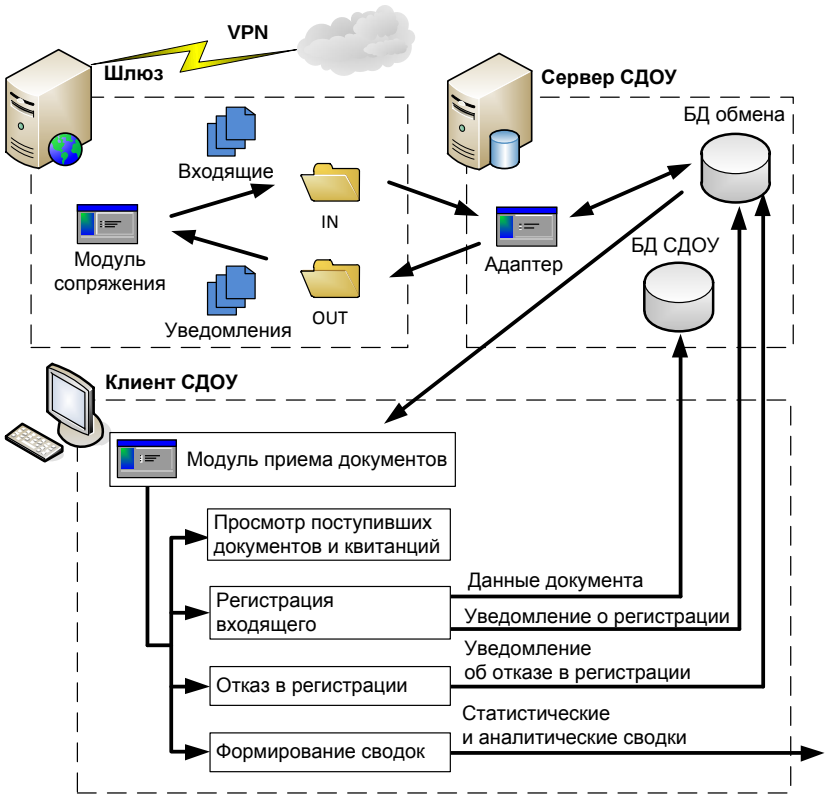


Рис. 1: Функциональная схема модуля приема документов.

Модуль приема документов расположен на клиентской стороне системы. Помимо него в состав программного комплекса взаимодействия входят следующие элементы. Шлюз - программно-аппаратный комплекс, на вход которого через VPN поступают документы из сторонних организаций. БД СДОУ - база данных СДОУ организации. БД обмена - БД, в которой сохраняются все поступившие на шлюз документы, уведомления, и квитанции

а также документы, уведомления и квитанции, выгруженные из СДОУ организации для отображения журналов и формирования статистических и аналитических сводок. Адаптер - загружает из папки IN в БД обмена данные документов, уведомлений и квитанций, поступивших на шлюз; извлекает из БД обмена и отправляет исходящие документы, уведомления и квитанции. Разработанный модуль на основании данных БД обмена формирует и выводит на экран список поступивших документов и квитанций. В ходе формирования списка, модуль анализирует формат каждого электронного документа и в случае обнаружения в нем ошибок, автоматически формирует уведомление об отказе в регистрации. Если ошибок формата не обнаружено, модуль проверяет, не зарегистрирован ли данный документ в СДОУ. В случае, если документ уже присутствует в БД СДОУ, формируется уведомление об отказе в регистрации. В результате пользователь получает список документов, доступных для регистрации, может просмотреть все заполненные реквизиты XML-карточки документа, а также образ документа. Если по каким-либо причинам документ не соответствует требованиям инструкции по делопроизводству (направлен не по принадлежности, образ нечитаемый или не содержит подписи и т.п.), пользователь может вручную сформировать уведомление об отказе, указав причину. Переход в режим регистрации осуществляется по соответствующей команде пользователя. При этом на экране отображается регистрационная карточка документа, в которую автоматически переносятся необходимые реквизиты из XML-карточки документа, а также присоединяется образ документа. После успешного завершения регистрации данные документа заносятся в БД СДОУ, а для корреспондента формируется соответствующее уведомление, содержащее входящий номер и дату документа. Сформированные уведомления записываются в БД обмена и отправляются адресатам посредством адаптера. В состав модуля также входит блок формирования статистических и аналитических сводок, который позволяет получать справки по документообороту между организациями.

4. Реализация пользовательского интерфейса

Пользовательский интерфейс модуля реализован в виде экрана журнала поступивших документов, представленного на рис. 2. В интерфейсе предусмотрена возможность подбора документов по номеру и дате поступления. В верхней части экрана расположены основные функциональные кнопки - печати списка, регистрации или отказа в регистрации документа. В нижней части экрана отображается краткое содержание выбранного документа. Контекстное меню позволяет просмотреть реквизиты документа и его образ. Список поступивших документов может быть отсортирован по любой из колонок, значение поля кСтатусь дублируется цветом: белый - документ доступен для регистрации, синий - документ зарегистрирован, зеленый - отказано в регистрации, красный - ошибка формата XML-карточки документа.

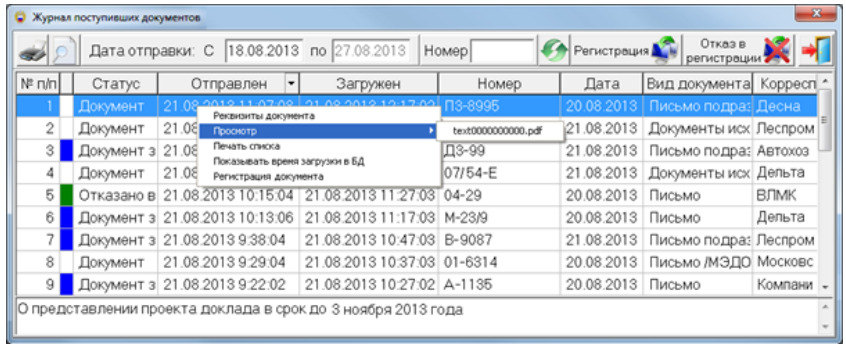


Рис. 2: Экран журнала поступивших документов.

5. Заключение

В результате работы спроектирован и реализован модуль приема электронных документов. Разработанный модуль позволяет получать статистические и аналитические данные о поступивших электронных документах, а также существенно упрощает процедуру регистрации электронных документов за счет автоматического переноса реквизитов в регистрационную карточку. В случае выявления документов, не подлежащих регистрации, формируются соответствующие уведомления, что позволяет обеспечить оперативную обратную связь между организацией и корреспондентом.

ЛИТЕРАТУРА

1. Волчков Д.В. Разработка структуры программного комплекса обмена документами между корпоративными СДОУ // Distributed Computer and Communication Networks: Control, Computation, Communications (DCCN-2013), Moscow: JSC TECHNOSPHERA, 2013. - 464p. - с.329-331.
2. Баканова Н.Б., Цапаева Ю.А., Гурвиц А.Л., Сурпин В.П. Создание корпоративной системы сбора и анализа информации на основе Web-технологий. Международный семинар "Распределенные компьютерные телекоммуникационные сети". М.: Техносфера, 2005.
3. Волчков Д.В., Сурпин В.П. Разработка механизма синхронизации данных в распределенных системах документационного обеспечения управления // Управление развитием крупномасштабных систем (MLSD-2011): Материалы Пятой международной конференции. Том II. М.: Учреждение Российской академии наук Институт проблем управления им. В.А. Трапезникова РАН, 2011. - с.213-216.
4. Таненбаум Э.С., Ван Стеен М. Распределенные системы. Принципы и парадигмы. С-Пб.: Питер, 2003. - 877 с.

To a question of management of TCP in a wireless mesh network on the basis of neural networks

Vorobev V.M.
Research & Development Company «Information and Networking Technologies», Moscow, Russia

Abstract

For effective work of TCP in a wireless cellular network has to undergo changes. The arising problems when using classical TCP and the prerequisite of use of additional management of the protocol on the basis of a neural network are discussed.

К ВОПРОСУ ОБ УПРАВЛЕНИИ TCP В БЕСПРОВОДНОЙ ЯЧЕИСТОЙ СЕТИ НА БАЗЕ НЕЙРОННЫХ СЕТЕЙ

В.М. Воробьев
Научно-производственное объединение «Информационные и сетевые технологии», Москва
svetazarobn@bk.ru

Аннотация

Для эффективной работы TCP в беспроводной ячеистой сети должен претерпеть изменения. Обсуждаются возникающие проблемы при использовании классического TCP и предпосылки использования дополнительного управления протоколом на базе нейронной сети.

1. Введение

Беспроводная ячеистая сеть («wireless mesh network», WMN) представляет собой сеть связи (или передачи данных), состоящую из множества беспроводных узлов, в с двумя по меньшей мере маршрутами для осуществления связи с каждым узлом. Инфраструктура беспроводной ячеистой сети состоит из точек доступа/маршрутизаторов, являющихся узлами сети, соединенных беспроводными каналами связи. Беспроводная ячеистая сеть соединяется с другими проводными сетями, такими как Интернет или корпоративные локальные сети, через ячейку-узел межсетевого шлюза. И для передачи данных используется протокол TCP [1].Он изначально предназначался для работы в проводных сетях, в которых наиболее значима проблема перегруженности буферов в узлах сети.

Стандартные механизмы TCP не позволяют эффективно обнаруживать потери пакетов, не связанные с перегрузками в сети. Механизмы борьбы с потерями пакетов и предотвращения перегрузок в TCP являются ключевой темой многих исследований. С момента разработки протокола TCP было разработано множество алгоритмов борьбы с перегрузками в сети, однако, но не все они были успешными. Ключевыми механизмами борьбы с перегрузками являются: TCP Tahoe, TCP Reno, TCP NewReno, TCP SACK, TCP Vegas и др.

Данные механизмы корректно работают в условиях проводной среды, где потери пакетов происходят из-за перегрузок в сети, полагая, что состояние канала связи постоянное. В беспроводных же сетях связи, где канал является по своей природе нестабильным, потеря данных, как правило, происходит из-за ухудшения характеристик канала. Поэтому, механизмы борьбы с потерей пакетов, эффективно работающие в проводных сетях передачи данных (ПД), для беспроводных сетей во многих случаях являются неэффективными и только замедляют работу.

Например, абонент находится в состоянии хэндовера и пакеты данных не доставляются вследствие того, что на какое-то непродолжительное время связь становится недоступной для данного абонента. В данном случае, стандартные механизмы борьбы с потерями пакетов предполагают, что в сети произошла перегрузка и запускают механизмы борьбы с ней, снижая тем самым скорость передачи данных во много раз. Через некоторое время, оказавшись в радиусе действия другой БС, абонент снова может передавать данные, и ему доступна достаточная полоса пропускания для передачи. Но из-за того что были запущены алгоритмы борьбы с перегрузкой в сети абонент может передавать данные очень небольшими порциями. Вследствие чего, канал оказывается незагруженным, а абонент терпит неудобства из-за медленной передачи данных, хотя ему доступна достаточная полоса пропускания.

Были разработаны TCP, предназначенные специально для работы в беспроводной среде ПД (TCP Westwood, TCP WestwoodNew). Данные механизмы значительно усовершенствованы по сравнению с предыдущими версиями, однако, они также не берут во внимание реальное состояние сети, а лишь прогнозируют ее состояние по динамически изменяющимся параметрам, таким как RTT и др.

Ситуация с использованием транспортного протокола усугубляется в случае беспроводных ячеистых сетей. В беспроводной ячеистой сети узлы могут быть постоянными или мобильными, новые узлы могут быть добавлены на лету и существующие узлы могут выйти из сети. Как следствие, в сети происходит динамическое изменение маршрутизации следования передаваемой информации. Кроме того, передача данных часто связана с различными видами потерь пакета.

Применение классического TCP в WMN осложняется тем, что протокол не может отличить случаи потери пакета из-за перегруженности буферов в

узлах от потерь, вызванных ухудшением состояния радиоканала. Поскольку TCP был разработан, исходя из медленно меняющегося значения RTT, его работа существенно ухудшается в беспроводной ячеистой сети.

В ряде проведенных исследований были предложены различные подходы к решению проблемы [2]. Предложено несколько вариантов изменений или дополнений к протоколу TCP. В основном предлагается использовать заранее выбранный фиксированный алгоритм управления поведением TCP, который не отслеживает текущее состояние сети и каналов. Измененный TCP должен отличать две ситуации: перегрузку буферов в узлах и потери пакетов из-за изменений в каналах. Из-за того, что изменения в каналах происходят непрерывно, необходим алгоритм классификатора, который анализировал вновь поступающую информацию, находил в ней закономерности, адаптировался и давал прогноз на ближайшее будущее. Т.е. TCP должен получать уведомление, что потеря пакетов произошла из-за изменения радиоканала, а не переполнения буферов в узлах сети. Наиболее подходящим инструментом для этой цели представляется использование искусственных нейронных сетей (ИНС) [3]. Статистические методы для управления транспортного протокола WMN ,по-видимому, не дадут хороших результатов из-за очень быстро изменяющейся структуры сети и изменения параметров радиолиний.

ИНС же позволяет не только выполнять заранее запрограммированную последовательность действий на заранее определенном наборе данных, но и анализировать вновь поступающую информацию, находить в ней закономерности, адаптироваться и проводить прогнозирование. Работу ИНС можно построить таким образом, чтобы она непрерывно обучалась на основе предыдущих значений.

Необходимо заметить, что по данным [4] для определения оптимального размера пакета в сети, построенной на базе стандарта 802.11b, возникает необходимость непрерывно анализировать до 12 различных параметров среды.

Для непрерывного отслеживания изменения параметров сети алгоритм прогнозирования оптимального размера пакета был построен на базе нелинейной искусственной нейронной сети. При этом, программа написанная на языке C имеет объем 19 кбайт и может функционировать как часть операционной системы, как программная часть в составе оборудования 802.11b и на уровне приложения.

Для сети с более сложными условиями такими, как WMN, предложение об использовании ИНС для управления TCP опубликовано в [5]. С точки зрения ИНС дополнение к TCP решает задачу кластеризации. Предполагается, что каждое наблюдение может относиться только к одному кластеру. Для решения поставленной задачи алгоритм должен работать в on-line режиме. Алгоритмов, предназначенных для решения этих задач, известно сравнительно немного [6]–[8], при этом они реализуют вероятностный

подход на основе рекуррентной оптимизации принятой нечеткой целевой функции.

ЛИТЕРАТУРА

1. Akyildiz, I.F. and Wang, X. and Wang, W., Wireless mesh networks: a survey // Computer Networks Elsevier, 2005.
2. Sajeeb Saha, Uzzal Kumar Acharjee, Md. Tahzib-Ul-Islam. A survey on wireless mesh network and its challenges at the transport layer // International Journal of Computer Engineering & Technology (IJCET), Volume 5, Issue 8, August (2014), pp. 169–177
3. В. Д. Семейкин, А. В. Скупченко. Метод управления компьютерной сетью на базе нейронных сетей // Вестник АГТУ Серия Управление, вычислительная техника и информатика, 2009, С.161–165
4. Nonlinear neural nets smooth WiFi packets eetimes 5/4/2004, `http://www.eetimes.com/document.asp?doc_id=1150104`
5. A. B. Alim Al Islam ,Vijay Raghunathan. iTCP: an intelligent TCP with neural network based end-to-end congestion control for ad-hoc multi-hop wireless mesh networks // Journal Wireless Networks, Volume 21, Issue 2, February 2015, pp.581–610
6. Krishnapuram R., Keller J. M. Fuzzy and possibilistic clustering methods for computer vision // Neural Fuzzy Systems. — 1994. — 12. — P. 133–159.
7. Park D. C., Dagher I. Gradient based fuzzy c-means (GBFCM) algorithm // Proc. IEEE Int. Conf. on Neural Networks. — 1984. — P. 1626–1631.
8. Chung F. L., Lee T. Fuzzy competitive learning // Neural Networks. — 1994. — 7. — є 3. — P. 539–552

ON A THREE-SERVER FINITE QUEUING SYSTEM WITH ORDERED ENTRY AND POISSON ARRIVALS

R. Razumchik[1], *I. Zariadov*[2]

[1] Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, Moscow, Russia
[2] Peoples' Friendship University, Moscow, Russia
rrazumchik@ipiran.ru, izariadov@gmail.com

Abstract

The purpose of this short note is to present some analytical results concerning markovian multi-server queueing system with ordered entry and finite capacity queue at each server. In general such systems serve as mathematical models of different types of conveyor systems used in industry. Here we consider the simplest case of three-server queuing system with ordered entry. Three servers have a finite (of equal size) waiting room capacity and serve customers according to exponential distribution with the same rate. The arrival process is Poisson. Each arriving customer tries to join the queue in front of server 1 first (i.e. 'ordered entry' selection): if the queue is full he tries to join the queue in front of server 2 etc. If all three queues full the customer is lost. Having joined one of the three queues a customer eventually leaves the system after being served. Although the described system fits the general QBD framework, we seek the efficient (not matrix-geometric) method for exact computation of stationary probability that the arriving customer is blocked. It is shown that the blocking probability can be found in two steps: firstly by analyzing corresponding two-server system using generating functions, and secondly by finding inter-overflow time distribution and applying well-known results for finite-capacity $GI/M/1$ queue.

Keywords: queueing system, multi-server, ordered entry, finite capacity

1. Introduction

Finite-capacity queueing system with ordered entry, being one type of overflow models have been a subject of extensive research and find their application in the study of conveyor system. See classical papers [3, 4] and quite recent ones [1, 2, 5] and references therein. There are numerous types of conveyors currently being used in industry but from practical point of view, as mentioned in [6], working conveyor is "*simply a mechanical device on which departures from one system are carried to a distant point at which they become arrivals to another system*". The motivation for studying the system considered in this short note is two-fold. The first one is the application of the model to the analysis of stability conditions of certain class of closed-loop conveyors as those described

in [7]. Without going into the details the stability condition for such systems can be formulated in terms of arrival rate and the stationary probability that the arriving customer is lost. The second motivation is pure theoretical. Although the considered model fits the QBD framework, in [8] it is shown that for the corresponding two-server queueing system with ordered entry one can carry out efficient recursive computation of the whole joint stationary distribution without resorting to matrix approach. Thus was of interest to understand if the methodology in [8] is suitable for arbitrary number of servers. Our research shows that apparently the methodology of [8] fails in the general case. Nevertheless, as one will see it is still applicable to the three-server case.

In what follows we give the detailed description of the system and briefly state the obtained results.

2. Description of the system

Consider a queuing system with ordered input and three servers (labelled by numbers 1, 2 and 3). Each server has a buffer of finite capacity M in front of it. Customers arrive according to a Poisson process at the constant rate λ. Upon arrival a new customer goes to the server no. 1 or, if it is busy, occupies a place in the queue in front of it. If upon arrival a new customer sees the queue in front of the server no. 2 full, it goes to the server no. 2 or, if it is busy, enters its queue. Finally if upon arrival a new customer sees queues in front of the servers no. 1 and no. 2 full, it goes to the server no. 3 or, if it is busy, in its queue; otherwise it leaves the system (it is considered to be lost). Customers from each queue are served according to FIFO discipline and each server serves customers for exponentially distributed times with the same parameter μ. Denote by p_{ijk}, $0 \leq i, j, k \leq M$, the stationary probability of the fact that there are i in the server no. 1 and its queue, j and k customers in the server no. 2 plus its queue and the server no. 3 plus its queue, respectively. By p_0 we denote the stationary probability of the empty system. We are interested in finding an efficient (not matrix-geometric) method to compute the stationary blocking probability, i.e. p_{MMM}.

3. Computation of p_{MMM}

The first step in computation of p_{MMM} is to notice that the evolution of customer contents in server no. 1 and server no. 2 (and their queues) does not depend on the evolution of the customer contents in the server no. 3 and its queue. The opposite statement is clearly not true due to the ordered-entry discipline. Thus one can find the joint stationary distribution q_{ij} of number of customers in the server no. 1 plus its queue and server no. 2 plus its queue by simply analysing two-server ordered entry queue with the same capacity constraint M, service rate μ and arrival rate λ. The efficient (not matrix-geometric) method for recursive computation of q_{ij} is presented in [8]. Obviously the obtained probabilities q_{ij} are equal to probabilities p_{ijk} summed over k, i.e. $\sum_{k=0}^{M} p_{ijk} = p_{ij\cdot} = q_{ij}$. By "$\cdot$" we further denote the

summation over all possible values of the discrete argument. Strictly speaking, the validity of this equality can also be seen from the fact that probability generating function (PGF) $P(u,v,z) = \sum_{i=0}^{M}\sum_{j=0}^{M}\sum_{k=0}^{M} p_{ijk}u^i v^j z^k$, when $z = 1$, coincides with the PGF of the sequence q_{ij} derived in [8].

The next step is the calculation of inter-overflow time distribution in the two-server system. Due to the exponentially distributed service times and Poisson arrivals, the overflow process from the two-server system is a renewal process and its distribution in terms of Laplace-Stieltjes transform, say $A(s)$, can be found by solving system of linear equations. Due to the fact that the times between overflows in the two-server system are the inter-arrival times to server no. 3., server no. 3 and its queue can be viewed as $GI/M/1/M-1$ system with the Laplace-Stieltjes transform of inter-arrival times given by $A(s)$.

The final step is the computation of loss probability π in $GI/M/1/M-1$ queue, which can be done, using Miyazawa's result in [9]. What is left to be noticed, is that $p_{MMM} = \pi p_{MM\cdot}$.

Acknowledgments.

This work was supported by the Russian Foundation for Basic Research (grant 15-07-03007, 13-07-00223).

REFERENCES

1. Karlof J.K., Jenkins J. The behavior of a multichannel queueing system under three queue disciplines // http://people.uncw.edu/karlof/publications/jenkins.pdf, 2002.
2. Isguder H. O., Celikoglu C.C. Minimizing the loss probability in GI/M/3/0 queueing system with ordered entry // Scientific Research and Essays Vol. 7(8), 2012. P. 963-968.
3. Disney R.L. Some multichannel queueing problems and ordered entry // J. Industrial Eng. 13, 1962. P. 46-48.
4. Elsayed E.A., Proctor C.L. Ordered entry and random choice conveyors with multiple Poisson input // Int. J. Prod. Res. 15, 1977. P. 439-451.
5. Masayuki M. Manufacturing and Service Enterprise with Risks // International Series in Operations Research and Management Science, Vol. 125, 2009, XI, 265 p
6. Phillips D.,Skeith R. Ordered Entry Queueing Networks with Multiple Servers and Multiple Queues // AIEE Transactions, 1969. Vol. 1, Iss. 4.
7. Nanwijn W.M. On a two-server finite queueing system with ordered entry and deterministic arrivals // Euro. J. Opnl. Res. 18, 1984. P. 388-395.
8. Zaryadov I., Meykhanadzhyan L., Milovanova T., Razumchik R. On the method of calculating the stationary distribution in the finite a two-channel system with ordered input // Systems and Means of Informatics, 2015. Vol. 25. Issue 1. P. 1–14. (in Russian)
9. Miyazawa M. Complementary generating functions for the $M^X/GI/1/k$ and $GI/M^Y/1/k$ queues and their application to the comparison for loss probabilities // Journal of Applied Probability, 1990. Vol. 27. P. 684–692.

THE NEW PRINCIPLES OF ORGANIZATION OF DISTRIBUTED COMPUTING IN LARGE NETWORKS

Yu. Zatuliveter, E. Fishchenko
V.A. Trapeznikov Institute of Control Sciences of Russian Academy of Sciences, Moscow, Russia

Abstract

The general-system problems arising because of heterogeneity of large computer networks at the hardware, program and system levels are analyzed. Influence of the growing complexity of a heterogeneous computer environment on problems of information security in three aspects is considered: technical (computer cybersecurity), social (impact of avalanche growth of information) and personal (on the example of a phenomenon of digital dementia). The root causes of heterogeneity of computer information and the computer environment are identified. The principles of the common solution of these problems by formation in resources of large networks of uniform, seamless programmable and cyber security algorithmic space of the distributed computing are offered. Problems and ways of implementation of reliable computing in this space in the conditions of nondeterministic composition of network resources are offered.

НОВЫЕ ПРИНЦИПЫ ОРГАНИЗАЦИИ РАСПРЕДЕЛЕННЫХ ВЫЧИСЛЕНИЙ В БОЛЬШИХ СЕТЯХ

Ю.С. Затуливетер, Е.А. Фищенко
Институт проблем управления им. В.А. Трапезникова РАН, Москва
zvt@ipu.rssi.ru, elena.fish@mail.ru

Аннотация

Анализируются общесистемные проблемы, возникающие из-за разнородности больших компьютерных сетей на аппаратном, программном и системном уровнях. Рассматривается влияние растущей сложности разнородной компьютерной среды на проблемы информационной безопасности в трех аспектах: техническом (компьютерная кибербезопасность), социальном (влияние лавинного роста информации на социальную среду) и личностном (на примере феномена цифровой деменции). Выявляются первопричины разнородности компьютерной информации и компьютерной среды. Предлагаются принципы общего решения этих проблем путем формирования в ресурсах больших сетей единого, бесшовно программируемого и кибербезопасного алгоритмического пространства распределенных вычислений.

Предлагаются пути реализации надежных вычислений в этом пространстве в условиях недетерминированного состава сетевых ресурсов.

Ключевые слова: большие сети, разнородность, интеграция, комбинаторная сложность, информационная безопасность, распределенные вычисления, единое алгоритмическое пространство, бесшовное программирование, кибербезопасность, социальная среда, цифровая деменция, недетерминированная компьютерная среда, надежные вычисления.

1. Введение

Компьютерные сети широко используются для распределенных вычислений и управления большими системами. К таким относятся системы подвижных аппаратов, разнообразных установок, машин, приборов, бытовой и другой техники массового потребления, технологических процессов, интеллектуальных датчиков и т.п. Компоненты таких систем оснащаются встроенными компьютерными средствами управления, которые позволяют программировать и управлять их поведением во всем диапазоне заложенных возможностей. Число компонентов в интегрированных системах достигает сотен, тысяч и более. В перспективе ограничения на количество компонентов должны сниматься. Их расположение может покрывать большие пространства, измеряемые десятками, сотнями, тысячами километров и более.

В соответствии с концепцией Интернет вещей (Internet of Things) в глобальных сетях такие системы могут охватывать сильносвязными взаимодействиями практически неограниченное количество и разнообразие объектов. Дальнейшее ее развитие — концепция Всеобъемлющий Интернет (Internet of Everything) — предполагает интеграцию в сильносвязном взаимодействии, как объектов, так и субъектов социальной среды.

Современный компьютерно-сетевой инструментарий, такой как Grid- и Cloud-технологии, для отдельных классов профилированных комплексов задач позволяют лишь в ограниченных масштабах создавать сетевые системы распределенных вычислений и/или управления того или иного назначения.

Принципиальные недостатки существующих методов и средств создания распределенных систем обработки данных состоят в следующем:

- функциональная и системная интеграция разнородных ресурсов многовариантна, ее комбинаторная сложность становится непреодолимым препятствием для наращивания размеров и масштабов таких систем;
- отсутствие средств «бесшовного» программирования распределенных вычислений сколь угодно больших сетей;
- опережающий рост сложности проблем информационной безопасности;

- необходимость жестких ограничений состава и типов объектов, их функций.

Разрабатываемые с помощью существующих технологий системы распределенной обработки данных в глобальных сетях в большинстве случаев имеют корпоративное назначение. Они строятся посредством дополнительных многослойных программных надстроек над стандартными операционными системами (ОС). В результате получаются разнородные, громоздкие, уязвимые в отношении кибератак программные решения с чрезмерным числом внутренних степеней свободы выведенных на человека. Трудоемкость, а значит, себестоимость и сроки их разработки по мере увеличения масштабов систем имеют тенденцию к опережающему росту. Эксплуатация требует сложного администрирования на многочисленных, крайне разнородных системных и сетевых уровнях.

Такие решения, как правило, трудно адаптируются к структурным изменениям внешнего контекста глобального информационного пространства, тенденции развития которого не подчиняется внутрикорпоративным интересам. В результате длительных сроков разработки таких систем информационная среда, в которой они должны функционировать, может существенно изменить свои свойства, что приводит к утрате актуальности и обесцениванию инвестиций и результатов усилий многих разработчиков.

2. Новые вызовы

Одним их главных факторов деструктивного воздействия разнородной компьютерной среды на мировую социосистему является беспрецедентный феномен глобальной информационной сильносвязности, который можно выразить формулой «всѣ влияет на всѣ и сразу». В существующей компьютерной среде глобализация действий передачи и хранения распределчнной информации не уравновешивается полномасштабной глобализацией свойств еѣ универсально программируемой переработки в целях управления функционированием и устойчивым развитием социосистем.

Несбалансированная глобализация компьютерной среды приводит к экспоненциальному росту интенсивности и объемов слабо переработанной и, потому, плохо организованной информации. Чрезмерные потоки такой информации ведут к неконтролируемому росту глобального «информационного шума». Прежние методы и способы управления функционированием и устойчивым развитием социосистем утрачивают свою действенность. В этом одна из главных причин непрерывной череды глобальных кризисов с нарастающей амплитудой, неподдающихся известным методам политического и финансово-экономического регулирования.

Еще один фундаментальный фактор воздействия — изначально неустранимая разнородность существующих сетевых ресурсов. Системная и функциональная интеграция сетевых ресурсов и программирование распределенных вычислений становятся многовариантными задачами, которые, как известно, обладают комбинаторной сложностью своих решений. Уровни

сложности таких систем быстро растут с увеличением размеров вычислительных сред, что нашло выражение в термине «комбинаторное проклятие» размерности.

Системные свойства существующих разнородных компьютерных сред и технологий их программирования не отвечают требованиям свободной масштабируемости распределенных вычислений и процессов управления, их универсальной и общедоступной (бесшовной) программируемости. Преодоление комбинаторной сложности в ходе интеграции и программирования сетевых ресурсов с увеличением размеров сетей требует неограниченного роста средств и времени.

Увеличение внутрикомпьютерной сложности приводит к практически неконтролируемому росту киберугроз. Существующие программные средства системного уровня (ОС, промежуточное ПО, все уровни сетевых протоколов) отличаются крайней разнородностью и, как следствие, обилием внутренних нестыковок, неучтенных уязвимостей. В ходе их обновления число уязвимостей только возрастает. Защита от вредоносных воздействий посредством программных решений в этих условиях становится все менее эффективной.

Прямым следствием комбинаторной сложности разнородной компьютерной среды становится рост негативного воздействия на здоровье массовых пользователей. Растущая по мере увеличения масштабов применения сетей сложность внутрикомпьютерных проблем неминуемо отражается в интерфейсах массового взаимодействия человека с машиной. Рядовые пользователи вынуждены брать на себя растущую нагрузку по переработке все более интенсивных потоков плохо организованной информации, управлению многочисленными настройками, сервисами. Психологическая нагрузка на пользователей растет и уже превышает природные защитные барьеры.

Количество пользователей при массовом использовании гаджетов исчисляется миллиардами. В результате — новейшие виды массовых заболеваний. Один из примеров — цифровая деменция («digital dementia»). Все больше молодых пользователей страдают потерей памяти, расстройством внимания, когнитивными нарушениями, депрессией, снижением самоконтроля. В мозгу пациентов происходят изменения, схожие с теми, что появляются после черепно-мозговой травмы или на ранней стадии возрастного слабоумия.

Источник новых вызовов — глобальная компьютерная среда, крайне разнородная и несбалансированная. Полномасштабные ответы на них возможны только на путях совершенствования ее фундаментальных системных качеств.

3. Корневые проблемы совершенствования компьютерных сред

Выделим и рассмотрим некоторые из таких проблем.

Первая проблема. Основным инструментом переработки глобально распределенной информации являются системы распределенных вычислений, которые решают задачи с вовлечением большого количества связанных сетями компьютерных устройств различных классов — от «умной пыли», смартфонов и ПК до суперкомпьютеров.

Ведущие технологии в этой сфере — Grid-системы и «облачные» вычисления. Они позволяют интегрировать для осуществления распределенных вычислений от десятков до десятков тысяч компьютерных устройств. Следует отметить, что в условиях разнородности этими технологиями покрывается лишь ничтожная часть совокупного вычислительного потенциала глобальных сетей, что не удовлетворяет опережающий спрос на расширение масштабов переработки экспоненциально растущих потоков глобально распределенной информации.

Вторая проблема. В настоящее время отсутствуют регулярные средства формирования в общедоступных ресурсах глобальных сетей единого, бесшовно программируемого пространства распределенных вычислений. Это является причиной глобального дисбаланса в развитии компьютерной среды. Он проявляет себя в неконтролируемом экспоненциальном росте потоков и объемов крайне разнородной и плохо организованной глобально распределенной информации. Перепроизводство слабо структурированной информации (из-за недостаточной глубины переработки) становится одним из главных барьеров на путях устойчивого функционирования и развития мировой социосистемы в условиях глобальной информационной сильносвязности.

Третья проблема. По мере увеличения масштабов интеграции разнородных сетевых ресурсов системная сложность систем распределенных вычислений растет опережающими темпами. В отсутствие формально строгих методов и средств композиции единого целого из разнородных фрагментов внутренняя структура таких систем обретает вид «лоскутного» одеяла. В отсутствие системной гарантии полноты и непротиворечивости функций в таких системах неизбежно накапливаются неучтенные ошибки и нестыковки. В них на разных уровнях — от аппаратных средств, инструментов программирования, ОС и до сетевых протоколов, неизбежно возникают каналы нелегального доступа к вычислительным ресурсам.

Неконтролируемый рост системной сложности больших систем в разнородных сетях является основой для несанкционированного использования сетевых ресурсов. Поэтому, чем масштабнее системы распределенной обработки, тем сложнее обеспечивать их кибербезопасность. Их защита от несанкционированного вмешательства обретает чрезвычайную остроту.

Существующие способы обеспечения защиты посредством многослойных обновлений (патчей) носят несистемный характер латания дыр. Методы нанесения заплаток на заплатки явно не адекватны темпам катастрофического роста системной сложности компьютерной среды.

4. Принципы построения бесшовно программируемого и кибербезопасного алгоритмического пространства распределенных вычислений

Дальнейшая борьба с системной сложностью все более масштабных систем путем лобового преодоления комбинаторной сложности разнородных программных решений посредством добавления все более дорогостоящих и менее надежных слоев промежуточного ПО в условиях глобальной сильносвязности не имеют долгосрочных перспектив. Сложность разнородных программных системных слоев достигла критических уровней. Качественное развитие компьютерной среды становится практически невозможным — ни в части функциональной интеграции, ни в части обеспечения кибербезопасности.

Необходим общий системный подход к решению проблем сложности разнородной компьютерной среды. Он состоит в следующем:

- выявление и устранение первопричин непрерывного воспроизводства разнородных форм представления компьютерной информации (программ и данных) и способов работы с ней;
- формирование единого, бесшовно программируемого и кибербезопасного алгоритмического пространства распределенных вычислений в сетях.

Первопричины разнородности компьютерной среды кроются в классической аксиоматике универсальной вычислений (модели Дж. фон Неймана). Это однозадачная модель универсальных машинных вычислений, в которой свойство универсальной программируемости замкнуто во внутренних ресурсах компьютера. В ней изначально отсутствует защита памяти от прямого вмешательства одной программы со стороны другой (незащищенное адресное пространство оперативной памяти). Многозадачность и другие системные функции управления вычислительными ресурсами привносится программно — на уровне ОС. Работа сетей также опирается на системные программы, обеспечивающие реализацию сетевых протоколов обмена данными. В чрезмерной сложности разнородных системных программ кроются причины уязвимости современных компьютерных систем и сетей.

В рамках нового подхода [1] в классической модели были выявлены первопричины непрерывного воспроизводства разнородных форм представления компьютерной информации (данных и программ). Они скрыты в постулатах универсального машинного счета классической модели Дж. фон Неймана в виде двух *избыточных степеней свободы управления вычислениями*, открытых для нерегламентированного вмешательства программистов. В классической модели компьютеров, лежащей и в основе микропроцессорных архитектур, выбираемые программистами структуры данных формируются потоками адресов к оперативной памяти. Эта модель позволяет программистам, во-первых, *произвольным образом выстраивать*

структуры данных и, во-вторых, *по собственному усмотрению алгоритмически кодировать* их в потоках адресов. Следовательно, в управлении машинным счетом имеются две степени свободы, открытых программистам. В этих степенях свободы скрываются причины неконтролируемого воспроизводства разнородных форм представления данных и программ.

Чрезмерное разнообразие трудно совместимых форм представления данных и программ с точки зрения алгоритмической универсальности является заведомо избыточным. Оно становится одним из главных факторов роста кинформационного шумань и комбинаторно растущей сложности создания и интеграции больших систем распределенной обработки данных.

В [1] посредством математического обобщения классической модели вычислений проведена ее *минимальная коррекция*, которая предоставляет математически замкнутую форму регламентации структур данных и программ и на уровне обновленной аксиоматики устраняет избыточные степени свободы, а вместе с ними и первопричины комбинаторного сопротивления глобальной интеграции. Модель построена в виде компьютерного исчисления древовидных структур. В этом формализме универсальный объект исчисления — деревья (простейшая связная структура из возможных). Они являются математически однородным структурным объектом представления программ и данных.

Предложенное исчисление — это функционально полный и математически замкнутый на множестве двоичных деревьев набор простейших операций произвольного преобразования деревьев, который представляет новый компьютерный базис [1]. Он позволяет минимальной коррекцией классической модели устранить обе избыточные степени свободы и при этом сохранить достоинства классической модели — простоту логических правил процедурного управления счетом и способов их реализации, что дает наибольшую простоту аппаратных реализаций компьютеров и их массовую применимость. Ключевой результат: *новый базис позволяет бесшовно распространить свойство универсальной программируемости с внутрикомпьютерных ресурсов на любые совокупности компьютеров связанных сетями.*

Новая модель вычислений составляет аксиоматику глобально распределенных вычислений, которая становится основой для формирования в сколь угодно больших сетях универсального, математически однородного и бесшовно программируемого алгоритмического пространства распределенных вычислений. Важное достоинство нового пространства — наследование функциональных возможностей существующих компьютерных и программных платформ, реализованных в рамках классической аксиоматики Дж. фон Неймана. Это системное качество позволяет использовать в перспективных системах распределенных вычислений существующие наработки.

В новой модели становится возможным эффективный перенос ключевых системных функций, которые сегодня реализуются в ядре ОС, на ап-

паратный уровень [2], что позволяет устранить необходимость в многослойном, крайне разнородном, непомерно раздувшемся и, потому, трудно контролируемом системном ПО (ОС, промежуточное ПО, многослойные сетевые протоколы). Тем самым, открываются возможности для кардинального снижения внутрикомпьютерной и системной сложности, повышения эффективности компьютерной среды в целом и выведения ее на качественно новые уровни кибербезопасности.

5. Аппаратный уровень обеспечения кибербезопасности

Обеспечение кибербезопасности — это не только защита от зловредных программ, но и, прежде всего, борьба с причинами чрезмерной системно-технической сложности, которая приводит к утрате контроля над внутренними степенями свободы больших систем и росту уязвимостей компьютерной среды.

Бесшовно программируемое алгоритмическое пространство дает стратегическое направление качественного совершенствования глобальной компьютерной среды. Особенность такого пространства — разнесение системных и прикладных функций на разные уровни.

В новом компьютерном базисе прикладные функции обработки распределенной информации обеспечивают произвольные алгоритмические преобразования сколь угодно больших древовидных структур, компоненты которых размещаются в оперативной памяти компьютеров, связанных сетями. При этом ключевые системные функции, образующие системный базис управления машинными ресурсами, предлагается реализовать на аппаратном уровне посредством универсальных сетевых компьютеров с немикропроцессорной архитектурой [2]. Этим обеспечивается как высокая эффективность управления машинными и сетевыми ресурсами, так и абсолютная защита от нелегального вмешательства на системные уровни со стороны прикладных программ. Защита достигается тем, что со стороны прикладных программ отсутствуют каналы нелегального доступа (в обход математически замкнутых операций исчисления древовидных структур) к внутренним (аппаратно реализованным) механизмам системных уровней.

В этом состоит стратегия кардинального решения вопросов кибербезопасности компьютерной среды. Аппаратная реализация системных функций в рамках математически замкнутого компьютерного базиса не имеет «неучтенных» степеней свободы и «скрытых функций» управления машинными ресурсами, что позволяет устранить неконтролируемые каналы внешнего несанкционированного проникновения.

6. Пути решения проблем организации надежных вычислений

Недетерминированность ресурсов глобальных сетей определяется неустранимыми причинами изменения состава дееспособных компьютеров и

связей между ними. Она характеризуется множеством разноплановых проявлений нестационарности текущего состава и состояния компьютерных устройств и сетевых коммуникаций. Отметим следующие [3]:

- случайным образом меняющийся состав и конфигураций связи компьютеров ведущих вычисления (включение/ выключение, отказы);
- внешние (стихийные или спланированные) деструктивные воздействия на привлеченные к распределенным вычислениям сетевые ресурсы.

Наличие избыточных связей между компьютерами сетей дает необходимый резерв для программных методов повышения надежности распределенных вычислений в сетях в условиях их неустранимой недетерминированности.

Универсальная программируемость единого алгоритмического пространства распределенных вычислений открывает возможности повышения надежности распределенных вычислений за счет автоматического внесения на уровне прикладных программ (на этапах их трансляции или интерпретации) вычислительной избыточности. Она реализуется посредством привлечения к вычислениям (через сети) работоспособных компьютерных устройств.

Повышение надежности распределенных вычислений обеспечивается путем одновременного выполнения многих копий программ с идентичными начальными состояниями и входными данными на разных компьютерных устройствах. Надежность достигается посредством сравнения в контрольных точках и мажорирования текущего состояния процессов, исполняемых на различных компьютерах. Выполнение процессов, текущее состояние которых отличается от состояния большинства аналогичных процессов, переносится с отказавших компьютеров на дееспособные (модель резервирования с заменой) [3]. Исследования показали экспоненциальный характер роста надежности при линейном увеличении количества идентичных процессов.

7. Заключение

Глобальная компьютерная среда в современном своем состоянии становится не столько инструментом развития, сколько генератором беспрецедентных вызовов и глобальных кризисов, на которые пока не существует методов и средств полномасштабных ответов. Это и экспоненциальный рост потоков плохо организованной информации, превышающих пропускную способность человека и социумов в части их переработки, и трудно разрешимые в рамках существующих компьютерных принципов и технологий проблемы интеграции и информационной безопасности.

Путь к качественному обновлению глобальной компьютерной среды — формирование в ней единого алгоритмического пространства на основе новых принципов организации распределенных вычислений.

ЛИТЕРАТУРА

1. Затуливетер Ю.С. Проблемы глобализации парадигмы управления в математически однородном поле компьютерной информации // Проблемы управления. 2005. №1. Ч. I.С. 1-12; №2. Ч. II. С. 13-23.
2. Затуливетер Ю.С., Фищенко Е.А., Семенов С.С. Принципы формирования универсального алгоритмического пространства распределенных и параллельных вычислений на основе немикропроцессорных компьютерно-сетевых архитектур // Вестник компьютерных и информационных технологий. 2013. №6. С. 3-10.
3. Затуливетер Ю.С., Топорищев А.В., Фищенко Е.А., Ходаковский И.А. Принципы реализации моделей повышения надежности распределенных вычислений в системе программирования ПАРСЕК // Датчики и системы. 2009. №12. С. 11-16.