DE GRUYTER OPEN

Int. J. Appl. Math. Comput. Sci., 2014, Vol. 24, No. 3, 485–501
DOI: 10.2478/amcs-2014-0036

amcs

ANALYSIS OF AN $MMAP/PH_1, PH_2/N/\infty$ QUEUEING SYSTEM OPERATING IN A RANDOM ENVIRONMENT

CHESOONG KIM [*], ALEXANDER DUDIN [**], SERGEY DUDIN [**], OLGA DUDINA [**]

[*] Department of Business Administration
Sangji University, Wonju, Kangwon, 220-702, Korea
e-mail: dowoo@sangji.ac.kr

[**]Department of Applied Mathematics and Computer Science
Belarusian State University, 4, Nezavisimosti Ave., Minsk, 220030, Belarus
e-mail: dudin@bsu.by, dudin85@mail.ru, dudina_olga@gmail.com

A multi-server queueing system with two types of customers and an infinite buffer operating in a random environment as a model of a contact center is investigated. The arrival flow of customers is described by a marked Markovian arrival process. Type 1 customers have a non-preemptive priority over type 2 customers and can leave the buffer due to a lack of service. The service times of different type customers have a phase-type distribution with different parameters. To facilitate the investigation of the system we use a generalized phase-type service time distribution. The criterion of ergodicity for a multi-dimensional Markov chain describing the behavior of the system and the algorithm for computation of its steady-state distribution are outlined. Some key performance measures are calculated. The Laplace–Stieltjes transforms of the sojourn and waiting time distributions of priority and non-priority customers are derived. A numerical example illustrating the importance of taking into account the correlation in the arrival process is presented.

Keywords: random environment, marked Markovian arrival process, phase-type distribution, Laplace–Stieltjes transform.

1. Introduction

A contact center is a facility used by companies to manage all client contacts through a variety of mediums such as telephone, fax, letter, e-mail, online live chat, etc. In contrast to call centers, which purely handle telephone correspondence, contact centers have a variety of roles combined to provide all encompassing solutions to clients, and customer contacts. Contact centers, along with call centers and communication centers, all fall under a larger umbrella labeled as the contact center management industry. This is becoming a rapidly growing recruitment sector in itself, as the capabilities of contact centers expand and thus require ever more complex systems as highly skilled operational and management staff.

Efficient operation of a contact center is very important for every company and so a lot of research is being done on optimization of design and work of call centers. For background information and the present state of the art in the study of call centers, the reader is referred to the survey by Aksin *et al*. (2007), the papers of Jouini *et al.* (2011; 2009), Kim and Park (2010), Dudin *et al.* (2013b) and the references therein.

The models of call centers in the overwhelming majority of existing papers assume that the arrival flow of customers is described by a stationary Poisson arrival process. This assumption greatly simplifies the study of systems but at the same time reduces the adequacy of the model, because the arrival flows in modern telecommunication networks and contact centers in particular do not possess the properties of stationarity, memorylessness and ordinarity. Only recently have some papers appeared where the arrival flow is described by the *MMAP* (Marked Markovian Arrival Process) or *MAP* (Markovian Arrival Process). The *MAP* was introduced as a Versatile Markovian Point Process (*VMPP*) by Neuts in the 1970s. The original development of the *VMPP* contained an extensive notation; however, it was simplified greatly by Lucantoni (1991) and ever since this process bears the name Markovian arrival process. The class of *MAP*s includes many input flows considered previously, such as stationary Poisson (*M*), Erlangian

(E_k), hyper-Markovian (*HM*), phase-type (*PH*), or the Markov Modulated Poisson Process (*MMPP*). Generally speaking, the *MAP* is correlated, so it is ideal to model correlated and or bursty traffic in modern telecommunication networks. The *MMAP* is an essential generalization of the *MAP* to the case of heterogeneous customers.

The most similar queueing model with the respect to the one studied in the present paper was recently analyzed by Dudin *et al.* (2013b). This quite a general model has, however, two important shortcomings. The first one consists in the imposed assumption that the service times of both types of customers have an exponential distribution. This assumption drastically decreases the complexity of the mathematical analysis of the model. But this assumption is not good from the practical point of view because it implies that the most probable value of the service time is equal to zero. Actually, the service of a customer's request in a contact center is implemented non-instantaneously. It consists of a random sequence of questions and answers (phases). So, definitely a much more adequate model of the service time is given by the so-called *PH* distribution. For more details about the properties of this distribution and its partial cases, see, e.g., the book by Neuts (1981).

The second shortcoming, typical for an overwhelming majority of the queueing literature, consists in the suggestion that the parameters of the distributions characterizing the arrival and service processes are fixed and do not fluctuate in time. In many real life systems these parameters may change randomly during the system operation due to different external reasons. The intensity of requests processed in the contact center of a bank may essentially depend on a season, day of the week, day time, expected instants of the changes in interest or exchange rates. If the exchange rate of a currency essentially fluctuates during a day, the rate of requests about a current exchange rate has sharp peaks during the periods adjusted to the scheduled instants of the possible changes of the rate. Correspondingly, the service rate may be lower than the average one during these peak times because of customers' complaints about long waiting time.

Some references to the literature about systems operating in a random environment can be found, e.g., in the works of Dudin *et al.* (2013a); Kim *et al.* (2010a; 2013a; 2007; 2010b) or Krieger *et al.* (2005).

In this paper, we consider the queueing model of a contact center where the behavior of the system depends on the current state of the Markovian random process with a finite state space. Under the fixed state of this process (which is called a random environment), the arrival flow is modeled by the *MMAP* and the service time distributions are of the *PH* type. Alternating the states of the random environment implies the corresponding change in the parameters of the arrival and service processes immediately.

The rest of the paper is organized as follows. In Section 2, the mathematical model is described. The multi-dimensional continuous time Markov chain (including the number of customers in the system and the priority customers in a buffer, the state of the random environment, and the states of the underlying *MMAP* and *PH* processes) describing the dynamics of the system is defined in Section 3. In Section 4, a necessary and sufficient condition for ergodicity of this Markov chain is presented and an algorithm for computing the steady state distribution of the Markov chain is outlined. The expressions for the computation of various performance measures of the system are presented in Section 5. Section 6 contains the results of an analysis of the distribution of the waiting and sojourn times of an arbitrary priority customer in the system, and Section 7 is devoted to investigation of the distribution of the waiting and sojourn times of an arbitrary non-priority customer. The results of the numerical experiment giving some insights into the behavior of the system are presented in Section 8. Section 9 concludes the paper.

2. Mathematical model

We consider an N-server queueing system with two types of customers and an infinite buffer as a model of a contact center. The structure of the system under study is presented in Fig. 1.

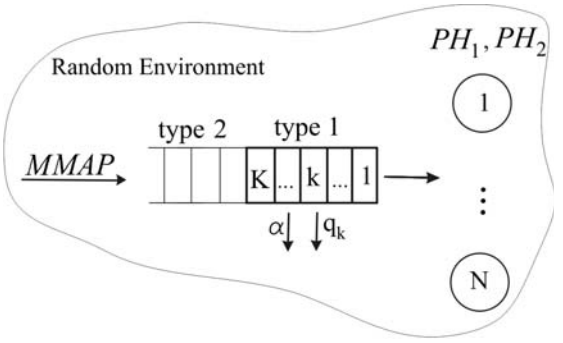


Fig. 1. Structure of the system.

The behavior of the system depends on the state of the random environment. The environment is given by the stochastic process r_t, $t \geq 0$, which is an irreducible continuous time Markov chain with the state space $\{1, 2, \ldots, R\}$ and the infinitesimal generator H.

The customers arrive to the system according to the *MMAP*. That means the following. The arrival of customers is directed by a stochastic process ν_t, $t \geq 0$, with the state space $\{0, 1, \ldots, W\}$. Under the fixed state r of the random environment this process is an irreducible continuous time Markov chain. The sojourn time of this

chain in the state ν is exponentially distributed with a positive finite parameter $\lambda_\nu^{(r)}$. When the sojourn time in the state ν expires, with probability $p_0^{(r)}(\nu,\nu')$, the process ν_t jumps to the state ν' without generation of customers, $\nu,\nu'=\overline{0,W},\ \nu\neq\nu',\ r=\overline{1,R}$. The notation $\nu=\overline{0,W}$ means that the parameter ν takes values in the set $\{0,1,\ldots,W\}$. With probability $p_1^{(r)}(\nu,\nu')$, the process ν_t jumps to the state ν' (probably the same) with generation of a type 1 customer, and with probability $p_2^{(r)}(\nu,\nu')$ the process ν_t jumps to the state ν' with generation of a type 2 customer, $\nu,\nu'=\overline{0,W},\ r=\overline{1,R}$.

The behavior of the *MMAP* under the fixed state of the random environment r is completely characterized by the matrices $D_0^{(r)}$, $D_1^{(r)}$, and $D_2^{(r)}$ defined by the entries

$$(D_0^{(r)})_{\nu,\nu}=-\lambda_\nu^{(r)},\quad \nu=\overline{0,W}, \tag{1}$$

$$(D_0^{(r)})_{\nu,\nu'}=\lambda_\nu^{(r)}p_0^{(r)}(\nu,\nu'),\ \nu,\nu'=\overline{0,W},\quad \nu\neq\nu', \tag{2}$$

$$(D_l^{(r)})_{\nu,\nu'}=\lambda_\nu^{(r)}p_l^{(r)}(\nu,\nu'),\quad \nu,\nu'=\overline{0,W},\quad r=\overline{1,R},\quad l=1,2. \tag{3}$$

The square matrix $D^{(r)}(1)=D_0^{(r)}+D_1^{(r)}+D_2^{(r)}$ of the dimension $\bar{W}=W+1$ represents the generator of the process $\nu_t, t\geq 0$, under fixed $r,\ r=\overline{1,R}$.

The average arrival rate $\lambda^{(r)}$ under the fixed state of the random environment r is given as

$$\lambda^{(r)}=\boldsymbol{\theta}^{(r)}(D_1^{(r)}+D_2^{(r)})\mathbf{e}, \tag{4}$$

where $\boldsymbol{\theta}^{(r)}$ is the invariant vector of a stationary distribution of the Markov chain $\nu_t,\ t\geq 0$, under the fixed state r. The vector $\boldsymbol{\theta}^{(r)}$ is the unique solution to the system $\boldsymbol{\theta}^{(r)}D^{(r)}(1)=\mathbf{0}$, $\boldsymbol{\theta}^{(r)}\mathbf{e}=1$. Here $\mathbf{e}$ is a column vector of the appropriate size consisting of ones and $\mathbf{0}$ is a row vector of appropriate size consisting of zeroes. The average arrival rate $\lambda_l^{(r)}$ of type l customers under the fixed state of the random environment r is defined by $\lambda_l^{(r)}=\boldsymbol{\theta}^{(r)}\,D_l^{(r)}\mathbf{e}$, $l=1,\ 2,\ r=\overline{1,R}$.

The squared coefficient of variation $c_{\rm var}^{(r)}$ of intervals between successive arrivals is given as

$$c_{\rm var}^{(r)}=2\lambda^{(r)}\boldsymbol{\theta}^{(r)}(-D_0^{(r)})^{-1}\mathbf{e}-1. \tag{5}$$

The squared coefficient of variation $c_{\rm var}^{(r,l)}$ of inter-arrival times of type l customers under the fixed state of the random environment r is defined by

$$c_{\rm var}^{(r,l)}=2\lambda_l^{(r)}\boldsymbol{\theta}^{(r)}(-D_0^{(r)}-D_{\bar{l}}^{(r)})^{-1}\mathbf{e}-1,\quad \bar{l}\neq l,\quad \bar{l},l=1,2. \tag{6}$$

The correlation coefficient $c_{\rm cor}^{(r)}$ of two successive intervals between arrivals under the fixed state of the random environment r is given (see Chakravarthy, 2001), as

$$c_{\rm cor}^{(r)}=\frac{1}{c_{\rm var}^{(r)}}(\lambda^{(r)}\boldsymbol{\theta}^{(r)}(-D_0^{(r)})^{-1}(D^{(r)}(1)-D_0^{(r)})\times(-D_0^{(r)})^{-1}\mathbf{e}-1). \tag{7}$$

The correlation coefficient $c_{\rm cor}^{(r,l)}$ of two successive intervals between type l customers' arrivals under the fixed state of the random environment r is computed by

$$c_{\rm cor}^{(r,l)}=\frac{1}{c_{\rm var}^{(r,l)}}(\lambda_l^{(r)}\boldsymbol{\theta}^{(r)}(D_0^{(r)}+D_{\bar{l}}^{(r)})^{-1}\times D_l^{(r)}(D_0^{(r)}+D_{\bar{l}}^{(r)})^{-1}\mathbf{e}-1),\quad \bar{l}\neq l,\ \bar{l},l=1,2. \tag{8}$$

We assume that during the epochs of the transitions of process $r_t,\ t\geq 0$, the states of the process $\nu_t,\ t\geq 0$, do not change, and only the intensities of the transitions of this process do.

We assume that type 1 customers have a non-preemptive priority over type 2 customers. The priority customers are impatient, so they must be serviced as soon as possible, while the non-priority customers can be stored and processed later. Thus, we suggest that a non-priority customer be chosen for service only if there are no priority customers in the buffer during the service completion epoch. Some references to the papers devoted to an analysis of the queues where one type of customers has a time priority while the second one has a space priority are given by Al-Begain *et al*. (2009), who study a single-server system with stationary Poisson inputs. It is worth mentioning that some other practical interpretations (not in terms of contact centers) of the model considered are presented in the above work and the papers cited therein. Al-Begain *et al*. (2006) also suggest the batch Markovian arrival process. However, the buffer space is assumed to be finite for both the types of customers. No analysis of waiting times is presented in the work.

If there is a free server during an arbitrary customer arrival epoch, this customer is admitted to the system and occupies a free server. If all servers are busy during a priority customer arrival epoch and k ($k=\overline{0,K}$), priority customers are present in the buffer, then this customer leaves (balks) the system with probability $q_k^{(r)}$ under the fixed state r of the random environment or moves to the buffer with a complementary probability. If all servers are busy during a priority customer arrival epoch and K priority customers are already present in the buffer, this customer is rejected, i.e., $q_K^{(r)}=1$. Note that a similar mechanism of customers dropping was analysed

for a single server queue with one type of customers by Chydziński and Chróst (2011).

If during a non-priority customer arrival epoch all servers are busy, this customer is placed into the buffer without any limitations. So, if $i \geq 0$ customers are present in the buffer in such a non-priority customer arrival epoch, then this customer is always admitted to the system and placed in the position number $i+1$ in the buffer.

The priority customers can be impatient, i.e., under the fixed state r of the random environment the customer leaves the system due to a lack of service after its arrival in an arbitrary time distributed exponentially with the parameter $\alpha^{(r)}$, $0 < \alpha^{(r)} < \infty$.

The service time of a priority customer by each server under the fixed state r of the random environment has a *PH* distribution with an irreducible representation $(\boldsymbol{\beta}_p^{(r)}, S_p^{(r)})$. This service time can be interpreted as the time until the underlying Markov process $\eta_t^{(1)}, t \geq 0$, with a finite state space $\{1, \ldots, M_1, M_1+1\}$ reaches the single absorbing state M_1+1 conditioned on the fact that the initial state of this process is selected from among the states $\{1, \ldots, M_1\}$ according to the probabilistic row vector $\boldsymbol{\beta}_p^{(r)} = (\beta_{p,1}^{(r)}, \ldots, \beta_{p,M_1}^{(r)})$. The transition rates of the process $\eta_t^{(1)}$ within the set $\{1, \ldots, M_1\}$ are defined by the sub-generator $S_p^{(r)}$ and the transition rates into the absorbing state (which lead to the service completion) are given by the entries of the column vector $\mathbf{S}_{p,0}^{(r)} = -S_p^{(r)}\mathbf{e}$.

The mean service time is calculated as $b_{p,1}^{(r)} = \boldsymbol{\beta}_p^{(r)}(-S_p^{(r)})^{-1}\mathbf{e}$. The squared coefficient of variation is given by $c_{p,var}^{(r)} = b_{p,2}^{(r)}/(b_{p,1}^{(r)})^2 - 1$, where $b_{p,2}^{(r)} = 2\boldsymbol{\beta}_p^{(r)}(-S_p^{(r)})^{-2}\mathbf{e}$.

The service time of a non-priority customer by each server under the fixed state of the random environment r has a PH distribution with M_2+1 states and an irreducible representation $(\boldsymbol{\beta}_n^{(r)}, S_n^{(r)})$ as well as the underlying Markov process $\eta_t^{(2)}, t \geq 0$. The mean service time is calculated as $b_{n,1}^{(r)} = \boldsymbol{\beta}_n^{(r)}(-S_n^{(r)})^{-1}\mathbf{e}$.

We assume that during the epochs of the transitions of the process r_t, $t \geq 0$, the states of the processes $\eta_t^{(1)}$, $\eta_t^{(2)}$, $t \geq 0$, do not change, and only the intensities of further transitions of these processes do.

3. Process of system states

In order to facilitate the investigation of the system, instead of a separate discussion of two phase type distributions of the service time of priority and non-priority customers, we propose to consider one phase type distribution of a generalized service time. We call this distribution a *generalized phase type* distribution with an irreducible representation $(\boldsymbol{\beta}_p^{(r)}, \boldsymbol{\beta}_n^{(r)}, S^{(r)})$.

The generalized service time can be interpreted as a time until the underlying Markov process $\eta_t, t \geq 0$, with a finite state space $\{1, \ldots, M_1, M_1+1, \ldots, M, M+1\}$, where $M = M_1 + M_2$, reaches the single absorbing state $M+1$. The initial state of this process is selected from among the states $\{1, \ldots, M\}$ depending on the type of the customer which is chosen for the service. If a priority customer is chosen for service under the fixed state of the random environment r, the initial state of this process is selected according to the probabilistic row vector $\boldsymbol{\beta}_1^{(r)} = (\boldsymbol{\beta}_p^{(r)}, \mathbf{0}_{M_2})$, and if a non-priority customer is chosen for service, the initial state of this process is selected according to the probabilistic row vector $\boldsymbol{\beta}_2^{(r)} = (\mathbf{0}_{M_1}, \boldsymbol{\beta}_n^{(r)})$. The transition rates of the process η_t within the set $\{1, \ldots, M\}$ are defined by the sub-generator

$$S^{(r)} = \begin{pmatrix} S_p^{(r)} & O \\ O & S_n^{(r)} \end{pmatrix}, \tag{9}$$

and the transition rates into the absorbing state (which lead to the service completion) are given by the entries of the column vector $\boldsymbol{S}_0^{(r)} = -S^{(r)}\mathbf{e}$.

Let i_t be the number of customers in the system, $i_t \geq 0$, r_t be the state of the random environment, $r_t = \overline{1, R}$, k_t be the number of priority customers in the buffer,

$$k_t = \overline{0, \min\{\max\{0, i-N\}, K\}},$$

ν_t be the state of the directing process of the $MMAP$, $\nu_t = \overline{0, W}$, and $\eta_t^{(m)}$ be the number of servers at phase m of service,

$$m = \overline{1, M}, \quad \eta_t^{(m)} = \overline{0, \min\{i_t, N\}},$$

$$\sum_{m=1}^{M} \eta_t^{(m)} = \min\{i_t, N\},$$

during the epoch t, $t \geq 0$. Then the behavior of the system under consideration can be described in terms of the regular irreducible continuous-time Markov chain

$$\xi_t = \{i_t, r_t, k_t, \nu_t, \eta_t^{(1)}, \ldots, \eta_t^{(M)}\}, \quad t \geq 0. \tag{10}$$

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

- I is an identity matrix; $\otimes$ indicates the symbol of the Kronecker product (see, e.g., Graham, 1981).
- $T_i = \binom{i+M-1}{M-1}$, $i = \overline{0, N}$;
- $\tilde{D}_0^{(i)} = \mathrm{diag}\{D_0^{(r)}, r = \overline{1, R}\} \otimes I_{T_i}$, $i = \overline{0, N}$; here $\mathrm{diag}\{D_0^{(r)}, r = \overline{1, R}\}$ means a diagonal matrix having the diagonal blocks $D_0^{(r)}$, $r = \overline{1, R}$;
- $\bar{D}_0^{(k)} = \mathrm{diag}\{I_{k+1} \otimes D_0^{(r)}, r = \overline{1, R}\} \otimes I_{T_N}$, $k = \overline{1, K}$;

- $\tilde{D}_1^{(k)} = \text{diag}\{\tilde{Q}_k^{(r)} \otimes D_1^{(r)}, r = \overline{1,R}\} \otimes I_{T_N}$, $k = \overline{0,K}$;
- $\tilde{Q}_k^{(r)} = \text{diag}\{q_0^{(r)}, q_1^{(r)}, \ldots, q_k^{(r)}\}$, $k = \overline{0,K}$;
- $A_i = \text{diag}\{I_{\bar{W}} \otimes A_i(N, S^{(r)}), r = \overline{1,R}\}$, $i = \overline{0,N}$;
- $\bar{A}_k = \text{diag}\{I_{(k+1)\bar{W}} \otimes A_N(N, S^{(r)}), r = \overline{1,R}\}$, $k = \overline{1,K}$;
- $\Delta_i = -\text{diag}\{I_{\bar{W}} \otimes \text{diag}\{A_i(N, S^{(r)})\mathbf{e} + L_{N-i}(N, \tilde{S}^{(r)})\mathbf{e}\}, r = \overline{1,R}\}$, $i = \overline{1,N}$, $\Delta_0 = O_{\bar{W}R}$.
- $\bar{\Delta}_k = -\text{diag}\{I_{(k+1)\bar{W}} \otimes \text{diag}\{A_N(N, S^{(r)})\mathbf{e} + L_0(N, \tilde{S}^{(r)})\mathbf{e}\}, r = \overline{1,R}\}$, $k = \overline{1,K}$,
- $\tilde{S}^{(r)} = \begin{pmatrix} 0 & \mathbf{0} \\ \boldsymbol{S}_0^{(r)} & S^{(r)} \end{pmatrix}$, $r = \overline{1,R}$;
- $J_k = \text{diag}\{\alpha^{(r)} C_k, r = \overline{1,R}\} \otimes I_{\bar{W}T_N}$, $k = \overline{1,K}$;
- $C_k = \text{diag}\{0, 1, \ldots, k\}, k = \overline{1,K}$;
- $L_{N-i} = \text{diag}\{I_{\bar{W}} \otimes L_{N-i}(N, \tilde{S}^{(r)}), r = \overline{1,R}\}$, $i = \overline{1,N}$;
- $\bar{C}_k = \text{diag}\{E_k^- \otimes I_{\bar{W}} \otimes L_0(N, \tilde{S}^{(r)}) P_{N-1}(\boldsymbol{\beta}_1^{(r)}) + \bar{E}_k \otimes I_{\bar{W}} \otimes L_0(N, \tilde{S}^{(r)}) P_{N-1}(\boldsymbol{\beta}_2^{(r)}), r = \overline{1,R}\}$, $k = \overline{1,K}$;
- E_k^-, $k = \overline{1,K}$, is the matrix of size $(k+1) \times k$ with all zero entries except the entries $(E_k^-)_{l,l-1}$, $l = \overline{1,k}$, which are equal to 1;
- $\bar{E}_k$, $k = \overline{1,K}$, is the matrix of size $(k+1) \times k$ with all zero entries except the entry $(\bar{E}_k)_{0,0} = 1$;
- $\bar{J}_k = \text{diag}\{\alpha^{(r)} C_k E_k^-, r = \overline{1,R}\} \otimes I_{\bar{W}T_N}$, $k = \overline{1,K}$;
- $\tilde{C} = \text{diag}\{I^- \otimes I_{\bar{W}} \otimes L_0(N, \tilde{S}^{(r)}) P_{N-1}(\boldsymbol{\beta}_1^{(r)}) + \bar{I} \otimes I_{\bar{W}} \otimes L_0(N, \tilde{S}^{(r)}) P_{N-1}(\boldsymbol{\beta}_2^{(r)}), r = \overline{1,R}\}$;
- I^- is the square matrix of size $K+1$ with all zero entries except the entries $(I^-)_{l,l-1}$, $l = \overline{1,K}$, which are equal to 1;
- $\bar{I}$ is the square matrix of size $K+1$ with all zero entries except the entry $(\bar{I})_{0,0} = 1$;
- $\tilde{J} = \text{diag}\{\alpha^{(r)} C_K I^-, r = \overline{1,R}\} \otimes I_{\bar{W}T_N}$, $k = \overline{1,K}$;
- $B_i = \text{diag}\{D_1^{(r)} \otimes P_i(\boldsymbol{\beta}_1^{(r)}) + D_2^{(r)} \otimes P_i(\boldsymbol{\beta}_2^{(r)}), r = \overline{1,R}\}$, $i = \overline{0,N-1}$.
- $\bar{D}_1^{(k)} = \text{diag}\{(I - \tilde{Q}_k^{(r)}) E_k^+ \otimes D_1^{(r)}, r = \overline{1,R}\} \otimes I_{T_N}$, $k = \overline{0,K-1}$;
- E_k^+, $k = \overline{0,K-1}$, is the matrix of size $k \times (k+1)$ with nonzero entries $(E_k^+)_{l,l+1}$, $l = \overline{0,k}$, which are equal to 1;
- $\bar{D}_2^{(k)} = \text{diag}\{\tilde{E}_k \otimes D_2^{(r)}, r = \overline{1,R}\} \otimes I_{T_N}$, $k = \overline{0,K-1}$;
- $\tilde{E}_k$, $k = \overline{1,K-1}$, is the matrix of size $k \times (k+1)$ with all zero entries except the entries $(\tilde{E}_k)_{l,l}$, $l = \overline{0,k}$, which are equal to 1;
- $\hat{D}_1 = \text{diag}\{(I - \tilde{Q}_K^{(r)}) I^+ \otimes D_1^{(r)}, r = \overline{1,R}\} \otimes I_{T_N}$;
- I^+ is the square matrix of size $K+1$ with all zero entries except the entries $(I^+)_{l,l+1}$, $l = \overline{0,K-1}$, which are equal to 1;
- $\hat{D}_2 = \text{diag}\{I_{K+1} \otimes D_2^{(r)}, r = \overline{1,R}\} \otimes I_{T_N}$.

A detailed description of the matrices $P_i(\boldsymbol{\beta}_l^{(r)})$, $i = \overline{0,N-1}$, $l = 1, 2$, $A_i(N, S^{(r)})$, $i = \overline{0,N}$, and $L_{N-i}(N, \tilde{S}^{(r)})$, $i = \overline{0,N}$, and the algorithms for their calculation are given by Kim *et al.* (2013b).

Let us enumerate the set of the states of this Markov chain having the fixed value i of the first component i_t in the direct lexicographic order of components (r, k, ν) and the reverse lexicographic order of components $(\eta_t^{(1)}, \ldots, \eta_t^{(M)})$, and refer to this set as a *level* i of the Markov chain.

Let Q be the generator of the Markov chain ξ_t, $t \geq 0$, consisting of blocks $Q_{i,j}$, which define the transition rates of the Markov chain ξ_t, $t \geq 0$, from level i to level j, $i, j \geq 0$.

Theorem 1. *The infinitesimal generator Q of the Markov chain ξ_t, $t \geq 0$, has a block-tridiagonal structure.*

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

$$Q_{i,i} = H \otimes I_{\bar{W}T_i} + \tilde{D}_0^{(i)} + A_i + \Delta_i, \quad i = \overline{0,N-1}, \tag{11}$$

$$Q_{N,N} = H \otimes I_{\bar{W}T_N} + \tilde{D}_0^{(N)} + A_N + \Delta_N + \tilde{D}_1^{(0)}, \tag{12}$$

$$Q_{i,i} = H \otimes I_{(i-N+1)\bar{W}T_N} + \bar{D}_0^{(i-N)} + \bar{A}_{i-N} + \bar{\Delta}_{i-N} + \tilde{D}_1^{(i-N)} - J_{i-N}, \quad i = \overline{N+1, N+K-1}, \tag{13}$$

$$Q_{i,i} = Q_0 = H \otimes I_{(K+1)\bar{W}T_N} + \bar{D}_0^{(K)} + \bar{A}_K + \bar{\Delta}_K + \tilde{D}_1^{(K)} - J_K, \quad i \geq N+K, \tag{14}$$

$$Q_{i,i-1} = L_{N-i}, \quad i = \overline{1,N}, \tag{15}$$

$$Q_{i,i-1} = \bar{C}_{i-N} + \bar{J}_{i-N}, \quad i = \overline{N+1,N+K}, \tag{16}$$

$$Q_{i,i-1} = Q^- = \tilde{C} + \tilde{J}, \quad i > N+K, \tag{17}$$

$$Q_{i,i+1} = B_i, \quad i = \overline{0,N-1}, \tag{18}$$

$$Q_{i,i+1} = \bar{D}_1^{(i-N)} + \bar{D}_2^{(i-N)}, \quad i = \overline{N,N+K-1}, \tag{19}$$

$$Q_{i,i+1} = Q^+ = \hat{D}_1 + \hat{D}_2, \quad i \geq N+K. \tag{20}$$

The proof of Theorem 1 proceeds using the analysis of 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 block matrix form.

4. Ergodicity condition: Calculation of stationary probabilities

It is obvious that the Markov chain $\xi_t, t \geq 0$, belongs to the class of quasi-birth-and-death processes with several boundary levels (cf. Neuts, 1981). It follows from Neuts (1981) that the ergodicity criterion can be expressed in terms of the matrices Q_0, Q^+, Q^- as follows.

Let $\mathbf{y}$ be the left stochastic eigenvector of the matrix $Q_0 + Q^+ + Q^- + I$. It satisfies the equations

$$\mathbf{y}(Q_0 + Q^+ + Q^-) = \mathbf{0}, \quad \mathbf{y}\mathbf{e} = 1. \tag{21}$$

The Markov chain $\xi_t, t \geq 0$, is ergodic if and only if the following inequality holds good:

$$\mathbf{y}Q^+\mathbf{e} < \mathbf{y}Q^-\mathbf{e}. \tag{22}$$

In what follows, we suggest that this condition is fulfilled. Then the following limits (stationary probabilities) exist:

$$\begin{aligned} &\pi(i,r,k,\nu,\eta^{(1)},\dots,\eta^{(M)}) \\ &= \lim_{t\to\infty} P\{i_t = i, r_t = r, k_t = k, \\ &\nu_t = \nu, \eta_t^{(m)} = \eta^{(m)}, \quad m = \overline{1,M}\}, \end{aligned} \tag{23}$$

$$i \geq 0, \quad r = \overline{1,R}, \quad k = \overline{0, \min\{\max\{0, i-N\}, K\}},$$
$$\nu = \overline{0,W}, \quad \eta^{(m)} = \overline{0, \min\{i,N\}},$$
$$\sum_{m=1}^{M} \eta^{(m)} = \min\{i,N\}.$$

We organize the stationary probabilities of the Markov chain into the row vectors $\boldsymbol{\pi}(i,r,k)$ and into the vectors $\boldsymbol{\pi}_i, \ i \geq 0$, according to the introduced enumeration of the states.

It is well known that the vectors $\boldsymbol{\pi}_i, \ i \geq 0$, are defined as a unique solution to the system

$$(\boldsymbol{\pi}_0, \boldsymbol{\pi}_1, \boldsymbol{\pi}_2, \dots)Q = \mathbf{0}, \tag{24}$$

$$(\boldsymbol{\pi}_0, \boldsymbol{\pi}_1, \boldsymbol{\pi}_2, \dots)\mathbf{e} = 1. \tag{25}$$

This system is infinite and its solution is much more difficult in comparison to solving similar finite systems, see, e.g., the system by Olwal *et al*. (2012).

A numerically stable algorithm for the computation of all these vectors, and in particular vectors $\boldsymbol{\pi}_i, \ i = \overline{0,N+K}$, is presented by Klimenok and Dudin (2006). The vectors $\boldsymbol{\pi}_i, \ i \geq N+K$, can be also computed by Neuts' formula,

$$\boldsymbol{\pi}_i = \boldsymbol{\pi}_{N+K}\mathcal{R}^{i-N-K}, \quad i \geq N+K, \tag{26}$$

where the matrix $\mathcal{R}$ is the minimal non-negative solution to the matrix equation

$$Q^+ + \mathcal{R}Q_0 + \mathcal{R}^2Q^- = O. \tag{27}$$

5. Performance measures

As soon as the vectors $\boldsymbol{\pi}_i, \ i \geq 0$, have been calculated, we are able to find various performance measures of the queueing system under consideration.

The stationary distribution of the number of customers in the system is

$$\lim_{t\to\infty} P\{i_t = i\} = \boldsymbol{\pi}_i\mathbf{e}, \quad i \geq 0. \tag{28}$$

The average number of customers in the system is

$$\tilde{L} = \sum_{i=1}^{\infty} i\boldsymbol{\pi}_i\mathbf{e}. \tag{29}$$

The average number of customers in the buffer is

$$N^{\text{buffer}} = \sum_{i=N+1}^{\infty} (i-N)\boldsymbol{\pi}_i\mathbf{e}. \tag{30}$$

The average number of priority customers in the buffer is

$$N_p^{\text{buffer}} = \sum_{i=N+1}^{\infty} \sum_{r=1}^{R} \sum_{k=1}^{\min\{i-N,K\}} k\boldsymbol{\pi}(i,r,k)\mathbf{e}. \tag{31}$$

The average number of non-priority customers in the buffer is

$$\begin{aligned} N_n^{\text{buffer}} &= \sum_{i=N+1}^{\infty} \sum_{r=1}^{R} \sum_{k=1}^{\min\{i-N,K\}} (i-N-k)\boldsymbol{\pi}(i,r,k)\mathbf{e} \\ &= N^{\text{buffer}} - N_p^{\text{buffer}}. \end{aligned} \tag{32}$$

The average number of busy servers is

$$N^{\text{server}} = \sum_{i=1}^{\infty} \min\{i, N\}\boldsymbol{\pi}_i \mathbf{e}. \tag{33}$$

The loss probability of an arbitrary priority customer at the entrance to the system due to the presence of K priority customers in the buffer is

$$P^{\text{ent}-\text{loss}} = \lambda_p^{-1} \sum_{i=N+1}^{\infty} \sum_{r=1}^{R} \boldsymbol{\pi}(i, r, K)(D_1^{(r)} \otimes I_{T_N})\mathbf{e}, \tag{34}$$

where the average arrival rate λ_p of priority customers is calculated as follows:

$$\lambda_p = \boldsymbol{\theta}\text{diag}\{D_1^{(r)},\, r = \overline{1, R}\}\mathbf{e}, \tag{35}$$

and the vector $\boldsymbol{\theta}$ is the unique solution to the following system:

$$\boldsymbol{\theta}(H \otimes I_{\bar{W}} + \text{diag}\{D_0^{(r)} + D_1^{(r)} + D_2^{(r)},\, r = \overline{1, R}\}) = \mathbf{0}, \tag{36}$$

$$\boldsymbol{\theta}\mathbf{e} = 1. \tag{37}$$

The average arrival rate λ_n of non-priority customers is calculated as

$$\lambda_n = \boldsymbol{\theta}\text{diag}\{D_2^{(r)},\, r = \overline{1, R}\}\mathbf{e}. \tag{38}$$

The probability $P^{\text{esc}-\text{loss}}$ that an arbitrary priority customer arrives when all servers are busy, the buffer is not full, and the customer does not join the buffer and leaves the system is given as

$$\begin{aligned} P^{\text{esc}-\text{loss}} &= \lambda_p^{-1} \sum_{i=N}^{\infty} \sum_{r=1}^{R} \sum_{k=0}^{\min\{i-N, K-1\}} q_k^{(r)} \boldsymbol{\pi}(i, r, k) \\ &\quad \times (D_1^{(r)} \otimes I_{T_N})\mathbf{e}. \end{aligned} \tag{39}$$

The intensity of the flow of customers, which obtain service in the system, is

$$\begin{aligned} \lambda^{\text{out}} &= \sum_{i=1}^{N} \boldsymbol{\pi}_i L_{N-i}\mathbf{e} \\ &\quad + \sum_{i=N+1}^{\infty} \boldsymbol{\pi}_i \text{diag}\{I_{(\min\{i-N,K\}+1)} \otimes I_{\bar{W}} \\ &\quad \otimes L_0(N, \tilde{S}^{(r)}),\, r = \overline{1, R}\}\mathbf{e}. \end{aligned} \tag{40}$$

The intensity of the output flow of priority customers is

$$\begin{aligned} \lambda_p^{\text{out}} &= \sum_{i=1}^{N} \boldsymbol{\pi}_i \text{diag}\{I_{\bar{W}} \otimes L_{N-i}(N, \tilde{S}_p^{(r)}),\, r = \overline{1, R}\}\mathbf{e} \\ &\quad + \sum_{i=N+1}^{\infty} \boldsymbol{\pi}_i \text{diag}\{I_{(\min\{i-N,K\}+1)} \otimes I_{\bar{W}} \\ &\quad \otimes L_0(N, \tilde{S}_p^{(r)}),\, r = \overline{1, R}\}\mathbf{e}, \end{aligned} \tag{41}$$

where

$$\tilde{S}_p^{(r)} = \begin{pmatrix} 0 & \mathbf{0} \\ ((\mathbf{S}_{p,0}^{(r)})^T, \mathbf{0}_{M_2})^T & S^{(r)} \end{pmatrix}.$$

The loss probability of an arbitrary priority customer is

$$P^{\text{loss}} = 1 - \frac{\lambda_p^{\text{out}}}{\lambda_p}. \tag{42}$$

The intensity of the output flow of non-priority customers is

$$\begin{aligned} \lambda_n^{\text{out}} &= \sum_{i=1}^{N} \boldsymbol{\pi}_i \text{diag}\{I_{\bar{W}} \otimes L_{N-i}(N, \tilde{S}_n^{(r)}),\, r = \overline{1, R}\}\mathbf{e} \\ &\quad + \sum_{i=N+1}^{\infty} \boldsymbol{\pi}_i \text{diag}\{I_{(\min\{i-N,K\}+1)} \otimes I_{\bar{W}} \\ &\quad \otimes L_0(N, \tilde{S}_n^{(r)}),\, r = \overline{1, R}\}\mathbf{e} \end{aligned} \tag{43}$$

where

$$\tilde{S}_n^{(r)} = \begin{pmatrix} 0 & \mathbf{0} \\ (\mathbf{0}_{M_1}, (\mathbf{S}_{n,0}^{(r)})^T)^T & S^{(r)} \end{pmatrix},$$

$$\mathbf{S}_{n,0}^{(r)} = -S_n^{(r)}\mathbf{e}.$$

Because the non-priority customers are never lost in the system, it is evident that

$$\lambda_n^{\text{out}} = \lambda_n. \tag{44}$$

This relation can be used for the control of the accuracy of computation of the stationary probabilities.

The probability $P^{\text{imp}-\text{loss}}$ that an arbitrary priority customer after arrival will go to the buffer and leave it due to impatience is computed by

$$P^{\text{imp}-\text{loss}} = P^{\text{loss}} - P^{\text{ent}-\text{loss}} - P^{\text{esc}-\text{loss}}. \tag{45}$$

Let us derive an alternative formula for calculation of this probability. To this end, let us introduce the probabilities $x(k, r, \eta^{(1)}, \ldots, \eta^{(M)})$, $k = \overline{1, K}$, that during the waiting time of a priority customer in the buffer this customer does not leave the buffer due to impatience conditioned on the fact that at its arrival epoch there are $k-1$ priority customers in the buffer, the state of the random environment is r, $r = \overline{1, R}$, and the states of the processes $\eta_t^{(1)}, \ldots, \eta_t^{(M)}$ are $\eta^{(1)}, \ldots, \eta^{(M)}$, correspondingly.

Lemma 1. *The column vectors $\mathbf{x}(k)$ consisting of the probabilities $x(k, r, \eta^{(1)}, \ldots, \eta^{(M)})$ enumerated in the reverse lexicographic order of the components $\eta^{(1)}, \ldots, \eta^{(M)}$ and the direct lexicographic order of the component r are given as follows:*

$$\mathbf{x}(1) = (\mathcal{J} - H \otimes I_{T_N} - \mathcal{A})^{-1}\mathcal{L}_p \mathbf{e}_{RK_N}, \tag{46}$$

$$\mathbf{x}(k) = (k\mathcal{J} - H \otimes I_{T_N} - \mathcal{A})^{-1} \times (\mathcal{L}_p + (k-1)\mathcal{J})\mathbf{x}(k-1), \quad k = \overline{2, K}, \tag{47}$$

where

$$\mathcal{J} = \operatorname{diag}\{\alpha^{(r)},\, r = \overline{1, R}\} \otimes I_{T_N}, \tag{48}$$

$$\mathcal{L}_p = \operatorname{diag}\{L_p^{(r)},\, r = \overline{1, R}\}, \tag{49}$$

$$L_p^{(r)} = L_0(N, \tilde{S}^{(r)})P_{N-1}(\boldsymbol{\beta}_1^{(r)}), \tag{50}$$

$$\mathcal{A} = \operatorname{diag}\{A^{(r)},\, r = \overline{1, R}\}, \tag{51}$$

$$A^{(r)} = A_N(N, S^{(r)}) - \operatorname{diag}\{A_N(N, S^{(r)})\mathbf{e} + L_0(N, \tilde{S}^{(r)})\mathbf{e}\}, \quad r = \overline{1, R}. \tag{52}$$

Corollary 1. *The probability $P^{\mathrm{imp-loss}}$ that an arbitrary priority customer will go to the buffer and leave it due to impatience is given by*

$$P^{\mathrm{imp-loss}} = \lambda_p^{-1} \sum_{i=N}^{\infty} \sum_{r=1}^{R} \sum_{k=0}^{\min\{i-N, K-1\}} (1 - q_k^{(r)}) \times \boldsymbol{\pi}(i, r, k)(D_1^{(r)}\mathbf{e} \otimes I_{K_N})(\mathbf{e}_{K_N} - \mathbf{x}(k+1, r)) \tag{53}$$

where $\mathbf{x}(k, r)$ are calculated as the sub-vectors of the vectors $\mathbf{x}(k)$, $k = \overline{1, K}$.

The existence of two different formulas (45) and (53) for computation of the probability $P^{\mathrm{imp-loss}}$ is also useful for controlling the accuracy of computation of the stationary probabilities.

6. Distribution of the sojourn time of an arbitrary priority customer in the system

Let $V_p(x)$ be the distribution function of the sojourn time of an arbitrary priority customer in the system and

$$v_p(s) = \int_0^{\infty} e^{-sx} \mathrm{d}V_p(x), \quad \operatorname{Re} s > 0,$$

be its Laplace–Stieltjes Transform (*LST*).

Let us tag an arbitrary priority customer and keep track of its staying in the system. We will derive the expression for the *LST* $v_p(s)$ using the method of collective marks (method of additional event, method of catastrophes) for reference (see, e.g., Kesten and Runnenburg, 1956; van Danzig, 1955).

To this end, we interpret the variable s as the intensity of some virtual stationary Poisson flow of catastrophes. Thus, $v_p(s)$ has the meaning of the probability that no catastrophe arrives during the sojourn time of the tagged customer.

To derive expression for the *LST* $v_p(s)$, we first need to formulate and prove two auxiliary results.

Let $y_p(s, r, m)$ be the probability that a catastrophe will not arrive during the rest of the tagged customer's service time in the system conditioned on the fact that at the given moment the state of the random environment is r, $r = \overline{1, R}$, and the state of service is m, $m = \overline{1, M_1}$.

Let us form the vectors

$$\mathbf{y}_p(s, r) = (y_p(s, r, 1), \ldots, y_p(s, r, M_1))^T, \quad r = \overline{1, R}, \tag{54}$$

$$\mathbf{y}_p(s) = (\mathbf{y}_p(s, 1), \ldots, \mathbf{y}_p(s, R))^T, \tag{55}$$

and introduce the following notation:

- $\hat{S}_p = \operatorname{diag}\{S_p^{(r)},\, r = \overline{1, R}\}$,
- $\hat{\mathbf{S}}_{p,0} = ((\mathbf{S}_{p,0}^{(1)})^T, \ldots, (\mathbf{S}_{p,0}^{(R)})^T)^T$.

Lemma 2. *The vector $\mathbf{y}_p(s)$ is determined by the formula*

$$\mathbf{y}_p(s) = (-\hat{S}_p - H \otimes I_{M_1} + sI)^{-1}\hat{\mathbf{S}}_{p,0}. \tag{56}$$

Proof. Based on a probabilistic sense of the *LST* and the law of total probability, it can be shown that the probabilities $y_p(s, r, m)$ satisfy the following system of linear algebraic equations:

$$y_p(s, r, m) = (-(S_p^{(r)})_{m,m} - H_{r,r} + s)^{-1} \times \left((\mathbf{S}_{p,0}^{(r)})_m + \sum_{r'=1,\, r' \neq r}^{R} H_{r,r'} y_p(s, r', m) + \sum_{m'=1,\, m' \neq m}^{M_1} (S_p^{(r)})_{m,m'} y_p(s, r, m') \right), \quad r = \overline{1, R}, \quad m = \overline{1, M_1}. \tag{57}$$

Using the notation introduced above, we can rewrite the system (57) in matrix form as

$$(\hat{S}_p + H \otimes I_{M_1} - sI)\mathbf{y}_p(s) = -\hat{\mathbf{S}}_{p,0}, \tag{58}$$

from which the statement of lemma follows immediately because the matrix $\hat{S}_p + H \otimes I_{M_1}$ represents a subgenerator, so the matrix $(\hat{S}_p + H \otimes I_{M_1} - sI)^{-1}$ exists. ■

Let $w_p(s, l, r, \eta^{(1)}, \ldots, \eta^{(M)})$ be the probability that a catastrophe will not arrive during the rest of the tagged customer's sojourn time in the system conditioned on the fact that at the given moment the tagged customer has the position l ($l = \overline{1, K}$) in the buffer, the state of the random environment is r, $r = \overline{1, R}$, and the state of the service processes are $\eta^{(1)}, \ldots, \eta^{(M)}$.

Let us enumerate the probabilities $w_p(s, l, r, \eta^{(1)}, \ldots, \eta^{(M)})$ in the reverse lexicographic order of the components $\eta^{(1)}, \ldots, \eta^{(M)}$, and form from these probabilities column vectors $\mathbf{w}_p(s, l, r)$ and, then, vectors $\mathbf{w}_p(s, l) = (\mathbf{w}_p^T(s, l, 1), \ldots, \mathbf{w}_p^T(s, l, R))^T$.

Lemma 3. *The vectors* $\mathbf{w}_p(s,l)$, $l=\overline{1,K}$, *can be determined by the following recursive formulas:*

$$\begin{aligned}\mathbf{w}_p(s,1) &= (-\mathcal{A}+\mathcal{J}-H\otimes I_{T_N}+sI)^{-1}\\ &\quad\times(\mathcal{L}_p(I_R\otimes\mathbf{e}_{T_N})\boldsymbol{\beta}_p\mathbf{y}_p(s)+\mathcal{J}\mathbf{e}),\end{aligned} \tag{59}$$

$$\begin{aligned}\mathbf{w}_p(s,l+1) &= (-\mathcal{A}+(l+1)\mathcal{J}-H\otimes I_{T_N}+sI)^{-1}\\ &\quad\times(\mathcal{J}\mathbf{e}+(\mathcal{L}_p+l\mathcal{J})\mathbf{w}_p(s,l)),\\ &\quad l=\overline{1,K-1},\end{aligned} \tag{60}$$

where

$$\boldsymbol{\beta}_p=\operatorname{diag}\{\boldsymbol{\beta}_p^{(r)},\ r=\overline{1,R}\}. \tag{61}$$

Proof. Based on a probabilistic sense of the *LST* and the law of total probability, the probabilities $\mathbf{w}_p(s,l,r)$, $l=\overline{1,K}, r=\overline{1,R}$, can be found from the system of linear algebraic equations:

$$\begin{aligned}&\mathbf{w}_p(s,l,r)\\ &=((s+l\alpha^{(r)}-H_{r,r})I_{T_N}-A^{(r)})^{-1}\\ &\quad\times\Bigg(\delta_{l,1}L_0(N,\tilde{S}^{(r)})\mathbf{e}\boldsymbol{\beta}_p^{(r)}\mathbf{y}_p(s,r)\\ &\quad+(1-\delta_{l,1})(L_p^{(r)}+(l-1)\alpha^{(r)}I_{T_N})\mathbf{w}_p(s,l-1,r)\\ &\quad+\sum_{r'=1,\,r'\neq r}^{R}H_{r,r'}\mathbf{w}_p(s,l,r')+\alpha^{(r)}I_{T_N}\Bigg),\end{aligned} \tag{62}$$

where $\delta_{i,j}=0$, if $i=j$, and $\delta_{i,j}=1$, otherwise. This system can be rewritten in the following form:

$$\begin{aligned}&(-sI-l\mathcal{J}+H\otimes I_{T_N}+\mathcal{A})\mathbf{w}_p(s,l)+\delta_{l,1}\mathcal{L}_p\mathbf{e}\boldsymbol{\beta}_p\mathbf{y}_p(s)\\ &\quad+(1-\delta_{l,1})(\mathcal{L}_p+(l-1)\mathcal{J})\mathbf{w}_p(s,l-1)\\ &\quad+\mathcal{J}\mathbf{e}=\mathbf{0}^T,\quad l=\overline{1,K},\end{aligned} \tag{63}$$

from which the statement of the lemma follows immediately because the matrix $-l\mathcal{J}+H\otimes I_{T_N}+\mathcal{A}$ is a subgenerator, and so the inverse matrices in the lemma statement exist.

■

Theorem 2. *The LST* $v_p(s)$ *of the distribution of an arbitrary priority customer's sojourn time in the system is computed by*

$$\begin{aligned}&v_p(s)\\ &=P^{\mathrm{ent-loss}}+P^{\mathrm{esc-loss}}\\ &\quad+\lambda_p^{-1}\Bigg(\sum_{i=0}^{N-1}\sum_{r=1}^{R}\boldsymbol{\pi}(i,r)(D_1^{(r)}\otimes I_{T_i})\mathbf{e}\boldsymbol{\beta}_p^{(r)}\mathbf{y}_p(s,r)\\ &\quad+\sum_{i=N}^{\infty}\sum_{r=1}^{R}\sum_{k=0}^{\min\{i-N,K-1\}}(1-q_k^{(r)})\boldsymbol{\pi}(i,r,k)\\ &\quad(D_1^{(r)}\mathbf{e}\otimes I_{T_N})\mathbf{w}_p(s,k+1,r)\Bigg).\end{aligned} \tag{64}$$

Proof. The proof follows from the law of total probability, a probabilistic sense of the *LST*s, as well as Lemmas 2 and 3. ■

Corollary 2. *The average sojourn time* V_p^{soj} *of an arbitrary priority customer is*

$$\begin{aligned}&V_p^{\mathrm{soj}}\\ &=-v_p'(s)\big|_{s=0}=-\lambda_p^{-1}\Bigg(\sum_{i=0}^{N-1}\sum_{r=1}^{R}\boldsymbol{\pi}(i,r)(D_1^{(r)}\\ &\quad\otimes I_{T_i})\mathbf{e}\boldsymbol{\beta}_p^{(r)}\frac{\partial\mathbf{y}_p(s,r)}{\partial s}\Bigg|_{s=0}\\ &\quad+\sum_{i=N}^{\infty}\sum_{r=1}^{R}\sum_{k=0}^{\min\{i-N,K-1\}}(1-q_k^{(r)})\boldsymbol{\pi}(i,r,k)\\ &\quad(D_1^{(r)}\mathbf{e}\otimes I_{T_N})\frac{\partial\mathbf{w}_p(s,k+1,r)}{\partial s}\Bigg|_{s=0}\Bigg).\end{aligned} \tag{65}$$

Here the vectors

$$\frac{\partial\mathbf{w}_p(s,l,r)}{\partial s}\Big|_{s=0},\quad l=\overline{1,K},\quad r=\overline{1,R},$$

are calculated as the sub-vectors of the vectors

$$\begin{aligned}&\frac{\partial\mathbf{w}_p(s,1)}{\partial s}\Bigg|_{s=0}\\ &=(\mathcal{A}-\mathcal{J}+H\otimes I_{T_N})^{-1}\\ &\quad\times(\mathbf{e}-(\mathcal{L}_p(I_R\otimes\mathbf{e}_{T_N})\boldsymbol{\beta}_p\frac{\mathrm{d}\mathbf{y}_p(s)}{\mathrm{d}s}\Bigg|_{s=0})),\end{aligned} \tag{66}$$

$$\begin{aligned}&\frac{\partial\mathbf{w}_p(s,l+1)}{\partial s}\Bigg|_{s=0}\\ &=(\mathcal{A}-(l+1)\mathcal{J}+H\otimes I_{T_N})^{-1}\\ &\quad\times(\mathbf{e}-(\mathcal{L}_p+l\mathcal{J})\frac{\partial\mathbf{w}_p(s,l)}{\partial s}\Big|_{s=0}),\quad l=\overline{1,K-1},\end{aligned} \tag{67}$$

and the values $\frac{\partial\mathbf{y}_p(s,r)}{\partial s}\Big|_{s=0}$ *are calculated as the sub-vectors of the vector*

$$\frac{\mathrm{d}\mathbf{y}_p(s)}{\mathrm{d}s}\Big|_{s=0}=(\hat{S}_p+H\otimes I_{M_1})^{-1}\mathbf{e}. \tag{68}$$

Corollary 3. *The average waiting time* V_p^{wait} *of an arbitrary priority customer is determined from the formula*

$$\begin{aligned}&V_p^{wait}\\ &=-\lambda_p^{-1}\sum_{i=N}^{\infty}\sum_{r=1}^{R}\sum_{k=0}^{\min\{i-N,K-1\}}(1-q_k^{(r)})\\ &\quad\times\boldsymbol{\pi}(i,r,k)(D_1^{(r)}\mathbf{e}\otimes I_{T_N})\frac{\partial\mathbf{z}_p(s,k+1,r)}{\partial s}\Bigg|_{s=0}\end{aligned} \tag{69}$$

where the values

$$\left.\frac{\partial \mathbf{z}_p(s,l,r)}{\partial s}\right|_{s=0}, \quad l=\overline{1,K}, \quad r=\overline{1,R},$$

are calculated as the entries of the vector

$$\left.\frac{\partial \mathbf{z}_p(s,1)}{\partial s}\right|_{s=0} = (\mathcal{A}-\mathcal{J}+H\otimes I_{T_N})^{-1}\mathbf{e}, \tag{70}$$

$$\begin{aligned}&\left.\frac{\partial \mathbf{z}_p(s,l+1)}{\partial s}\right|_{s=0}\\ &=(\mathcal{A}-(l+1)\mathcal{J}+H\otimes I_{T_N})^{-1}\\ &\quad\times(\mathbf{e}-((\mathcal{L}_p+l\mathcal{J})\left.\frac{\partial \mathbf{z}_p(s,l)}{\partial s}\right|_{s=0})), \quad l=\overline{1,K}.\end{aligned} \tag{71}$$

Theorem 3. *The LST $v_p^{\mathrm{serv}}(s)$ of the distribution of the sojourn time of an arbitrary priority customer, which successfully received service in the system, is computed by*

$$\begin{aligned}&v_p^{\mathrm{serv}}(s)\\ &=(\lambda_p^{\mathrm{out}})^{-1}\Bigg(\sum_{i=0}^{N-1}\sum_{r=1}^{R}\boldsymbol{\pi}(i,r)(D_1^{(r)}\otimes I_{T_i})\mathbf{e}\boldsymbol{\beta}_p^{(r)}\mathbf{y}_p(s,r)\\ &\quad+\sum_{i=N}^{\infty}\sum_{r=1}^{R}\sum_{k=0}^{\min\{i-N,K-1\}}(1-q_k^{(r)})\boldsymbol{\pi}(i,r,k)\\ &\quad\times(D_1^{(r)}\mathbf{e}\otimes I_{T_N})\mathbf{f}_p(s,k+1,r)\Bigg),\end{aligned} \tag{72}$$

where the vectors $\mathbf{f}_p(s,l,r)$, $l=\overline{1,K}$, $r=\overline{1,R}$, can be calculated as sub-vectors of the following vectors:

$$\begin{aligned}\mathbf{f}_p(s,1) &= (-\mathcal{A}+\mathcal{J}-H\otimes I_{T_N}+sI)^{-1}\\ &\quad\times\mathcal{L}_p(I_R\otimes\mathbf{e}_{T_N})\boldsymbol{\beta}_p\mathbf{y}_p(s),\end{aligned} \tag{73}$$

$$\begin{aligned}&\mathbf{f}_p(s,l+1)\\ &=(-\mathcal{A}+(l+1)\mathcal{J}-H\otimes I_{T_N}+sI)^{-1}\\ &\quad\times(\mathcal{L}_p+l\mathcal{J})\mathbf{f}_p(s,l), \quad l=\overline{1,K-1}.\end{aligned} \tag{74}$$

Corollary 4. *The average sojourn time $V_p^{\mathrm{soj-serv}}$ of an arbitrary priority customer, which successfully received service in the system, is calculated as*

$$\begin{aligned}&V_p^{\mathrm{soj-serv}}\\ &=-(v_p^{\mathrm{serv}}(s))'|_{s=0}\\ &=-(\lambda_p^{\mathrm{out}})^{-1}\Bigg(\sum_{i=0}^{N-1}\sum_{r=1}^{R}\boldsymbol{\pi}(i,r)(D_1^{(r)}\otimes I_{T_i})\\ &\quad\mathbf{e}\boldsymbol{\beta}_p^{(r)}\left.\frac{\partial \mathbf{y}_p(s,r)}{\partial s}\right|_{s=0}\\ &\quad+\sum_{i=N}^{\infty}\sum_{r=1}^{R}\sum_{k=0}^{\min\{i-N,K-1\}}(1-q_k^{(r)})\\ &\quad\times\boldsymbol{\pi}(i,r,k)(D_1^{(r)}\mathbf{e}\otimes I_{T_N})\left.\frac{\partial \mathbf{f}_p(s,k+1,r)}{\partial s}\right|_{s=0}\Bigg).\end{aligned} \tag{75}$$

Here the vectors

$$\left.\frac{\partial \mathbf{f}_p(s,l,r)}{\partial s}\right|_{s=0}, \quad l=\overline{1,K}, \quad r=\overline{1,R},$$

are calculated as the sub-vectors of the vectors

$$\begin{aligned}&\left.\frac{\partial \mathbf{f}_p(s,1)}{\partial s}\right|_{s=0}\\ &\quad=-(\mathcal{A}-\mathcal{J}+H\otimes I_{T_N})^{-2}\\ &\qquad\times\mathcal{L}_p(I_R\otimes\mathbf{e}_{T_N})\boldsymbol{\beta}_p\mathbf{y}_p(0)\\ &\qquad+(-\mathcal{A}+\mathcal{J}-H\otimes I_{T_N})^{-1}\\ &\qquad\times\mathcal{L}_p(I_R\otimes\mathbf{e}_{T_N})\boldsymbol{\beta}_p\left.\frac{\mathrm{d}\mathbf{y}_p(s)}{\mathrm{d}s}\right|_{s=0},\end{aligned} \tag{76}$$

$$\begin{aligned}&\left.\frac{\partial \mathbf{f}_p(s,l+1)}{\partial s}\right|_{s=0}\\ &\quad=-(\mathcal{A}-\mathcal{J}+H\otimes I_{T_N})^{-2}\\ &\qquad\times(\mathcal{L}_p+l\mathcal{J})\mathbf{f}_p(0,l)\\ &\qquad+(-\mathcal{A}+\mathcal{J}-H\otimes I_{T_N})^{-1}\\ &\qquad(\mathcal{L}_p+l\mathcal{J})\left.\frac{\partial \mathbf{f}_p(s,l)}{\partial s}\right|_{s=0}, \quad l=\overline{1,K-1}.\end{aligned} \tag{77}$$

7. Distribution of the sojourn time of an arbitrary non-priority customer in the system

Let $V_n(x)$ be the distribution function of the sojourn time of an arbitrary non-priority customer in the system and

$$v_n(s)=\int_0^{\infty}e^{-sx}\mathrm{d}V_n(x), \quad \mathrm{Re}\ s>0,$$

be its *LST*. Let us tag an arbitrary non-priority customer and keep track of its staying in the system. Consequently, $v_n(s)$ has the meaning of the probability that no catastrophe arrives during the sojourn time of the tagged customer.

To derive the expression for the *LST* $v_n(s)$, we also need to formulate and prove two auxiliary results. Let $y_n(s,r,m)$ be the probability that a catastrophe will not arrive during the rest of the tagged customer's service time in the system conditioned on the fact that during the given moment the state of the random environment is r, $r = \overline{1,R}$, and the state of service is m, $m = \overline{1,M_2}$.

Let us form the vectors

$$\mathbf{y}_n(s,r) = (y_n(s,r,1),\dots,y_n(s,r,M_2))^T, \quad r = \overline{1,R}, \tag{78}$$

$$\mathbf{y}_n(s) = (\mathbf{y}_n(s,1),\dots,\mathbf{y}_n(s,R))^T, \tag{79}$$

and introduce the following notation:

- $\hat{S}_n = \mathrm{diag}\{S_n^{(r)},\ r = \overline{1,R}\}$;
- $\hat{\mathbf{S}}_{n,0} = ((\mathbf{S}_{n,0}^{(1)})^T,\dots,(\mathbf{S}_{n,0}^{(R)})^T)^T$.

Lemma 4. *The vector* $\mathbf{y}_n(s)$ *is computed by the formula*

$$\mathbf{y}_n(s) = (-\hat{S}_n - H\otimes I_{M_2} + sI)^{-1}\hat{\mathbf{S}}_{n,0}. \tag{80}$$

Proof. The proof is analogous to that of Lemma 2. ■

Let $w_n(s,l,r,k,\nu,\eta^{(1)},\dots,\eta^{(M)})$ be the probability that a catastrophe will not arrive during the rest of the tagged customer sojourn time in the system conditioned on the fact that, at the given moment, the number of non-priority customers in the buffer that arrived to the system earlier than the tagged customer is equal to $l-1$, $l \ge 1$, the state of the random environment is r, $r = \overline{1,R}$, the number of priority customers in the buffer is equal to k, $k = \overline{0,K}$, and the states of the processes $\nu_t, \eta_t^{(1)},\dots,\eta_t^{(M)}$, $t \ge 0$, are $\nu,\eta^{(1)},\dots,\eta^{(M)}$, $\nu = \overline{0,W}$, $\eta^{(m)} = \overline{0,N}$, $\sum_{m=1}^{M}\eta^{(m)} = N$, $m = \overline{1,M}$.

Let us enumerate the probabilities $w_n(s,l,r,k,\nu,\eta^{(1)},\dots,\eta^{(M)})$ in the reverse lexicographic order of the components $\eta^{(1)},\dots,\eta^{(M)}$ and the direct lexicographic order of the component ν, and form from these probabilities the column vectors $\mathbf{w}_n(s,l,r,k)$.

Let us introduce also the column vectors

$$\mathbf{w}_n(s,l,r) = (\mathbf{w}_n^T(s,l,r,0),\dots,\mathbf{w}_n^T(s,l,r,K))^T, \quad l \ge 1, \quad r = \overline{1,R}, \tag{81}$$

$$\mathbf{w}_n^T(s,l) = (\mathbf{w}_n^T(s,l,1),\dots,\mathbf{w}_n^T(s,l,R))^T, \quad l \ge 1. \tag{82}$$

Lemma 5. *The vectors* $\mathbf{w}_n(s,l)$, $l \ge 1$, *can be calculated by the following recursive formulas:*

$$\mathbf{w}_n(s,1) = (sI-\mathcal{V})^{-1}\tilde{C}_n(I_R\otimes \mathbf{e}_{(K+1)\bar{W}T_N})\boldsymbol{\beta}_n\mathbf{y}_n(s), \tag{83}$$

$$\mathbf{w}_n(s,l+1) = (sI-\mathcal{V})^{-1}\tilde{C}_n\mathbf{w}_n(s,l), \quad l \ge 1, \tag{84}$$

where

$$\mathcal{V} = H\otimes I_{(K+1)\bar{W}T_N} - J_K + \bar{D}_0^{(K)} + \tilde{C}_p + \tilde{J} + \hat{D}_1 + \tilde{D}_1^{(K)} + \hat{D}_2 + \bar{A}_K + \bar{\Delta}_K, \tag{85}$$

$$\tilde{C}_p = \mathrm{diag}\{I^-\otimes I_{\bar{W}}\otimes L_0(N,\tilde{S}^{(r)}) \times P_{N-1}(\boldsymbol{\beta}_1^{(r)}),\quad r=\overline{1,R}\}, \tag{86}$$

$$\tilde{C}_n = \mathrm{diag}\{\bar{I}\otimes I_{\bar{W}}\otimes L_0(N,\tilde{S}^{(r)}) \times P_{N-1}(\boldsymbol{\beta}_2^{(r)}),\quad r=\overline{1,R}\}, \tag{87}$$

$$\boldsymbol{\beta}_n = \mathrm{diag}\{\boldsymbol{\beta}_n^{(1)},\dots,\boldsymbol{\beta}_n^{(R)}\}. \tag{88}$$

Proof. Based on a probabilistic sense of the *LST* and the law of total probability, it can be shown that the probability vectors $\mathbf{w}_n(s,l,r,k)$ satisfy the following system of linear algebraic equations:

$$\begin{aligned}
&\mathbf{w}_n(s,l,r,k) = ((s+l\alpha^{(r)}-H_{r,r})I_{\bar{W}T_N} - D_0^{(r)}\oplus A^{(r)})^{-1}\\
&\times\Big(\delta_{l,1}\delta_{k,0}I_{\bar{W}}\otimes L_0(N,\tilde{S}^{(r)})\mathbf{e}\boldsymbol{\beta}_n^{(r)}\mathbf{y}_n(s,r)\\
&+(1-\delta_{l,1})\delta_{k,0}I_{\bar{W}}\\
&\otimes L_0(N,\tilde{S}^{(r)})P_{N-1}(\boldsymbol{\beta}_2^{(r)})\mathbf{w}_n(s,l-1,r,0)\\
&+(1-\delta_{k,0})I_{\bar{W}}\otimes(L_p^{(r)}+k\alpha^{(r)}I_{T_N})\\
&\times\mathbf{w}_n(s,l,r,k-1) + \sum_{r'=1,\,r'\ne r}^{R} H_{r,r'}\mathbf{w}_n(s,l,r',k)\\
&+(1-q_k^{(r)})D_1^{(r)}\otimes I_{T_N}\mathbf{w}_n(s,l,r,k+1)\\
&+(q_k^{(r)}D_1^{(r)}+D_2^{(r)})\otimes I_{T_N}\mathbf{w}_n(s,l,r,k)\Big).
\end{aligned} \tag{89}$$

This system can be rewritten in matrix form as

$$\begin{aligned}
&\Big((-s+H_{r,r})I_{(K+1)\bar{W}T_N} - \alpha^{(r)}C_K\otimes I_{\bar{W}T_N}\\
&+I_{K+1}\otimes(D_0^{(r)}\oplus A^{(r)}) + I^-\otimes I_{\bar{W}}\otimes L_p^{(r)}\\
&+\alpha^{(r)}C_KI^-\otimes I_{\bar{W}T_N}((I-\tilde{Q}_K^{(r)})I^+\otimes D_1^{(r)}\\
&+\tilde{Q}_K^{(r)}\otimes D_1^{(r)} + I_{K+1}\otimes D_2^{(r)})\otimes I_{T_N}\Big)\mathbf{w}_n(s,l,r)\\
&+\delta_{l,1}\bar{I}\otimes I_{\bar{W}}\otimes L_0(N,\tilde{S}^{(r)})\mathbf{e}\boldsymbol{\beta}_n^{(r)}\mathbf{y}_n(s,r)\\
&+(1-\delta_{l,1})\bar{I}\otimes I_{\bar{W}}\otimes L_n^{(r)}\mathbf{w}_n(s,l-1,r)\\
&+\sum_{r'=1,\,r'\ne r}^{R} H_{r,r'}\mathbf{w}_n(s,l,r') = \mathbf{0}^T
\end{aligned} \tag{90}$$

and, then, in the form

$$(-sI+\mathcal{V})\mathbf{w}_n(s,l)+\delta_{l,1}\tilde{C}_n\mathbf{e}\boldsymbol{\beta}_n\mathbf{y}_n(s) + (1-\delta_{l,1})\tilde{C}_n\mathbf{w}_n(s,l-1)=\mathbf{0}^T. \quad (91)$$

The statement of Lemma 5 immediately follows from the formula (91). ■

Theorem 4. *The LST $v_n(s)$ of the distribution of an arbitrary non-priority customer's sojourn time in the system is computed by the formula*

$$\begin{aligned} v_n(s) &= \lambda_n^{-1}\Bigg(\sum_{i=0}^{N-1}\sum_{r=1}^{R}\boldsymbol{\pi}(i,r)(D_2^{(r)}\otimes I_{T_i})\mathbf{e}\boldsymbol{\beta}_n^{(r)}\mathbf{y}_n(s,r) \\ &\quad+\sum_{i=N}^{\infty}\sum_{r=1}^{R}\sum_{k=0}^{\min\{i-N,K\}}\boldsymbol{\pi}(i,r,k)(D_2^{(r)}\otimes I_{T_N}) \\ &\quad\times\mathbf{w}_n(s,i-N-k+1,r,k)\Bigg). \end{aligned} \quad (92)$$

Proof. The proof follows from the law of total probability, a probabilistic sense of the *LST*s as well as Lemmas 4 and 5. ■

Corollary 5. *The average sojourn time V_n^{soj} of an arbitrary non-priority customer is calculated as*

$$\begin{aligned} V_n^{\mathrm{soj}} &= -v_n'(s)\big|_{s=0} \\ &= -\lambda_n^{-1}\Bigg(\sum_{i=0}^{N-1}\sum_{r=1}^{R}\boldsymbol{\pi}(i,r)(D_2^{(r)}\otimes I_{T_i}) \\ &\quad\times\mathbf{e}\boldsymbol{\beta}_n^{(r)}\frac{\partial\mathbf{y}_n(s,r)}{\partial s}\bigg|_{s=0} \\ &\quad+\sum_{i=N}^{\infty}\sum_{r=1}^{R}\sum_{k=0}^{\min\{i-N,K\}}\boldsymbol{\pi}(i,r,k)(D_2^{(r)}\otimes I_{T_N}) \\ &\quad\times\frac{\partial\mathbf{w}_n(s,i-N-k+1,r,k)}{\partial s}\bigg|_{s=0}\Bigg). \end{aligned} \quad (93)$$

Here the vectors

$$\frac{\partial\mathbf{w}_n(s,l,r)}{\partial s}\bigg|_{s=0},\quad l\geq 1,\quad r=\overline{1,R},$$

are calculated as the sub-vectors of the vectors

$$\begin{aligned} &\frac{\partial\mathbf{w}_n(s,1)}{\partial s}\bigg|_{s=0} \\ &=\mathcal{V}^{-1}(\mathbf{e}-\tilde{C}_n(I_R\otimes\mathbf{e}_{(K+1)\bar{W}T_N})\boldsymbol{\beta}_n\frac{\mathrm{d}\mathbf{y}_n(s)}{\mathrm{d}s}\bigg|_{s=0}), \\ &\frac{\partial\mathbf{w}_n(s,l+1)}{\partial s}\bigg|_{s=0} \\ &=\mathcal{V}^{-1}(\mathbf{e}-\tilde{C}_n\frac{\partial\mathbf{w}_n(s,l)}{\partial s}\bigg|_{s=0}),\quad l\geq 1, \end{aligned} \quad (94)$$

and the values

$$\frac{\partial\mathbf{y}_n(s,r)}{\partial s}\bigg|_{s=0}$$

are calculated as the sub-vectors of the vector

$$\frac{\mathrm{d}\mathbf{y}_n(s)}{\mathrm{d}s}\bigg|_{s=0}=(\hat{S}_n+H\otimes I_{M_2})^{-1}\mathbf{e}. \quad (95)$$

Corollary 6. *The average waiting time V_n^{wait} of an arbitrary non-priority customer is calculated by the formula*

$$\begin{aligned} V_n^{\mathrm{wait}} &= -\lambda_n^{-1}\sum_{i=N}^{\infty}\sum_{r=1}^{R}\sum_{k=0}^{\min\{i-N,K\}}\boldsymbol{\pi}(i,r,k)(D_2^{(r)}\otimes I_{T_N}) \\ &\quad\times\frac{\partial\mathbf{z}_n(s,i-N-k+1,r,k)}{\partial s}\bigg|_{s=0}, \end{aligned} \quad (96)$$

where the values

$$\frac{\partial\mathbf{z}_n(s,l,r,k)}{\partial s}\bigg|_{s=0},\quad l\geq 1,\quad r=\overline{1,R},\quad k=\overline{0,K},$$

are calculated as the entries of the vector

$$\frac{\partial\mathbf{z}_n(s,1)}{\partial s}\bigg|_{s=0}=\mathcal{V}^{-1}\mathbf{e}, \quad (97)$$

$$\begin{aligned} &\frac{\partial\mathbf{z}_n(s,l+1)}{\partial s}\bigg|_{s=0} \\ &\quad=\mathcal{V}^{-1}\Big(\mathbf{e}-\tilde{C}_n\frac{\partial\mathbf{z}_n(s,l)}{\partial s}\bigg|_{s=0}\Big),\quad l\geq 1. \end{aligned} \quad (98)$$

8. Numerical example

To demonstrate the feasibility of the developed algorithms and numerically show some features of the system under consideration, we present the results of a numerical experiment.

Let us assume that the random environment has two states and the changes of the states are defined by the infinitesimal generator

$$H=\begin{pmatrix}-0.01 & 0.01\\ 0.1 & -0.1\end{pmatrix}.$$

One may interpret that the first state of the random environment corresponds to a normal mode of the system operation while the second state—to a mode when the system is overloaded by the customers (peak time).

Under the first state of the random environment the service time distribution of priority customers is

characterized by the vector $\boldsymbol{\beta}_p^{(1)} = (0.2, 0.8)$ and the matrix

$$S_p^{(1)} = \begin{pmatrix} -0.4368 & 0.40768 \\ 0.42608 & -1.71809 \end{pmatrix}.$$

The average service time is $b_{p,1}^{(1)} = 1.934$ and the coefficient of variation of this time is $c_{p,\text{var}}^{(1)} = 1.93$. The service time distribution of non-priority customers is characterized by the vector $\boldsymbol{\beta}_n^{(1)} = (0.2, 0.8)$ and the matrix

$$S_n^{(1)} = \begin{pmatrix} -1.2132 & 1.1235 \\ 2.5423 & -3.173 \end{pmatrix}.$$

The average service time is $b_{n,1}^{(1)} = 3.89$ and the coefficient of variation is $c_{n,\text{var}}^{(1)} = 1.14$. The intensity of impatience is $\alpha^{(1)} = 0.2$, the probabilities of balking are $q_k^{(1)} = 0.02(k+1)$, $k = \overline{0, K-1}$.

Under the second state of the random environment the service time distribution of priority customers is characterized by the vector $\boldsymbol{\beta}_p^{(2)} = (0.7, 0.3)$ and the matrix

$$S_p^{(2)} = \begin{pmatrix} -0.6402 & 0.56812 \\ 0.22308 & -1.34325 \end{pmatrix}.$$

The average service time is $b_{p,1}^{(2)} = 2.18$ and the coefficient of variation is $c_{p,\text{var}}^{(2)} = 0.91$. The service time distribution of non-priority customers is characterized by the vector $\boldsymbol{\beta}_n^{(2)} = (0.9, 0.1)$ and the matrix

$$S_n^{(2)} = \begin{pmatrix} -2.9453 & 2.8435 \\ 0.2967 & -0.6235 \end{pmatrix}.$$

The average service time is $b_{n,1}^{(2)} = 3.47$ and the coefficient of variation is $c_{n,\text{var}}^{(2)} = 0.9$. The intensity of impatience is $\alpha^{(2)} = 0.3$, the probabilities of balking are $q_k^{(2)} = 0.03(k+1)$, $k = \overline{0, K-1}$.

We assume that under the first state of the random environment the average arrival rate of customers $\lambda^{(1)} = 1$ (the average arrival rate of priority customers $\lambda_1^{(1)} = 5/6$, and the average arrival rate of non-priority customers $\lambda_2^{(1)} = 1/6$). Under the second state of the random environment, the average arrival rate of customers is three times higher: $\lambda^{(2)} = 3$ (the average arrival rate of priority customers $\lambda_1^{(2)} = 5/2$, and the average arrival rate of non-priority customers $\lambda_2^{(2)} = 1/2$).

To demonstrate that the performance measures of the system essentially depend not only on the average arrival rates but also on correlation of inter-arrival times, let us consider three different sets of *MMAP* flows having the same average arrival rate but different coefficients of correlation of inter-arrival times.

In the first set, coded as *MMAP*^{0}, the arrival flow under the first state of the random environment is defined by the matrices $D_0^{(1)} = -1$, $D_1^{(1)} = 5/6$ and $D_2^{(1)} = 1/6$, and under the first state of the random environment—defined by the matrices $D_0^{(2)} = -3$, $D_1^{(2)} = 5/2$ and $D_2^{(2)} = 1/2$. It this set, the arrival flows have the coefficients of correlation $c_{\text{cor}}^{(r)} = c_{\text{cor}}^{(r,l)} = 0$ and the coefficients of variation $c_{\text{var}}^{(r)} = c_{\text{var}}^{(r,l)} = 1$, $r, l = 1, 2$. Here, the arrival processes of priority and non-priority customers are defined as the stationary Poisson processes.

In the second set, coded as *MMAP*$^{0.2}$, we assume that the arrival flow under the first state of the random environment is defined by the matrices

$$D_0^{(1)} = \begin{pmatrix} -1.3518 & 0 \\ 0 & -0.04388 \end{pmatrix},$$

$$D_1^{(1)} = \begin{pmatrix} 1.119 & 0.00748 \\ 0.02037 & 0.01618 \end{pmatrix},$$

$$D_2^{(1)} = \begin{pmatrix} 0.2238 & 0.00152 \\ 0.00407 & 0.00326 \end{pmatrix},$$

and under the second state of the random environment—by the matrices

$$D_0^{(2)} = 3D_0^{(1)}, \quad D_1^{(2)} = 3D_1^{(1)}, \quad D_2^{(2)} = 3D_2^{(1)}.$$

It this set, the coefficients of correlation are $c_{\text{cor}}^{(r)} = 0.2$, $c_{\text{cor}}^{(r,1)} = 0.178, c_{\text{cor}}^{(r,2)} = 0.042$ and the coefficients of variation are $c_{\text{var}}^{(r)} = 12.34$, $c_{\text{var}}^{(r,1)} = 11.19$, $c_{\text{var}}^{(r,2)} = 3.97$, $r = \overline{1, 2}$.

In the third set, coded as *MMAP*$^{0.4}$, we assume that the arrival flow under the first state of the random environment is defined by the matrices

$$D_0^{(1)} = \begin{pmatrix} -3.41713 & 0.0201 \\ 0.001 & -0.11082 \end{pmatrix},$$

$$D_1^{(1)} = \begin{pmatrix} 2.81795 & 0.01288 \\ 0.01015 & 0.08133 \end{pmatrix},$$

$$D_2^{(1)} = \begin{pmatrix} 0.56358 & 0.00262 \\ 0.00202 & 0.01632 \end{pmatrix},$$

and under the second state of the random environment—by the matrices

$$D_0^{(2)} = 3D_0^{(1)}, \quad D_1^{(2)} = 3D_1^{(1)}, \quad D_2^{(2)} = 3D_2^{(1)}.$$

It has the coefficients of correlation $c_{\text{cor}}^{(r)} = 0.4$, $c_{\text{cor}}^{(r,1)} = 0.39, c_{\text{cor}}^{(r,2)} = 0.23$ and the coefficients of variation $c_{\text{var}}^{(r)} = 12.33$, $c_{\text{var}}^{(r,1)} = 12.06$, $c_{\text{var}}^{(r,2)} = 7.9$, $r = \overline{1, 2}$.

Let us fix the number of servers $N = 5$ and vary the priority customer's buffer capacity K in the interval $[0, 10]$.

Figures 2 and 3 illustrate the dependence of the average number of non-priority customers N_n^{buffer} and

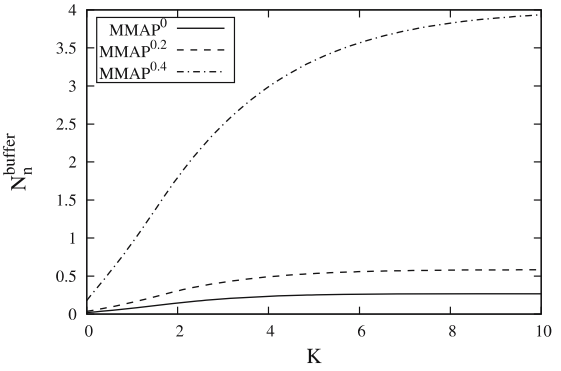


Fig. 2. Dependence of the average number of non-priority customers in the buffer N_n^{buffer} on the buffer capacity K.

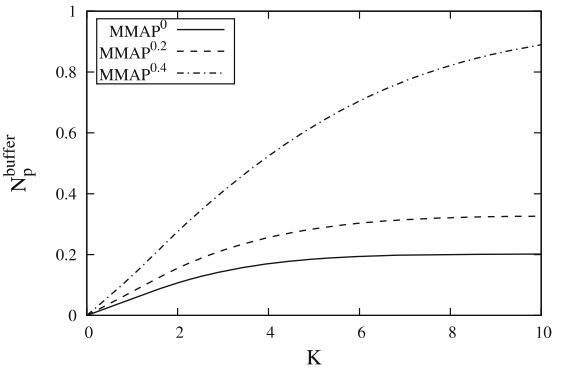


Fig. 3. Dependence of the average number of priority customers in the buffer N_p^{buffer} on the buffer capacity K.

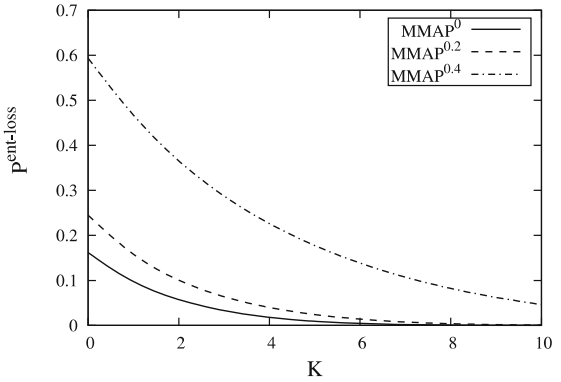


Fig. 4. Dependence of the probability $P^{\text{ent}-\text{loss}}$ on the buffer capacity K.

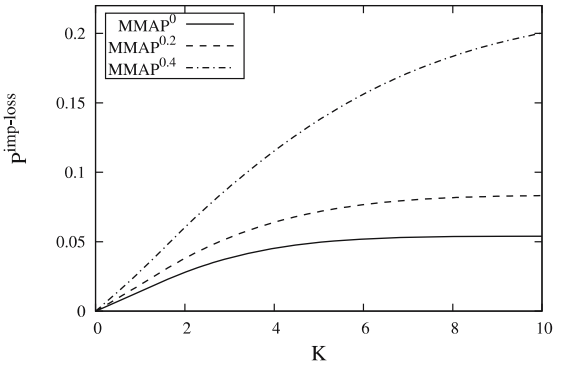


Fig. 5. Dependence of the probability $P^{\text{imp}-\text{loss}}$ on the buffer capacity K.

priority customers N_p^{buffer} in the buffer on the buffer capacity K for three sets of arrival processes.

The dependence of the loss probability $P^{\text{ent}-\text{loss}}$ of an arbitrary priority customer at the entrance to the system, the probability $P^{\text{imp}-\text{loss}}$ that an arbitrary priority customer will go to the buffer and leave it due to impatience and the loss probability P^{loss} of an arbitrary priority customer on the buffer capacity K is presented in Figs. 4–6, respectively.

Figures 7 and 8 show the dependence of the average sojourn and waiting times of an arbitrary priority customer on the buffer capacity K.

Figures 9 and 10 illustrate the dependence of the the average sojourn and waiting times of an arbitrary non-priority customer on the buffer capacity K.

It is worth mentioning that the analysis of the computational results shows that the following versions of Little's formula hold for the system under study:

$$V_p^{\text{wait}} = (\lambda_p)^{-1} N_p^{\text{buffer}}, \tag{99}$$

$$V_n^{\text{wait}} = (\lambda_n)^{-1} N_n^{\text{buffer}}. \tag{100}$$

The dependence of the average waiting time $V_p^{\text{soj}-\text{serv}}$ of an arbitrary priority customer, which successfully received service in the system, on the buffer capacity K is presented in Fig. 11.

Three important conclusions can be drawn from the analysis of the results of the numerical experiment.

- The capacity K of the buffer for priority customers has essential impact on the performance measures of the system. An increase in K leads to a decrease in the loss probability of a priority customer and probability of its loss at the entrance of the system. However, it also implies an increase in the probability of the loss of a priority customer due to impatience and the average waiting and sojourn times of a priority customer. Thus, our results can be useful for solving various optimization problems, e.g., for finding the value of K providing the minimal value of a loss probability under the restriction on the maximal admissible average waiting time or providing the minimal value of the average waiting time under the restriction on the admissible value of the loss probability.

- An increase in K leads to worse performance characteristics of service of non-priority customers. Accordingly, possible formulations of optimization problems may include also restrictions on the average sojourn time of non-priority customers.

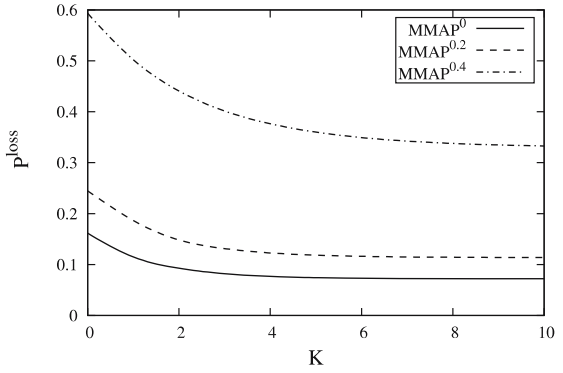


Fig. 6. Dependence of the loss probability P^{loss} on the buffer capacity K.

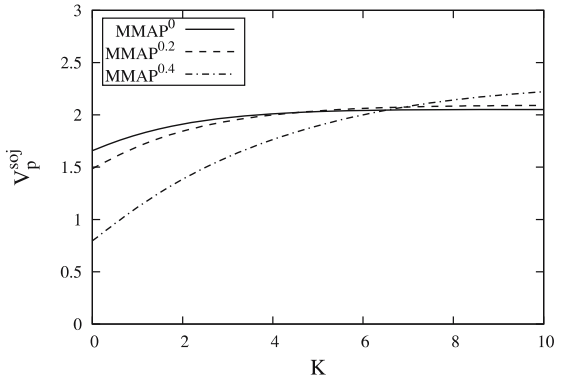


Fig. 7. Dependence of the average sojourn time of an arbitrary priority customer V_p^{soj} on the buffer capacity K.

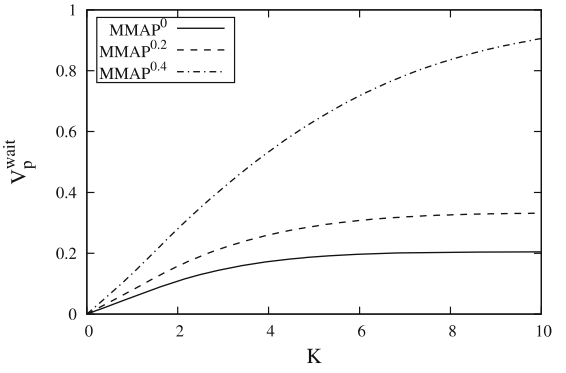


Fig. 8. Dependence of the average waiting time of an arbitrary priority customer V_p^{wait} on the buffer capacity K.

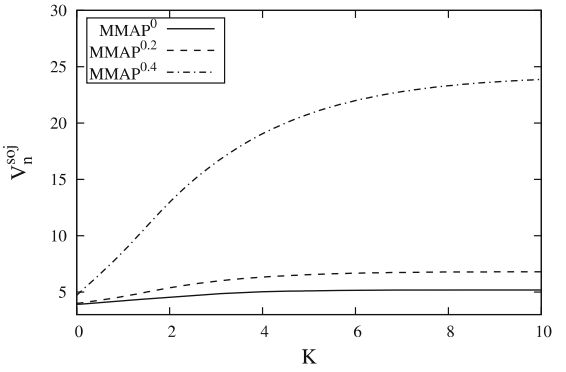


Fig. 9. Dependence of the average sojourn time of an arbitrary non-priority customer V_n^{soj} on the buffer capacity K.

- Correlation in the arrival process has an essential negative effect and ignorance of the correlation (which occurs, e.g., if arrival flows under the fixed values of the random environment are assumed to be stationary Poisson) leads to significant errors in the prediction of the values of the performance characteristics of the system.

It should be noted from comparison of Figs. 8 and 11 that an increase in the average sojourn time of an arbitrary serviced priority customer $V_p^{\text{soj}-\text{serv}}$ with an increase in the parameter K is much more essential (especially in the case of high correlation in arrival process) than an increase in the average sojourn time of an arbitrary priority customer V_p^{soj}. This is explained by the fact that in computation of the average sojourn time of an arbitrary priority customer it is suggested that the sojourn time be equal to zero in the case of the loss of a priority customer while this probability decreases when K grows.

9. Conclusion

A multi-server queueing system with a heterogeneous Markovian arrival process and phase-type distributions of the service time operating under the influence of some external random environment was analyzed. One type of customer has a time priority but restricted access to the buffer while another type has a space priority (unlimited access to the buffer). Customers having the time priority are impatient and can balk the system when all servers are busy at an arrival instant. The behavior of the system was described by a multi-dimensional continuous time Markov chain. The generator of this chain is written in block matrix form. The chain belongs to the class of quasi-birth-and-death processes with many boundary states, so the problem of computation of its stationary distribution is solved by standard techniques. Based on ergodic theorems for Markov chains, expressions for a bunch of performance measures of the system were derived. The technique of a matrix analog of the so-called method of catastrophes was successfully applied for the derivation of algorithms for computation of Laplace–Stieltjes transforms of the distributions of waiting and sojourn time for customers having different priorities. A numerical example illustrating the feasibility of the developed algorithms and dependence of some performance measures of the system on the capacity of a buffer space available for time priority customers was presented. The importance of taking into account correlation in the arrival process (which is possible

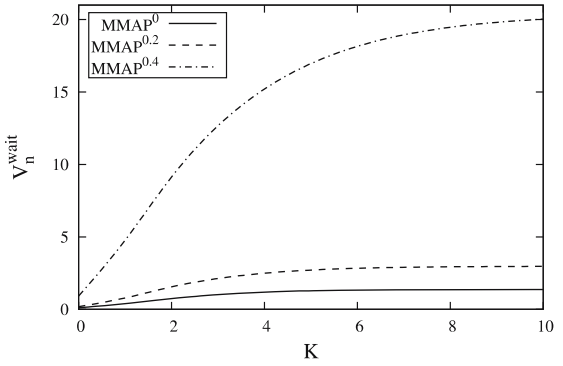


Fig. 10. Dependence of the average waiting time of an arbitrary non-priority customer V_n^{wait} on the buffer capacity K.

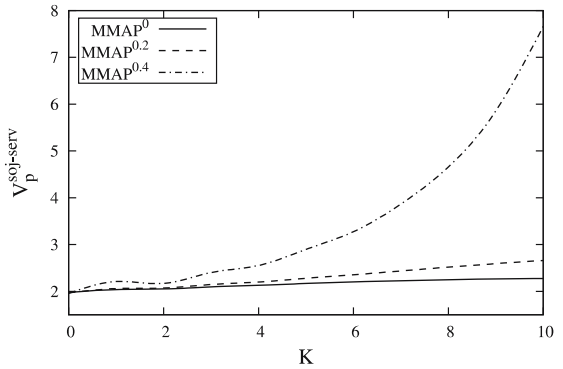


Fig. 11. Dependence of the average sojourn time of an arbitrary serviced priority customer $V_p^{\text{soj}-\text{serv}}$ on the buffer capacity K.

due to the use of the marked Markov arrival process instead of the stationary Poisson process) was numerically illustrated. The validity of variants of Little's formula for the system under study were empirically justified.

The results can be useful for modeling a lot of different real world systems, including call centers and mobile networks with HSDPA protocols. They can be applied for solving different optimization problems, e.g., selection of a required number of servers, sharing a buffer space, dynamical adjusting of the required service rate to the randomly changing arrival rate, etc.

Acknowledgment

This research has been supported by a 2013 Sangji University research grant.

References

Aksin, O., Armony, M. and Mehrotra, V. (2007). The modern call centers: A multi-disciplinary perspective on operations management research, *Production and Operation Management* **16**(6): 655–688.

Al-Begain, K., Dudin, A., Kazimirsky, A. and Yerima, S. (2009). Investigation of the $M_2/G_2/1/\infty, N$ queue with restricted admission of priority customers and its application to HSDPA mobile systems, *Computer Networks* **53**(6): 1186–1201.

Al-Begain, K., Dudin, A. and Mushko, V. (2006). Novel queueing model for multimedia over downlink in 3.5g wireless network, *Journal of Communication Software and Systems* **2**(2): 68–80.

Chakravarthy, S. (2001). The batch Markovian arrival process: A review and future work, *in* A. Krishnamoorthy, N. Raju and V. Ramaswami (Eds.), *Advances in Probability Theory and Stochastic Processes,* Notable Publications, Neshanic Station, NJ, pp. 21–49.

Chydziński, A. and Chróst, Ł. (2011). Analysis of AQM queues with queue size based packet dropping, *International Journal of Applied Mathematics and Computer Science* **21**(3): 567–577, DOI: 10.2478/v10006-011-0045-7.

Dudin, A., Osipov, E., Dudin, S. and Schelen, O. (2013a). Socio-behavioral scheduling of time-frequency resources for modern mobile operators, *Communications in Computer and Information Science* **356**: 69–82.

Dudin, S., Kim, C. and Dudina, O. (2013b). $MMAP/M/N$ queueing system with impatient heterogeneous customers as a model of a contact center, *Computers and Operations Research* **40**(7): 1790–1803.

Graham, A. (1981). *Kronecker Products and Matrix Calculus with Applications*, Ellis Horwood, Chichester.

Jouini, O., Aksin, Z. and Dallery, Y. (2011). Call centers with delay information: Models and insights, *Manufacturing & Service Operations Management* **13**(4): 534–548.

Jouini, O., Dallery, Y. and Aksin, Z. (2009). Queuing models for flexible multi-class call centers with real-time anticipated delays, *International Journal of Production Economics* **120**(2): 389–399.

Kesten, H. and Runnenburg, J. (1956). *Priority in Waiting Line Problems*, Mathematisch Centrum, Amsterdam.

Kim, C., Dudin, A., Klimenok, V. and Khramova, V. (2010a). Erlang loss queueing system with batch arrivals operating in a random environment, *Computers and Operations Research* **36**(3): 674–697.

Kim, C., Klimenok, V., Mushko, V. and Dudin, A. (2010b). The $BMAP/PH/N$ retrial queueing system operating in Markovian random environment, *Computers and Operations Research* **37**(7): 1228–1237.

Kim, C., Dudin, S., Dudin, A. and Dudina, O. (2013a). Queueing system $MAP/PH/N/R$ with session arrivals operating in random environment, *Communications in Computer and Information Science* **370**: 406–415.

Kim, C., Dudin, S., Taramin, O. and Baek, J. (2013b). Queueing system $MMAP/PH/N/N+R$ with impatient heterogeneous customers as a model of call center, *Applied Mathematical Modelling* **37**(3): 958–976.

Kim, C., Klimenok, V., Lee, S. and Dudin, A. (2007). The $BMAP/PH/1$ retrial queueing system operating in random environment, *Journal of Statistical Planning and Inference* **137**(12): 3904–3916.

Kim, J. and Park, S. (2010). Outsourcing strategy in two-stage call centers, *Computers and Operations Research* **37**(4): 790–805.

Klimenok, V. and Dudin, A. (2006). Multi-dimensional asymptotically quasi-Toeplitz Markov chains and their application in queueing theory, *Queueing Systems* **54**(4): 245–259.

Krieger, U., Klimenok, V., Kazimirsky, A., Breuer, L. and Dudin, A. (2005). A $BMAP/PH/1$ queue with feedback operating in a random environment, *Mathematical and Computer Modelling* **41**(8–9): 867–882.

Lucantoni, D. (1991). New results on the single server queue with a batch Markovian arrival process, *Communication in Statistics-Stochastic Models* **7**(1): 1–46.

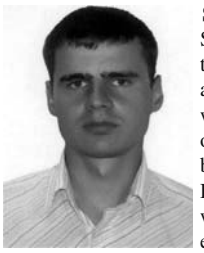
Neuts, M. (1981). *Matrix-geometric Solutions in Stochastic Models—An Algorithmic Approach*, Johns Hopkins University Press, Baltimore, MD.

Olwal, T., Djouani, K., Kogeda, O. and van Wyk, B. (2012). Joint queue-perturbed and weakly coupled power control for wireless backbone networks, *International Journal of Applied Mathematics and Computer Science* **22**(3): 749–764, DOI: 10.2478/v10006-012-0056-z.

van Danzig, D. (1955). Chaines de markof dans les ensembles abstraits et applications aux processus avec regions absorbantes et au probleme des boucles, *Annals de l'Institute H. Pioncare* **14**(3): 145–199.

Chesoong Kim obtained his M.Sc. and Ph.D. degrees in engineering from the Department of Industrial Engineering at Seoul National University in 1989 and 1993, respectively. He was a visiting scholar in the Department of Mechanical Engineering at the University of Queensland, Australia, from September 1998 to August 1999. He was a foreign scientist at the School of Mathematics & Statistics at Carleton University, Canada, from July 2003 to August 2004. He was also a visiting professor in the Department of Industrial Engineering at the University of Washington, USA, from August 2004 to August 2005. He has had scientific visits to Belarusian State University in Belarus and the University of Debrecen in Hungary. He was selected for *Who's Who in Asia* and *Who's Who in the World* in 2007. He is currently a full professor and the head of the Department of Industrial Engineering at Sangji University. His research interests are in stochastic processes, queueing theory, with particular emphasis on computer and wireless communication networks, queueing network modeling and their applications. He has published around 60 papers in internationally refereed journals. He has been the recipient of a number of grants from the Korean Science and Engineering Foundation (KOSEF) and the Korea Research Foundation (KRF).

Alexander N. Dudin obtained his Ph.D. degree in probability theory and mathematical statistics in 1982 from Vilnius University and the Doctor of Science degree in 1992 from Tomsk University. He is the head of the Laboratory of Applied Probabilistic Analysis at Belarusian State University, and a professor at the Probability Theory and Mathematical Statistics Department. He is an author of 300 publications, including more than 80 papers in top level journals. He is the chairman of the IPC of the Belarusian Winter Workshops on *Queueing Theory*, held since 1985. He serves as a member of IPCs of several international conferences. His field of scientific interests includes random processes in queueing systems, controllable queueing systems and their optimization, queueing systems in random environment, retrial queueing systems, applications of queueing theory to telecommunication. He has been invited for lecturing and research to the USA, the UK, Germany, France, Holland, Japan, South Korea, India, China, Italy and Sweden.



Sergey A. Dudin graduated from Belarusian State University in 2007. In 2010, he received the Ph.D. degree at the same university in system analysis, control and information processing, and works currently as a senior scientific researcher of the Research Laboratory of Applied Probabilistic Analysis at Belarusian State University. His main fields of interests are queueing systems with session arrivals and controlled tandem models.

Olga S. Dudina graduated from Belarusian State University in 2007. In 2010, she received her Ph.D. degree at the same university in probability theory and mathematical statistics, and works currently as a senior scientific researcher of the Research Laboratory of Applied Probabilistic Analysis at Belarusian State University. Her main fields of interests are queueing tandem queueing models with correlated arrival flows, non-Markovian queueing systems.

Received: 19 August 2013
Revised: 24 February 2014