Operations Research Perspectives 5 (2018) 245–255

Contents lists available at ScienceDirect

Operations Research Perspectives

journal homepage: www.elsevier.com/locate/orp

Analysis of queueing model with processor sharing discipline and customers impatience

A.N. Dudin[*,a,b], S.A. Dudin[a,b], O.S. Dudina[a,b], K.E. Samouylov[b]

[a] Department of Applied Mathematics and Computer Science, Belarusian State University, 4, Nezavisimosti Ave., Minsk 220030, Belarus
[b] Department of Applied Probability and Informatics, RUDN University, 6, Miklukho-Maklaya st., Moscow 117198, Russia

ARTICLE INFO

Keywords:
Processor sharing
Admission control
Markovian arrival process
Impatience

ABSTRACT

Queueing systems with processor sharing represent the adequate models for sharing the resources, e.g., components of a computer or a bandwidth of communication systems. In this paper, we consider a queueing system with processor sharing discipline under quite general assumptions about the arrival and service processes. Arrivals are defined by the Markovian arrival process. The service time has a phase type distribution. Possible impatience of customers is taken into account. The number of customers, which can simultaneously obtain service, is limited. We compare two approaches for monitoring service of customers, namely, the approach counting the number of customers at each phase of service and the approach counting the phase of service of each customer and show the significant advantage of the former approach. We obtain the joint distribution of the number of customers in the system and the states of the underlying arrival and service processes as well as the loss probabilities. It is shown that the sojourn time in the system of an arbitrary customer has phase type distribution and an irreducible representation of this distribution is obtained. Numerical examples are presented. A possibility of optimal choice of the server capacity (e.g., multi-programming level) is numerically illustrated. An opportunity of increasing the speed of computations via the use of the graphics processing unit is discussed.

1. Introduction

Processor sharing discipline is very popular in computers, communication systems and networks. For references and examples of real-world applications see, e.g. [16], the surveys [32,33] as well as the recent papers [19,31]. In particular, this discipline is very popular for tasks scheduling in multi-programming computer systems. The model considered in our paper significantly extends possibility of adequate modelling of such systems. We do not impose restrictive assumptions like an exponential distribution of all times characterizing the behavior of the system and a flow of tasks as well as on the number of tasks that share the computer resources. The presented results allow to consider a task processing in a computer or communication system as a whole sequence of various operations, e.g., using CPU, GPU, RAM, I/O devices, etc, not just a single operation duration of which has an exponential distribution.

In the classical settings, a processor can be shared by the unlimited number of users and the majority of the existing literature is devoted to the analysis of queueing systems under this assumption. However, in many applications of this discipline in computer systems and communication networks this assumption is not fulfilled because a certain minimal share of the bandwidth of the computer or channel has to be guaranteed to provide acceptable quality of service to a customer. Therefore, the *limited* processor sharing or processor sharing with a finite capacity is often considered. This kind of processor sharing suggests that the maximal number, say N, $N < \infty$, of users who may obtain service simultaneously is fixed. Customers arriving when the capacity of the server is not exhausted immediately start service with the rate which is, in general, inversely proportional to the number of customers in service. The majority of the existing research is addressed to analysis of the simple $M/M/1$ type queues where it is assumed that the arrivals are described by the stationary Poisson process and the service time distribution is exponential. However, both these assumptions look quite artificial in many real-world systems. In particular, it is already well recognized that the stationary Poisson arrival process is not a good descriptor of the real-world information flows and the Markovian arrival process (*MAP*) suits much better for the description of such flows, see, e.g. [6,17,30]. An exponential distribution is a very particular case of the phase type (*PH*) distribution successfully used for approximation of an arbitrary distribution, see, e.g. [1]. In our paper, to provide the advanced model, we assume that the arrival process is defined by the *MAP* and the service time distribution is of phase type. A short list of related papers, in which at least one of the unrealistic assumptions that

* Corresponding author at: Belarusian state university, 4, Independence Ave., Minsk 220030, Belarus.
E-mail addresses: dudin@bsu.by (A.N. Dudin), dudin85@mail.ru (S.A. Dudin), dudina@bsu.by (O.S. Dudina), ksam@sci.pfu.edu.ru (K.E. Samouylov).

https://doi.org/10.1016/j.orp.2018.08.003
Received 8 June 2018; Received in revised form 19 August 2018; Accepted 20 August 2018
Available online 22 August 2018

the arrivals are defined by the stationary Poisson process and the service time distribution is exponential is omitted, is as follows. The model with the infinite capacity of the server and the *MAP* is considered, e.g., in [10,18,20]. The model with the finite capacity and the *MAP* is considered, e.g., in [7,26]. It is worth to note that as a rule the problem of computation of the stationary distribution of the number of customers in the system under processor sharing discipline has a known solution which coincides with the solution for the corresponding system with first-in-first-out service discipline. The problem of computation of the sojourn time distribution is more complicated. This problem for the *M/M/*1 and *MAP/M/*1 systems with an infinite capacity was addressed in [20,34], correspondingly. The moments of the sojourn time distribution for the unreliable *MAP/M/*1 system with a finite capacity are computed in [26]. In all cited above papers, it was assumed that the service time has an exponential distribution. This assumption is more or less suitable for modelling the systems with the coefficient of variation of the service time equal to 1. However, in some real-world systems, including cellular wireless communication networks, the distribution of the service time may have higher variation, see, e.g. [23] and the hyper-exponential distribution describes the duration of holding times in such networks better. The hyper-exponential distribution as well as the Erlangian distribution is very particular case of the *PH* distribution. The model of *M/PH/*1 type with unlimited processor sharing was considered in the paper [27]. The mathematical technique exploited for analysis there can be hardly used in the case of the *MAP* arrival process.

In this paper, we consider the *MAP/PH/*1 queue with limited processor sharing. The very recent paper [28] is devoted to detailed consideration of an analogous system along with a survey of the related research. However, there are three essential differences between our paper and [28]. (i) We assume that a customer arriving when the capacity of the server is exhausted is lost while in [28] it is assumed that such a customer joins the buffer of an infinite capacity to obtain service later. It seems that the model with customer loss better suits, e.g., for modelling bandwidth sharing in wireless communication networks. (ii) In real-world systems, customers may be impatient and leave the system before service completion due to long processing. When the processor is shared by many customers, service of each customer becomes slower and importance of account of an impatience phenomenon increases. In our model, we account possible impatience of customers. (iii) We use another description of the system states by the multi-dimensional Markov chain. This description allows to compute characteristics of the system faster and for much larger capacity N of the server. E.g., even in the case when the state spaces of the underlying Markov processes of the *MAP* arrival process and the *PH* distribution consists of only two states, it is more or less realistic to compute characteristics of the system based on the classical description of the system states only for N up to 12. The effective description applied in our paper allows to make computations even for N equal to 1000.

The rest of the paper is organized as follows. In Section 2, the mathematical model of the system under study is described. The stationary distribution of the number of customers in the system is analysed in Section 3. The dynamics of the system is described by the multi-dimensional Markov chain, the generator of which is derived and equilibrium equations are written down. Formulas for the throughput of the system and the customer loss probabilities (due to the server capacity exhausting and due to impatience) are presented. In Section 4, it is shown that the sojourn time of an arbitrary customer has a phase type distribution. Section 5 contains the numerical results illustrating the dependence of the key performance measures of the system on its capacity, correlation in the arrival process and variance of the service times. An optimization problem is considered in brief. An advisability of using for computations the graphics processing unit (GPU) is discussed. Section 6 concludes the paper.

2. Description of the model

We consider a single-server queueing system without a buffer. The arrival process is the *MAP*. Arrivals are controlled by the underlying irreducible continuous-time Markov chain $\nu_t, t \geq 0$, with a finite state space $\{0, 1, ..., W\}$. The *MAP* is defined by the square matrices D_k, $k = 0, 1$, of size $W + 1$ consisting of the intensities of transitions of the Markov chain ν_t accompanied by the arrival of k customers. The matrix $D_0 + D_1$ is an infinitesimal generator of the process ν_t. The stationary distribution vector $\boldsymbol{\theta}$ of this process is the unique solution of the system $\boldsymbol{\theta}(D_0 + D_1) = \boldsymbol{0}$, $\boldsymbol{\theta}\mathbf{e} = 1$ where $\mathbf{e}$ is a column vector consisting of1's', and $\boldsymbol{0}$ is a zero row vector. The average intensity λ (fundamental rate) of the *MAP* is given by $\lambda = \boldsymbol{\theta}D_1\mathbf{e}$. We assume that $\lambda < \infty$. For more detailed and exact definition of the *MAP* and motivation of its importance for description of the correlated bursty arrival flows in modern communication networks see [6,17,30].

The service time of an individual customer (service in absence of other customers) has a *PH* distribution with an irreducible representation $(\boldsymbol{\beta}, S)$. This service time can be interpreted as the time until the underlying Markov process $\eta_t, t \geq 0$, with a finite state space $\{1, \cdots, M, M + 1\}$ reaches the single absorbing state $M + 1$, conditioned on the fact that the initial state of this process is selected among the transient states $\{1, \cdots, M\}$ with probabilities defined by the entries of the probabilistic row vector $\boldsymbol{\beta} = (\beta_1, ..., \beta_M)$. The transition rates of the process η_t within the set $\{1, \cdots, M\}$ are defined by the sub-generator S and the transition rates into the absorbing state (which leads to service completion) are given by the entries of the column vector $\boldsymbol{S}_0 = -S\mathbf{e}$. The Laplace-Stieltjes transform of the distribution having an irreducible representation $(\boldsymbol{\beta}, S)$ is defined as $\boldsymbol{\beta}(sI - S)^{-1}\boldsymbol{S}_0$, $Re\, s > 0$. For more detailed information about the *PH* distribution see [21].

The problem of constructing the matrices D_0, D_1, S and the vector $\boldsymbol{\beta}$ based on traces of real arrival and service processes is extensively addressed in the literature and may be more or less easily solved based on the results from, e.g. [4,5,22].

We assume that up to N customers can be served simultaneously. The number N is called the capacity of the server. If during an arbitrary customer arrival epoch the number of customers in service is less than N, the customer is admitted and immediately starts obtaining service. If the number of customers in service is equal to N, the arriving customer leaves the system permanently (is lost). The most well-known results relating to the systems with processor sharing assume the exponential distribution of individual customer service time. Let us denote the parameter of this distribution (rate) by μ. It is assumed that when i customers simultaneously receive service each customer is served with the rate $\frac{\mu}{i}$. Because here we assume the *PH* distribution of service time, it is necessary at first to specify the interaction of simultaneous services. It is reasonable to do this in the following way. It follows from the description of the *PH* distribution given above that the service time of a customer can be interpreted as the walking time of a customer in the open network consisting of M nodes. The customer starts walking from the node m with the probability β_m, $m = \overline{1, M}$. Here, denotation like $m = \overline{1, M}$ means that the variable m takes the values from the set $\{1, \cdots, M\}$. Then, the customer makes the transitions within this network. The intensities of the transitions are given by the entries of the matrix S. Then, the customer leaves the network with the intensities given by the entries of the column vector $\boldsymbol{S}_0$. From this interpretation, it is clear that the starting phase of the service of any customer should be chosen independently of other customers receiving service. The individual intensities of the transitions within the network during the periods when i customers present in the system are defined by the components of the sub-generator $\frac{1}{i}S$, $i = \overline{1, N}$. The intensities of transitions leading to service completion are defined by the components of the vector $\boldsymbol{S}_{0,i} = -\frac{1}{i}S\mathbf{e}$, $i = \overline{1, N}$.

It is worth to note that the presented below results can be easily extended to the case of more general, than the supposed above, inversely proportional dependence of the intensities of the transitions between the phases on the number i of customers presenting in the system.

As it was mentioned in Introduction, account of customers impatience is very important in analysis of the processor sharing discipline

and we assume that the customers are impatient. Customers do not have information about the number of other customers receiving service. Therefore, the patience time of the customer does not depend on other customers. If the duration of the mth phase of service of a customer exceeds a random time having an exponential distribution with the parameter α_m, then the customer terminates service and leaves the system permanently (is lost), $\alpha_m \geq 0,\ m = \overline{1, M}$.

The goals of our analysis are to find the stationary distributions of the number of customers in the system and the sojourn time of an arbitrary customer, the loss probability of an arbitrary customer and the throughput of the system and to find the optimal value of the server capacity N.

3. Stationary distribution of the number of customers in the system

Let $i_t,\ i_t = \overline{0, N}$, be the number of customers receiving service in the system at the moment $t, t \geq 0$. The process i_t is non-Markovian. To study this process, at first we have to construct the multi-dimensional Markov process that includes i_t as the component. It is clear that this process has to also include the state ν_t of the underlying process of the *MAP* and the components that keep track of the service underlying processes of customers presenting in the system. There are two opportunities to keep track of the service processes. One of them, called in [12] as TPFS (track-phase-for-server), counts a current phase of service of each customer. The second one called in [12] as CSFP (count-server-for-phase) counts the number of customers receiving the service at the certain phase. The approach, which uses CPFS, is traced back to the papers [24,25]. The key information about this approach is as follows.

Let us consider an arbitrary queueing system where up to N *PH* service processes defined by an irreducible representation $(\boldsymbol{\beta}, S)$ run independently of each other. Here, the size of the row vector $\boldsymbol{\beta}$ and the sub-generator S is assumed to be equal to M. Let the current number $i_t,\ i_t = \overline{0, N}$, of running processes be equal to i and $h_t^{(m)}$ be the number of processes that currently stay at the phase $m,\ h_t^{(m)} \in \{0, ..., i\}$, $m = \overline{1, M},\quad \sum_{m=1}^{M} h_t^{(m)} = i.$ Introduce the vector process $\mathbf{h}_t = (h_t^{(1)}, \cdots, h_t^{(M)})$ where the components $h_t^{(m)},\ m = \overline{1, M}$, are assumed to be enumerated in the reverse lexicographic order. Denote by $A_i(N, S)$ the matrix, the non-diagonal entries of which define the intensities of transitions of the process $\mathbf{h}_t$ that do not lead to the change of the number i of the running processes and the diagonal entries are equal to 0. By $L_{N-i}(N, \widetilde{S})$ we denote the matrix, entries of which define the intensities of transitions of the process $\mathbf{h}_t$ when the number i of the running processes decreases by one. By $P_i(\boldsymbol{\beta})$ we denote the matrix, entries of which define the transition probabilities of the process $\mathbf{h}_t$ when the number i of the running processes increases by one. Here, the square matrix $\widetilde{S}$ of size $M + 1$ is defined by $\widetilde{S} = \begin{pmatrix} 0 & \mathbf{0} \\ \mathbf{S}_0 & O_M \end{pmatrix}$ where O_M is the zero square matrix of size M.

Formulas and algorithms for recursive computation of matrices $A_i(N, S)$, $L_{N-i}(N, \widetilde{S})$ and $P_i(\boldsymbol{\beta})$ can be found, e.g., in [14,15,24,25]. The number $M_i = \binom{i+M-1}{M-1} = \frac{(i+M-1)!}{i!\,(M-1)!}$ defines the number of possible states of the process $\mathbf{h}_t$ when the number of running in parallel processes is equal to $i,\ i = \overline{0, N}$.

In the paper [28] devoted to consideration of a similar system, but with the buffer of an infinite capacity and without account of customers impatience, the TPFS approach was used. The choice of the authors of [28] is explained by the reasonings that the form of the blocks of the generator of Markov chain under study in that approach is more transparent. Our choice of the CSFP approach, which requires derivation and computer realization of more involved formulas, is explained by the fact that the TPFS approach allows to compute the stationary distribution of the system under study only for relatively small capacity N of the server. Essential computational advantages of CSFP over TPFS were illustrated, e.g., in [12,15]. For instance, the CSFP approach requires operation with the blocks of size up to $M_N = \binom{N+M-1}{M-1}$ while the TPFS approach requires operation with the blocks of size up to M^N. For instance, if $M = 2$, the CSFP approach requires operation with the blocks of size up to $(N + 1)$ while the TPFS approach requires operation with the blocks of size up to 2^N. If $N = 20$, these sizes are 21 and 1 048 576, correspondingly.

Let us adopt the outlined above CSFP approach originally developed for multi-server queues with *PH* distribution of service time for analysis of the considered *MAP/PH/*1 system with the limited processor sharing discipline and impatient customers. Let the number of customers in service at the moment t be equal to i, the number of customers in service at phase m be $h_t^{(m)},\ m = \overline{1, M}$, and $\mathbf{h}_t = (h_t^{(1)}, \cdots, h_t^{(M)})$. Comparing to the situation, for which this approach was explained, we have two distinguishing features: (a) in our queueing model, the running in parallel service processes are not independent (due to processor sharing); (b) each of the running in parallel processes may be terminated ahead of the schedule (due to customers impatience).

The feature (b) is accounted via the introducing and using an additional denotation. Let $\widehat{L}_{N-i}(N, \widehat{S})$ be the matrix, the entries of which define the intensities of transitions of the process $\mathbf{h}_t$ when one of i customers receiving service leaves the system due to impatience. It is obvious that the matrices $\widehat{L}_{N-i}(N, \widehat{S})$ can be computed in the same way as the matrices $L_{N-i}(N, \widetilde{S})$ by replacing the matrix $\widetilde{S}$ with the matrix

$$\widehat{S} = \begin{pmatrix} 0 & \mathbf{0} \\ \boldsymbol{\alpha} & O_M \end{pmatrix}$$

where $\boldsymbol{\alpha}$ is the column vector having the entries $\alpha_m,\ m = \overline{1, M}$.

Accounting also the feature (a), we prove the following statement.

Lemma 1. *Let the number of customers in service be equal to $i,\ i = \overline{1, N}$. Then the intensities of transitions of the vector process $\mathbf{h}_t$ that do not lead to the change of the number of customers in service are given by the entries of the square matrix $\frac{1}{i}A_i(N, S)$ of size M_i. The intensities of transitions of the process $\mathbf{h}_t$ when the number of customers in service decreases from i to $i - 1$ are defined by the entries of the matrix*

$$\frac{1}{i}L_{N-i}(N, \widetilde{S}) + \widehat{L}_{N-i}(N, \widehat{S})$$

of size $M_i \times M_{i-1},\ i = \overline{1, N}$. The transition probabilities of the process $\mathbf{h}_t$ when the number of customers in service increases from i to $i + 1$ are defined by the entries of the matrix $P_i(\boldsymbol{\beta})$ of size $M_i \times M_{i+1},\ i = \overline{0, N-1}$. The matrix $P_0(\boldsymbol{\beta})$ is equal to the vector $\boldsymbol{\beta}$.

Proof of Lemma 1 is straightforward because, as it was noted in the previous section, the sharing of a single server by i customers implies that the transition rates between the phases and to the absorbing state are equal to the corresponding rates of the individual service of one customer divided by i. The rate of service termination of an arbitrary customer due to impatience does not depend on the number of other customers in service.

It is easy to see that the dynamics of the considered queueing system is described by the multi-dimensional process

$$\zeta_t = \{i_t, \nu_t, h_t^{(1)}, \cdots, h_t^{(M)}\},\ t \geq 0,$$

which is the continuous-time Markov chain. If $i_t = 0$, the components $\{h_t^{(1)}, \cdots, h_t^{(M)}\}$ are absent.

Let us introduce the following notation:

- *I is an identity matrix of the corresponding size. If the size is not clear from context, it is indicated by the suffix, i.e. $I_{\overline{W}}$ is the identity matrix of size $\overline{W} = W + 1$;*
- *$\otimes$ and $\oplus$ are symbols of Kronecker product and sum of matrices, see [11];*
- *$\mathrm{diag}\{\mathbf{b}\}$ is the diagonal matrix with the diagonal entries defined by the entries of the vector $\mathbf{b}$;*
- $\Delta^{(i)}(N) = \mathrm{diag}\left\{\frac{1}{i}(A_i(N, S)\mathbf{e} + L_{N-i}(N, \widetilde{S})\mathbf{e}) + \widehat{L}_{N-i}(N, \widehat{S})\mathbf{e}\right\},\ i = \overline{1, N}.$

Let us enumerate the states of the Markov chain $\zeta_t, t \geq 0$, in the direct lexicographic order of the component ν_t and the reverse lexicographic order of the components $h_t^{(1)}, \cdots, h_t^{(M)}$. We refer to the set of states of the Markov chain with value i of the first component as a level i. Let Q be the generator of the Markov chain $\zeta_t, t \geq 0$, consisting of the blocks $Q_{i,j}$, which define the transition rates of this chain from the level i to the level j. The diagonal entries of the matrices $Q_{i,i}$ are negative. The modulus of the diagonal entry of the blocks $Q_{i,i}$ defines the total intensity of departure from the corresponding state of the Markov chain $\zeta_t, t \geq 0$.

Lemma 2. *The generator of the Markov chain $\zeta_t, t \geq 0$, has the following block-tridiagonal form:*

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

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

$$Q_{0,0} = D_0,$$

$$Q_{i,i} = D_0 \oplus \frac{1}{i} A_i(N, S) - I_{\overline{W}} \otimes \Delta^{(i)}(N), \ i = \overline{1, N-1},$$

$$Q_{N,N} = D_0 \oplus \frac{1}{N} A_N(N, S) - I_{\overline{W}} \otimes \Delta^{(N)}(N) + D_1 \otimes I_{M_N},$$

$$Q_{i,i+1} = D_1 \otimes P_i(\boldsymbol{\beta}), \ i = \overline{0, N-1},$$

$$Q_{i,i-1} = I_{\overline{W}} \otimes \left(\frac{1}{i} L_{N-i}(N, \widetilde{S}) + \widehat{L}_{N-i}(N, \widehat{S}) \right), \quad i = \overline{1, N}.$$

Proof of Lemma 2 follows from analysis of all possible transitions of the components of the Markov chain ζ_t during the time interval having an infinitesimal length.

Since the Markov chain ζ_t is irreducible and has a finite state space, the stationary probabilities

$$\pi(0, \nu) = \lim_{t \to \infty} P\{i_t = 0, \nu_t = \nu\}, \ \nu = \overline{0, W},$$

$$\pi(i, \nu, h^{(1)}, \cdots, h^{(M)}) = \lim_{t \to \infty} P\{i_t = i, \nu_t = \nu, h_t^{(1)} = h^{(1)}, \cdots, h_t^{(M)} = h^{(M)}\},$$

$$\nu = \overline{0, W}, \quad \sum_{m=1}^{M} h^{(m)} = i, \ i = \overline{1, N},$$

exist for any set of the system parameters.

Let π_0 be the row vector formed by the probabilities $\pi(0, \nu)$, and π_i be the row vector formed by the probabilities $\pi(i, \nu, h^{(1)}, \cdots, h^{(M)})$, enumerated in the direct lexicographic order of the component ν and the reverse lexicographic order of the components $(h^{(1)}, \cdots, h^{(M)})$, $i = \overline{1, N}$. Let $\boldsymbol{\pi} = (\pi_0, \cdots, \pi_N)$.

Corollary 1. *The vectors $\pi_i, i = \overline{0, N}$, satisfy the following system of equilibrium equations*

$$\pi_0 D_0 + \pi_1 (I_{\overline{W}} \otimes (L_{N-1}(N, \widetilde{S}) + \widehat{L}_{N-1}(N, \widehat{S}))) = \mathbf{0}, \tag{1}$$

$$\pi_i \left(D_0 \oplus \frac{1}{i} A_i(N, S) - I_{\overline{W}} \otimes \Delta^{(i)}(N) \right) + \pi_{i+1} \left(I_{\overline{W}} \otimes \left(\frac{1}{i+1} L_{N-i-1}(N, \widetilde{S}) + \widehat{L}_{N-i-1}(N, \widehat{S}) \right) \right) + \pi_{i-1}(D_1 \otimes P_{i-1}(\boldsymbol{\beta})) = \mathbf{0}, \ i = \overline{1, N-1}, \tag{2}$$

$$\pi_N \left((D_0 + D_1) \oplus \frac{1}{N} A_N(N, S) - I_{\overline{W}} \otimes \Delta^{(N)}(N) \right) + \pi_{N-1}(D_1 \otimes P_{N-1}(\boldsymbol{\beta})) = \mathbf{0}. \tag{3}$$

System (1)-(3) obviously follows from the equilibrium (or Chapman-Kolmogorov) equation $\pi Q = \mathbf{0}$.

There exist many methods for solving the finite system of the linear algebraic Eqs. (1)–(3) supplemented with the normalization condition $\sum_{i=0}^{N} \pi_i \mathbf{e} = 1$. For instance, the algorithm described in [2, 9] and some references from [3], which effectively uses the block-tridiagonal structure of the generator Q, can be recommended.

Algorithm 1

Step 1. Compute the set of the matrices $G_i, \ i = \overline{1, N}$, from the recursion

$$G_1 = -Q_{1,0} Q_{0,0}^{-1},$$

$$G_i = -Q_{i,i-1} (Q_{i-1,i-1} + G_{i-1} Q_{i-2,i-1})^{-1}, \ i = \overline{2, N}.$$

Step 2. Compute the row vector $\boldsymbol{\psi}_N$ as the unique solution of the system

$$\boldsymbol{\psi}_N (Q_{N,N} + G_N Q_{N-1,N}) = \mathbf{0}, \quad \boldsymbol{\psi}_N \mathbf{e} = 1.$$

Step 3. Compute the set of the vectors $\boldsymbol{\psi}_i, \ i = \overline{0, N-1}$, from the backward recursion

$$\boldsymbol{\psi}_i = \boldsymbol{\psi}_{i+1} G_{i+1}, \ i = \overline{0, N-1}.$$

Step 4. Compute the normalizing constant $c = \sum_{i=0}^{N} \boldsymbol{\psi}_i \mathbf{e}$.

Step 5. Compute the probability vectors $\boldsymbol{\pi}_i, \ i = \overline{0, N}$, as

$$\boldsymbol{\pi}_i = \frac{1}{c} \boldsymbol{\psi}_i, \ i = \overline{0, N}.$$

Remark 1. It is well known in the literature that if the arrival flow is described by the stationary Poisson process and the service time distribution is exponential, system (1)–(3) has the same form and solution as the equilibrium equations for the stationary probabilities of the system *M/M/1/N* with a finite buffer and FIFO (first-in-first-out) service discipline. It can be checked that the similar fact is valid for the system with the *MAP* as well. However, consideration of the *PH* distribution of the service time drastically changes the situation. The problem of computation of the stationary distribution of the system states cannot be reduced to the analogous problem for the single-server queue with the finite buffer.

Corollary 2. *The average number $\widehat{N}$ of customers receiving service at an arbitrary moment is computed by*

$$\widehat{N} = \sum_{i=1}^{N} i \boldsymbol{\pi}_i \mathbf{e}.$$

Corollary 3. *The intensity T of output of customers that obtained service in the system is computed by*

$$T = \sum_{i=1}^{N} \boldsymbol{\pi}_i \left(I_{\overline{W}} \otimes \frac{1}{i} L_{N-i}(N, \widetilde{S}) \right) \mathbf{e}.$$

Corollary 4. *The probability $P_{ent-loss}$ that an arbitrary customer is lost because it arrives when the server capacity is exhausted is computed by the formula*

$$P_{ent-loss} = \frac{1}{\lambda} \boldsymbol{\pi}_N (D_1 \otimes I_{M_N}) \mathbf{e}.$$

Corollary 5. *The probability $P_{imp-loss}$ that an arbitrary customer is lost due to impatience is computed by the formula*

$$P_{imp-loss} = \frac{1}{\lambda} \sum_{i=1}^{N} \boldsymbol{\pi}_i (I_{\overline{W}} \otimes \widehat{L}_{N-i}(N, \widehat{S})) \mathbf{e}.$$

Corollary 6. *The probability P_{loss} that an arbitrary customer is lost (at the entrance or due to impatience) is computed by the formula*

$$P_{loss} = P_{ent-loss} + P_{imp-loss}$$

or by the formula

$$P_{loss} = 1 - \frac{T}{\lambda}.$$

Remark 2. Availability of two different formulas for computing the probability P_{loss} is helpful for control of the accuracy of computation of the stationary distribution of the system states.

Remark 3. In the case when the size of the vectors $\boldsymbol{\pi}_i$, $i = \overline{0, N}$, is large, the memory-efficient method developed in [3] for computing the stationary characteristics without preliminary computation of the stationary distribution of the Markov chain can be applied to compute the listed above performance measures of the system.

4. The stationary distribution of the sojourn time of an arbitrary customer in the system

The sojourn time distribution of an arbitrary customer is one of the most important characteristics of any queueing system. In this section, we show that the sojourn time of an arbitrary customer in the system under study has a phase type distribution and obtain the irreducible representation of this distribution. To this end, first we derive the Laplace-Stieltjes transform of the stationary distribution of the sojourn time of an arbitrary customer.

4.1. Computation of the Laplace-Stieltjes transform of the stationary distribution of the sojourn time of an arbitrary customer in the system

Let $V(x)$ be the distribution function of the sojourn time of an arbitrary customer in the system and $v(s) = \int_0^\infty e^{-sx} dV(x)$, $Re\ s > 0$, be the LST of this distribution. To obtain the expression for this LST, we tag an arbitrary customer and monitor its processing in the system. We separately count the state of the underlying process $\eta_t^{(1)}$ of service of this customer and the number of other customers at each phase of service.

Let $\mathbf{v}(s) = (\mathbf{v}_0(s), \cdots, \mathbf{v}_{N-1}(s))$ be the column vector sub-vectors $\mathbf{v}_i(s)$ of which define the LST of the sojourn time of the tagged customer conditioned on the fact that i other customers receive service, $i = \overline{0, N-1}$, and the processes ν_t, $\eta_t^{(1)}$, $h_t^{(1)}$, $\cdots, h_t^{(M)}$ have the corresponding states.

Let us introduce some auxiliary denotations and matrices:

- $\mathrm{diag}\{F_k,\ k = \overline{1, K}\}$ is the block-diagonal matrix with the diagonal blocks $F_1, \cdots, F_K$;
- $\mathrm{diag}_+\{F_k,\ k = \overline{1, K}\}$ is the matrix having all zero blocks except the updiagonal blocks defined by the matrices $F_1, \cdots, F_K$;
- $\mathrm{diag}_-\{F_k,\ k = \overline{1, K}\}$ is the matrix having all zero blocks except the subdiagonal blocks defined by the matrices $F_1, \cdots, F_K$;
- $I_+ = \mathrm{diag}_+\{D_1 \otimes I_M \otimes P_n(\boldsymbol{\beta}),\ n = \overline{0, N-2}\}$;
- $I_- = \mathrm{diag}_-\left\{I_{\overline{W}M} \otimes \left(\frac{1}{n} L_{N-n}(N-1, \widetilde{S}) + \widehat{L}_{N-1-n}(N-1, \widehat{S})\right),\ n = \overline{2, N}\right\}$;
- $\mathcal{H} = \mathrm{diag}\left\{D_0 \oplus S,\ D_0 \oplus \frac{1}{n+1} S \oplus A_n(N-1, S) - I_{\overline{W}M} \otimes \Delta^{(n)}(N-1),\ ;\right.$
  $\left. n = \overline{1, N-1}\right\} + I_+ + I_-$
- $\mathbf{H}_0 = -\mathcal{H}\mathbf{e}$.

Lemma 3. *The vector* $\mathbf{v}(s)$ *is computed by the formula*

$$\mathbf{v}(s) = (sI - \mathcal{H})^{-1}\mathbf{H}_0. \tag{4}$$

To prove this lemma, we derive the system of linear algebraic equations for the vectors $\mathbf{v}_i(s)$, $i = \overline{0, N-1}$. *In derivation, we use the so-called method of collective marks (method of additional event, method of catastrophes), for references see, e.g. [13, 29]. To this end, we interpret the variable* s *as the intensity of some virtual stationary Poisson flow of the so-called catastrophes. The use of this additional flow allows to obtain a system of equations for the conditional LSTs under study based on the transparent probabilistic derivations.*

Let us assume that the tagged customer arrives to the system when i *customers receive service. It is easy to understand that the components of the vector LST* $\mathbf{v}_i(s)$, $i = \overline{0, N-1}$, *define the probability that a catastrophe will not arrive during the stay of the tagged customer in the system when* i *other customers are servicing in the system. It is evident that if* $i = N$ *the tagged customer is lost and the probability that no catastrophe arrives during its sojourn time is equal to 1. Therefore, we consider in detail only the case* $i = \overline{0, N-1}$. *The structure of the tagged customer sojourn time is as follows. First, during a time interval of a length* $\tau, 0 < \tau < \infty$, *the components* $(\nu_t, \eta_t^{(1)}, h_t^{(1)}, \cdots, h_t^{(M)})$ *can make only transitions that do not lead to the new customer arrival, service completion or termination. It is well known that the probabilities of such transitions of these components and no catastrophe arrival during time* τ *are defined by the entries of the matrix exponent*

$$e^{\left(D_0 \oplus \frac{1}{i+1}(S \oplus A_i(N-1,S)) - I_{\overline{W}M} \otimes \Delta^{(i)}(N-1) - sI_{\overline{W}MM_i}\right)\tau}.$$

After the moment τ, *during the time interval of the infinitesimal length* $d\tau$ *the following events can occur:*

- *A new customer arrives and starts service (if* $i + 1 < N$*). The intensities of such transitions are given by the entries of the matrix* $D_1 \otimes I_M \otimes P_i(\boldsymbol{\beta})$. *After this event occurrence, the probabilities of no catastrophe arrival during the rest of the sojourn time of the tagged customer are defined by the vector* $\mathbf{v}_{i+1}(s)$.
- *Service of one of* i, $i = \overline{1, N-1}$, *non-tagged customers is finished or terminated. The intensities of such transitions are given by the entries of the matrix*

  $$I_{\overline{W}M} \otimes \left(\frac{1}{i+1} L_{N-1-i}(N-1, \widetilde{S}) + \widehat{L}_{N-1-i}(N-1, \widehat{S})\right).$$

  After this event, the probabilities of no catastrophe arrival during the rest of the sojourn time of the tagged customer are defined by the vector $\mathbf{v}_{i-1}(s)$.
- *Service of the tagged customer is finished or terminated. The intensities of such a transition are given by the entries of the matrix* $I_{\overline{W}} \otimes \left(\frac{1}{i+1}\mathbf{S}_0 + \boldsymbol{\alpha}\right) \otimes I_{M_i}$. *After this event, the probabilities of no catastrophe arrival during the rest of the sojourn time of the tagged customer are defined by the vector* $\mathbf{e}$.

Therefore, for $i = \overline{1, N-2}$ *we have the relation*

$$\mathbf{v}_i(s) = \int_0^\infty e^{\left(D_0 \oplus \frac{1}{i+1}(S \oplus A_i(N-1,S)) - I_{\overline{W}M} \otimes \Delta^{(i)}(N-1) - sI_{\overline{W}MM_i}\right)\tau} \Bigg[(D_1 \otimes I_M \otimes P_i(\boldsymbol{\beta}))\mathbf{v}_{i+1}(s) + \left(I_{\overline{W}M} \otimes \left(\frac{1}{i+1} L_{N-1-i}(N-1, \widetilde{S}) + \widehat{L}_{N-1-i}(N-1, \widehat{S})\right)\right)\mathbf{v}_{i-1}(s) + \mathbf{e}_{\overline{W}} \otimes \left(\frac{1}{i+1}\mathbf{S}_0 + \boldsymbol{\alpha}\right) \otimes \mathbf{e}_{M_i} \Bigg] d\tau,$$

from which we have the equation

$$\begin{aligned} &\left(D_0 \oplus \frac{1}{i+1}(S \oplus A_i(N-1,S)) - I_{\overline{W}M} \otimes \Delta^{(i)}(N-1) - sI_{\overline{W}MM_i}\right)\mathbf{v}_i(s) + (D_1 \otimes I_M \otimes P_i(\boldsymbol{\beta}))\mathbf{v}_{i+1}(s) \\ &+ \left(I_{\overline{W}M} \otimes \left(\frac{1}{i+1} L_{N-1-i}(N-1, \widetilde{S}) + \widehat{L}_{N-1-i}(N-1, \widehat{S})\right)\right)\mathbf{v}_{i-1}(s) \\ &+ \mathbf{e}_{\overline{W}} \otimes \left(\frac{1}{i+1}\mathbf{S}_0 + \boldsymbol{\alpha}\right) \otimes \mathbf{e}_{M_i} = \mathbf{0}^T,\ i = \overline{1, N-2}. \end{aligned} \tag{5}$$

Analogously, for $i = 0$ *we obtain the equation*

$$\left(D_0 \oplus S - sI_{\overline{W}M}\right)\mathbf{v}_0(s) + (D_1 \otimes I_M \otimes P_0(\boldsymbol{\beta}))\mathbf{v}_1(s) + \mathbf{e}_{\overline{W}} \otimes (\mathbf{S}_0 + \boldsymbol{\alpha}) = \mathbf{0}^T \quad (6)$$

and for $i = N-1$ we obtain the equation

$$\left(D_0 \oplus \frac{1}{N}(S \oplus A_{N-1}(N-1, S)) - I_{\overline{W}M} \otimes \Delta^{(N-1)} + D_1 \otimes I_{MM_{N-1}} - sI_{\overline{W}MM_{N-1}}\right)\mathbf{v}_{N-1}(s) + \left(I_{\overline{W}M} \otimes \left(\frac{1}{N}L_0(N-1, \widetilde{S}) + \widehat{L}_0(N-1, \widehat{S})\right)\right)\mathbf{v}_{N-2}(s) + \mathbf{e}_{\overline{W}} \otimes \left(\frac{1}{N}\mathbf{S}_0 + \boldsymbol{\alpha}\right) \otimes \mathbf{e}_{M_{N-1}} = \mathbf{0}^T. \quad (7)$$

Using the introduced above denotations, Eqs. (5)–(7) can be rewritten in form (4). Because the matrix $\mathcal{H}$ is the sub-generator, the diagonal entries of the matrix $sI - \mathcal{H}$ dominate in each row. Therefore, the matrix $sI - \mathcal{H}$ is non-singular. Lemma 3 is proved.

Theorem 1. *LST $v(s), Re\, s > 0$, of the sojourn time of an arbitrary customer in the system is computed as*

$$v(s) = P_{ent-loss} + \frac{1}{\lambda}\sum_{i=0}^{N-1}\boldsymbol{\pi}_i(D_1 \otimes \boldsymbol{\beta} \otimes I_{M_i})\mathbf{v}_i(s).$$

Proof easily follows from the formula of total probability.

Corollary 7. *LST $\widetilde{v}(s)$ of the sojourn time of an arbitrary customer, which is not lost at the entrance to the system, is computed as*

$$\widetilde{v}(s) = \frac{1}{\lambda(1 - P_{ent-loss})}\sum_{i=0}^{N-1}\boldsymbol{\pi}_i(D_1 \otimes \boldsymbol{\beta} \otimes I_{M_i})\mathbf{v}_i(s).$$

4.2. Computation of the distribution of the sojourn time of an arbitrary customer in the system

Above we defined a random variable having *PH* distribution with an irreducible representation $(\boldsymbol{\beta}, S)$ as the time until the underlying Markov process $\eta_t, t \geq 0$, with a finite state space $\{1, \cdots, M, M+1\}$ reaches the single absorbing state $M+1$, conditioned on the fact that the initial state of this process is selected among the *transient* states $\{1, \cdots, M\}$ with probabilities defined by the entries of the probabilistic row vector $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_M)$.

A bit more general definition of *PH* distribution assumes that the number φ_0 and the row vector $\boldsymbol{\varphi}$ such that the row vector $(\varphi_0, \boldsymbol{\varphi})$ is the stochastic one are fixed instead of the stochastic vector $\boldsymbol{\beta}$. With probability φ_0 the underlying Markov process selects the absorbing state as the initial state. The components of the vector φ_0 define the probabilities of the choice by the underlying Markov process of one of the transient states as the initial state. In this more general case, it is possible to say about an irreducible representation $(\varphi_0, \boldsymbol{\varphi}, S)$.

Theorem 2. *The sojourn time of an arbitrary customer in the system has a PH distribution with an irreducible representation $(\varphi_0, \boldsymbol{\varphi}, \mathcal{H})$ where*

$$\varphi_0 = P_{ent-loss},\ \boldsymbol{\varphi} = \frac{1}{\lambda}(\boldsymbol{\pi}_0(D_1 \otimes \boldsymbol{\beta} \otimes I_{M_0}), \boldsymbol{\pi}_1(D_1 \otimes \boldsymbol{\beta} \otimes I_{M_1}), \cdots, \boldsymbol{\pi}_{N-1}(D_1 \otimes \boldsymbol{\beta} \otimes I_{M_{N-1}})).$$

Proof of the theorem is obvious. Using the introduced denotations, the statement of Theorem 1 can be rewritten as

$$v(s) = \varphi_0 + \boldsymbol{\varphi}\mathbf{v}(s).$$

Taking into account (4), this formula transforms into

$$v(s) = \varphi_0 + \boldsymbol{\varphi}(sI - \mathcal{H})^{-1}\mathbf{H}_0.$$

This formula evidently defines the LST of the pH distribution with an irreducible representation $(\varphi_0, \boldsymbol{\varphi}, \mathcal{H})$. Theorem 2 is proved.

Corollary 8. *The distribution function of the sojourn time of an arbitrary customer has the form*

$$V(x) = 1 - \boldsymbol{\varphi}e^{\mathcal{H}x}\mathbf{e}.$$

4.3. Computation of the moments of the sojourn time distribution

Sometimes, in real-world systems, information about the distribution function of the sojourn time is redundant for managerial purposes and it is enough to compute only several moments of distribution. Here we present an algorithm for their computation.

Corollary 9. *The rth moment v_r, $r \geq 1$, of the distribution of the sojourn time of an arbitrary customer, including the lost customers, is computed as*

$$v_r = \frac{1}{\lambda}\sum_{i=0}^{N-1}\boldsymbol{\pi}_i(D_1 \otimes \boldsymbol{\beta} \otimes I_{M_i})\mathbf{v}_i^{(r)} = \boldsymbol{\varphi}\mathbf{v}^{(r)}$$

where the vectors $\mathbf{v}_i^{(r)}$, $i = \overline{0, N-1}$, are the components of the vector $\mathbf{v}^{(r)}$ defined by formula

$$\mathbf{v}^{(r)} = r!(-\mathcal{H})^{-r}\mathbf{e}, \quad r \geq 1. \quad (8)$$

Proof of the corollary evidently follows from the formula of total probability and properties of the LST. The average sojourn time of an arbitrary customer in the system is equal to v_1, the variance of the sojourn time is defined as $v_2 - v_1^2$.

Corresponding formula for an arbitrary non-lost customer is easily obtained by analogy with the statement of Corollary 7.

The disadvantage of explicit formula (8) from the computational point of view is necessity to compute the reverse matrix $\mathcal{H}^{-1}$ and the degrees of this matrix. The size of this matrix is equal to $\overline{W}M\sum_{m=0}^{N-1}M_m$ and can be large. Therefore, it is more preferable to elaborate another, recursive, way for computing the vectors $\mathbf{v}^{(r)}$, $r \geq 1$. This way is described as follows.

The vectors $\mathbf{v}^{(r)}$ are defined via the vector LST $\mathbf{v}(s)$ by well-known formula

$$\mathbf{v}^{(r)} = (-1)^r\mathbf{v}^{(r)}(0)$$

where $\mathbf{v}^{(r)}(0)$ means the rth derivative of the vector LST $\mathbf{v}(s)$ at the point $s = 0$.

It is easy to show that the derivatives $\mathbf{v}^{(r)}(0)$ can be computed recursively by the formula:

$$\mathcal{H}\mathbf{v}^{(r)}(0) = r\mathbf{v}^{(r-1)}(0), \quad r \geq 1, \quad (9)$$

with the initial condition $\mathbf{v}^{(0)}(0) = \mathbf{e}$.

When the vector in the right hand side of (9) is already known, say it is equal to a vector $\mathbf{b} = (\mathbf{b}_0, \cdots, \mathbf{b}_{N-1})^T$, the unknown vector $\mathbf{x} = \mathbf{v}^{(r)}(0)$ is defined as the solution to the system $\mathcal{H}\mathbf{x} = \mathbf{b}$. To solve such a system of linear algebraic equations with block-tridiagonal structure of the matrix $\mathcal{H}$, we recommend the following algorithm.

Algorithm 2

Step 1. Denote as $\mathcal{H}_{i,i}$, $i = \overline{0, N-1}$, the diagonal blocks of the matrix $\mathcal{H}$,$\mathcal{H}_{i,i-1}$, $i = \overline{1, N-1}$, the sub-diagonal blocks of the matrix $\mathcal{H}$, and $\mathcal{H}_{i,i+1}$, $i = \overline{0, N-2}$, the up-diagonal blocks of the matrix $\mathcal{H}$.

Step 2. Compute the sequences of the matrices $\widetilde{\mathcal{H}}_{i,i}$ and the vectors $\widetilde{\mathbf{b}}_i$ from the backward recursion

$$\widetilde{\mathcal{H}}_{i,i} = (\mathcal{H}_{i,i} - \mathcal{H}_{i,i+1}\widetilde{\mathcal{H}}_{i+1,i+1}\mathcal{H}_{i+1,i})^{-1}, \quad i = \overline{0, N-2},$$

$$\widetilde{\mathbf{b}}_i = \mathbf{b}_i - \mathcal{H}_{i,i+1}\widetilde{\mathcal{H}}_{i+1,i+1}\widetilde{\mathbf{b}}_{i+1}, \quad i = \overline{0, N-2},$$

with the initial conditions

$$\widetilde{\mathcal{H}}_{N-1,N-1} = \mathcal{H}^{-1}_{N-1,N-1},\ \widetilde{\mathbf{b}}_{N-1} = \mathbf{b}_{N-1}.$$

Step 3. Recursively compute the components $\mathbf{x}_i$, $i = \overline{0, N-1}$, of the unknown vector $\mathbf{x}$ by the formulas

$$\mathbf{x}_0 = -\widetilde{\mathcal{H}}^{-1}_{0,0}\widetilde{\mathbf{b}}_0,$$

$$\mathbf{x}_i = -\widetilde{\mathcal{H}}^{-1}_{i,i}(\widetilde{\mathbf{b}}_i - \mathcal{H}_{i,i-1}\mathbf{x}_{i-1}),\ i = \overline{1, N-1}.$$

All the inverted matrices are the irreducible sub-generators and, therefore, are non-singular. The reverse matrices have non-negative entries. This guarantees the numerical stability of Algorithm 2 and its easy computer implementation.

5. Numerical experiments

The goals of the experiments are to demonstrate the feasibility of the proposed algorithms and to show the importance of account of correlation in the arrival process and variance of the service time distribution.

Let us assume that the parameters of the system are fixed as follows. The *MAP* is defined by the matrices

$$D_0 = \begin{pmatrix} -13.52 & 0 \\ 0 & -0.43(8) \end{pmatrix},\ D_1 = \begin{pmatrix} 13.43 & 0.09 \\ 0.2(4) & 0.19(4) \end{pmatrix}.$$

This arrival process has the average arrival rate $\lambda = 10$, the coefficient of correlation of two successive intervals between arrivals $c_{cor} = 0.2$, and the squared coefficient of variation of the intervals between customer arrivals $c_{var} = 12.34$.

The service time distribution is Erlangian of order 2 and is defined by the vector $\boldsymbol{\beta} = (1, 0)$ and the sub-generator $S = \begin{pmatrix} -20 & 20 \\ 0 & -20 \end{pmatrix}$. The service time has the mean value equal to 0.1 and the squared coefficient of variation equal to 0.5.

Let the customers be patient during the first phase of the service, i.e., $\alpha_1 = 0$, while the impatience rate during the second phase of service be $\alpha_1 = 0.05$.

The experiments were implemented on PC having the following configuration: Intel Core i7-8700 CPU 3.20 GHz (6 cores), 16 Gb RAM, video card Nvidia Ge Force GTX 1050 Ti 4Gb, Cuda 8.0, Java 9.

The first conclusion from the implemented experiments consists of confirmation of essential advantage of CSFP approach for description of service process. Using the alternative, TPFS approach, due to RAM limitation, we succeeded to compute stationary distributions only for the server capacity N up to 12. Using the CSFP approach, we succeeded to compute the stationary distribution of the number of customers in the system for the server capacity N up to 1000. Computation of the both stationary distributions of the number of customers and of the sojourn time for $N = 500$ required about 10 min.

In Figs. 1–5, we present the dependencies of the following performance measures of the system: the average number $\widehat{N}$ of busy servers, the probability $P_{ent-loss}$ of an arbitrary customer loss at the entrance to the system, the probability $P_{imp-loss}$ of an arbitrary customer loss due to impatience, the loss probability P_{loss} of an arbitrary customer and the average sojourn time v_1 of an arbitrary customer on the server capacity N, which is varied from 1 to 150. The presented above analysis of stationary characteristics of the system seems to be not easy. However, the computation time is not very long. Computation for 150 values of N required only 5 min 11 s. Nevertheless, the natural question may arise: whether or not it is possible to approximate these performance

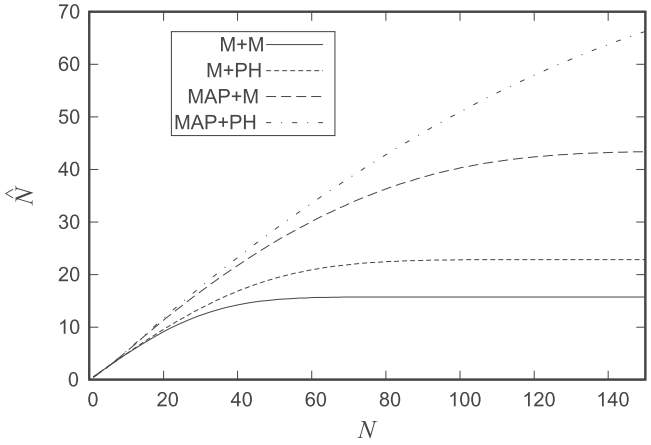

Fig. 1. Dependence of the average number $\widehat{N}$ of customers in service on the server capacity N.

measures by the corresponding measures of the simplified models. Besides the original model with the *MAP* arrival process and the *PH* service time distribution (we code further this model as $MAP + PH$), we consider three approximate models. The first one, coded as $M + PH$, assumes that arrivals occur according to the stationary Poisson process with the same rate $\lambda = 10$ as the initial *MAP* process. Two other models assume the following reasonable simple approximations of the service and impatience processes. It is assumed that the service time distribution is exponential with the same rate $\mu = 10$ as the original Erlangian service time distribution. The impatience rate α of an arbitrary customer is assumed to be fixed as the constant, $\alpha = 0.024969$, computed as the weighted sum of the rate 0 during the first phase of service and rate 0.05 during the second phase of service with the weights $\frac{20.05}{40.05}$ and $\frac{20}{40.05}$ defining the invariant probability vector of the generator $S - \mathrm{diag}\{\boldsymbol{\alpha}\} + (\mathbf{S}_0 + \boldsymbol{\alpha})\boldsymbol{\beta}$. The model with the *MAP* arrivals and approximated service is coded as $MAP + M$ and the model with the stationary Poisson process and approximated service is coded as $M + M$.

It is evident from Figs. 1–5, that the approximations, especially the simplest approximation $M + M$, may be quite poor. Bad feature of all approximations is that they are too optimistic under the given set of parameters of the system. E.g., based on $M + M$ approximation one may expect that the average sojourn time for $N = 140$ is less than 1.8 while indeed this time is more than 6.5. This justifies the analysis of the model with the *MAP* arrival process and the *PH* service time distribution if the precise evaluation of the system performance measures is required.

Fig. 2confirms an intuitively clear fact that the probability $P_{ent-loss}$ is the decreasing function of N while Fig. 3 confirms the fact that the

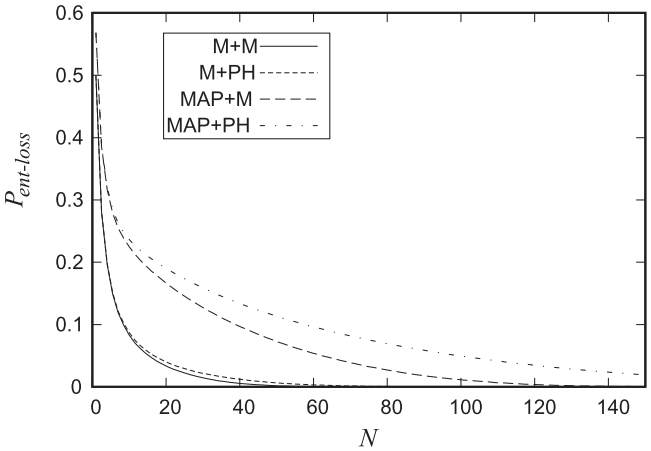

Fig. 2. Dependence of the probability $P_{ent-loss}$ of an arbitrary customer loss at the entrance to the system on the server capacity N.

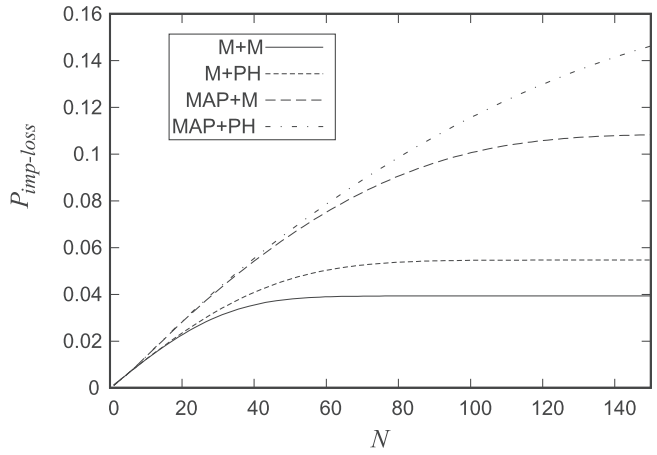

Fig. 3. Dependence of the probability $P_{imp-loss}$ of an arbitrary customer loss due to impatience on the server capacity N.

probability $P_{imp-loss}$ is the increasing function of N.

The probability P_{loss} of an arbitrary customer loss is equal to the sum of the probabilities $P_{ent-loss}$ and $P_{imp-loss}$. This probability is large when N is small. When N increases, this probability decreases. However, it can be verified that there exist points where the probability P_{loss} achieves the minimal value. E.g., for the $MAP + PH$ case such a minimal value is equal to 0.1645 and it is achieved when $N = 119$. When $N > 119$, the probability P_{loss} starts the increasing when N grows. E.g., for $N = 150$, $P_{loss} = 0.1655$.

To more convincingly demonstrate a possibility of using the obtained results for optimization purposes, we formulate the problem to find the optimal value of N, which minimizes the following cost function:

$$E(N) = a\lambda P_{ent-loss} + b\lambda P_{imp-loss}$$

where a and b are the charges paid per unit of time because of one customer loss at the entrance and due to impatience, correspondingly. We suppose that $b > a$, i.e., it is better to reject a customer from the early beginning than admit this customer for service, waste some system resources for its service but eventually lose this customer due to too long service time. Let as fix $a = 0.1$ and $b = 0.3$.

Fig. 6illustrates the dependence of the cost function $E(N)$ on N and demonstrates the existence of optimal values N^* of the system capacity. The optimal value N^* for the $MAP/PH/1$ system is equal to 14 and the optimal value of the cost function is $E(14) = 0.27992$. For the $MAP + M$ approximation, $N^* = 22$ and the optimal value of the cost function is

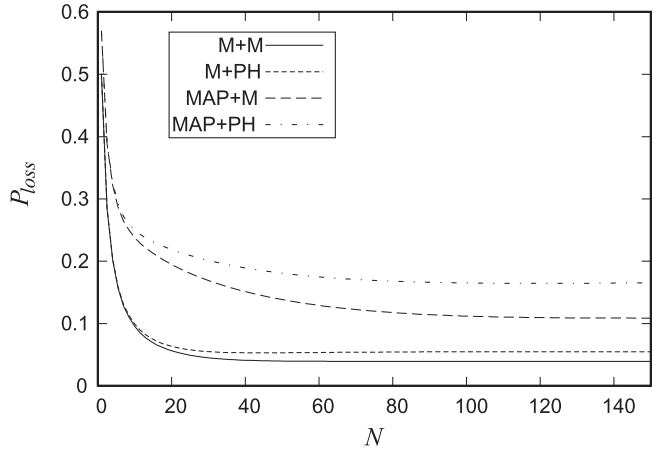

Fig. 4. Dependence of the probability P_{loss} of an arbitrary customer loss on the server capacity N.

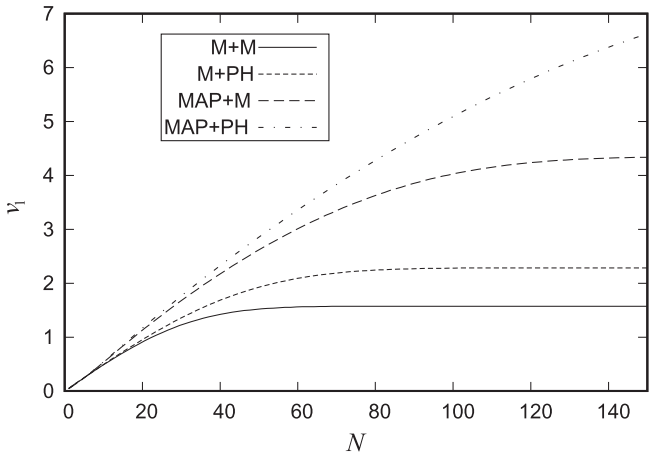

Fig. 5. Dependence of the average sojourn time v_1 of an arbitrary customer in the system on the server capacity N.

$E(22) = 0.250873$. For the $M + PH$ approximation, $N^* = 17$ and $E(17) = 0.10973$. For the $M + M$ approximation, $N^* = 20$ and $E(20) = 0.101739$. Again, we can conclude that the approximations give too optimistic prediction of the value of the cost function and the biased estimation of the optimal system capacity N^*. Another evident conclusion is that the choice of the system capacity in the proper way may provide an essential reduction of the value of the cost function.

Remark 4. It is worth noting that in all our extensive numerical experiments, Little's formula is valid in the form

$$v_1 = \lambda^{-1}\widehat{N}.$$

. Since the computation of the stationary distributions of the number of customers in the system and the sojourn time, respectively, are independent, we have implicitly checked the accuracy of our computational methods. If only the average sojourn time v_1, not the jitter or higher moments of the distribution of the sojourn time, is needed, the use of this relation allows to compute v_1 without the use of Corollary 9.

Above we presented the numerical results for an exponential distribution of the service time (having the squared coefficient of variation equal to 1) and the Erlangian distribution of the service time (having the squared coefficient of variation equal to 0.5). Let us consider the case when the service time has a hyper-exponential distribution. Let $\boldsymbol{\beta} = (0.6, 0.4)$, $S = \mathrm{diag}\{-20, -\frac{40}{7}\}$. As well as the exponential and Erlangian distribution of service time used above, this service time has mean value 0.1. The squared coefficient of variation is equal to 1.75.

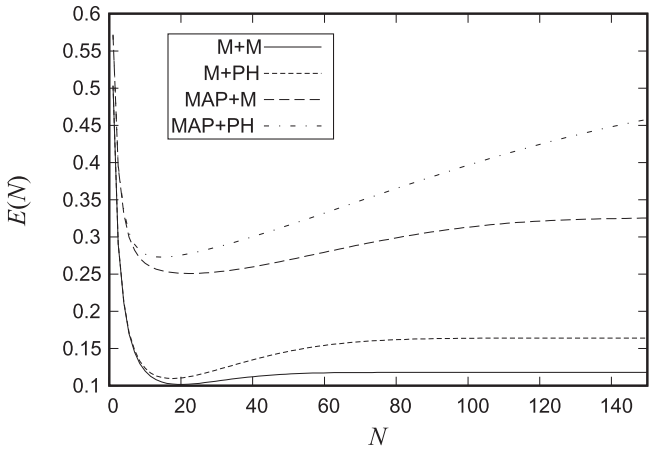

Fig. 6. Dependence of the cost function $E(N)$ on the server capacity N.

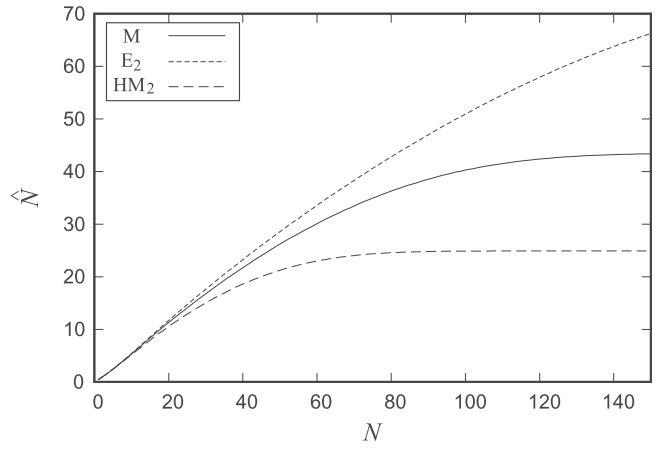

Fig. 7. Dependence of the average number $\widehat{N}$ of customers in service on the server capacity N for service times having different coefficients of variation.

Figs. 7and 8 show the dependence of the average number $\widehat{N}$ of customers in service and the probability P_{loss} of an arbitrary customer loss on the server capacity N for the fixed three distributions of the service time. The symbol M corresponds to the exponential distribution, the symbol E_2 corresponds to the Erlangian distribution, the symbol HM_2 corresponds to the hyper-exponential distribution.

It is seen from these figures that the system operates better when, under the same mean value, the service time has a higher variation. The intuitive explanation of this fact is as follows. The Erlangian distribution has the smallest (among considered three distributions) coefficient of variation. Therefore, the service time of an arbitrary customer is more or less close to its mean value 0.1. The Hyper-exponential distribution considered in this example assumes that about 60 % of customers have the service time around 0.05 (twice less than the mean value) and 40 % of customers have the service time around 0.175. Large percentage of customers, which obtain quick service, implies, in average, the presence in the system of the smaller number of customers and lower loss probability. Note also that, as it is seen from Fig. 7, for $N = 150$ the average number of customers in the system for the service times having the hyper-exponential, exponential and Erlangian distribution is less than 25, about 43 and more than 65, correspondingly. Definitely, the difference is large. Therefore, careful account of variation of the service time is very important for exact evaluation of system performance and assumption that this distribution is exponential made in the overwhelming majority of the relevant literature may cause big errors.

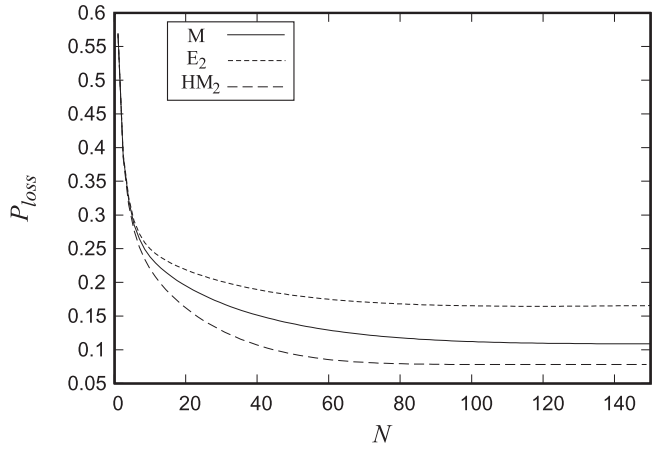

Fig. 8. Dependence of the probability P_{loss} of an arbitrary customer loss on the server capacity N for service times having different coefficients of variation .

Table 1
Information about the time required for multiplication of two square random double-precision matrices of fixed dimension using CPU and GPU.

Matrices dimension	CPU computation time	GPU computation time
1000	0.184 s	0.519 s
2000	1.722 s	0.769 s
4000	16.29 s	2.26 s
6000	56.23 s	6.04 s
8000	2 m 13 s	13.4 s
10,000	4 m 20 s	25 s

5.1. Advisability of using the graphics processing unit (GPU)

Algorithms 1 and 2 presented above operate with matrices the size of which may be large. In the numerical examples described above, we considered the *PH* distribution of the service time of order 2. Therefore, the maximum size of the blocks of the generator Q was $2(N + 1)$ where N is capacity of the server. If $N = 25$, this size is equal to 52.

Let now consider the *PH* distribution of the service time of order 4 having the same average service time equal to 0.1. Namely, we assume that the distribution is defined by the vector $\boldsymbol{\beta} = (1, 0, 0, 0)$ and the sub-generator $S = \begin{pmatrix} -40 & 40 & 0 & 0 \\ 0 & -40 & 40 & 0 \\ 0 & 0 & -40 & 40 \\ 0 & 0 & 0 & -40 \end{pmatrix}$. The vector $\boldsymbol{\alpha}$ defining the rates of customer impatience at various phases of service is fixed as $\boldsymbol{\alpha} = (0, 0, 0.25, 0.25)$. For this distribution of the service time, the maximum size of the blocks of the generator Q is $2\frac{(N+3)!}{N!3!}$. E.g., if $N = 25$, this size is equal to 6552. The work with matrices having large size may be quite slow. To accelerate calculation for operations (especially multiplication and inversion) with matrices, having large dimension, we used graphics processing unit (GPU) instead of central processing unit (CPU).

To justify such a choice, we compared the required computation time using CPU and GPU. Table 1 contains the information about the time required for multiplication of two square random double-precision matrices of fixed dimension using CPU and GPU. Note that for matrix multiplication on CPU we used multi-threaded multiplication with 12 threads. Time indicated for GPU indeed includes also time required for memory allocation and data transfer. Table 2 contains the information about the time required for the inversion of an arbitrary nonsingular double-precision matrix using CPU and GPU.

One can see from Tables 1 and 2 that using GPU is not reasonable for small size of matrices (less than 1500 for multiplication of matrices and less than 1200 for matrix inversion). However, for the larger size of matrices the use of GPU allows to significantly reduce the computation time comparing to the use of CPU. For block size 10,000, calculation time using GPU is about 10 times less than calculation time using CPU.

The maximal size of the block in the considered in this example model with *PH* distribution of the service time of order 4 is equal to 6552. The required time for computation of the stationary distribution of the system states and performance measures using CPU is equal to 18 min 11 s, while computation time using GPU is equal to 3 min 47 s, i.e., 4.8 times less. Therefore, it is reasonable to use GPU for

Table 2
Information about the time required for inversion of an arbitrary nonsingular double-precision matrix using CPU and GPU.

Matrices dimension	CPU computation time	GPU computation time
1000	0.82 s	1.15 s
2000	8 s	2 s
4000	1 m 15 s	11 s
6000	4 m 13 s	30 s
8000	9 m 55 s	1 m 7 s
10,000	19 m 22 s	2 m 6 s

Table 3
Probabilities $\boldsymbol{\pi}_i\mathbf{e}$, $i = \overline{0, 25}$, in the case of correlated and non-correlated arrival processes.

Number of customers i	Probability $\pi_i\mathbf{e}$ in the case of correlated arrival process	Probability $\pi_i\mathbf{e}$ in the case of stationary Poisson arrival process
0	0.19823123646679572	0.04040360491887042
1	0.01467929518482118	0.04039349191580746
2	0.011357331559540544	0.04036445739273102
3	0.011440149771210602	0.04031655680656745
4	0.01159098550418136	0.04024987212429469
5	0.011757519064994575	0.04016451158498454
6	0.011948953764007276	0.04006060933020177
7	0.01217705050755949	0.03993832483934601
8	0.012456440455182377	0.0397978420690468
9	0.012806065849807506	0.039639368140455745
10	0.0132509400235982	0.03946313133716315
11	0.013824426588429277	0.03926937805755197
12	0.014571272778732654	0.03905836819082612
13	0.015551692426169758	0.038830368129018335
14	0.016846892136214948	0.0385856402481709
15	0.01856657751378666	0.03832442713168408
16	0.020859181920464834	0.038046927981591844
17	0.02392585755438428	0.03775326344661735
18	0.02803970638002532	0.0374434233198957
19	0.033572389451116666	0.037117189001855
20	0.04103128267164104	0.03677401901564072
21	0.05111200725878111	0.03641288092319468
22	0.06477394005346315	0.0360320065222698
23	0.08335112174734381	0.03562853928936344
24	0.10871959758003577	0.03519803437595426
25	0.14355808578771082	0.03473376390689619

computations in the described example. It is worth to note that the reasonability of using GPU depends on the concrete configuration of a computer and programming language.

Table 3contains the values of the probabilities $\pi_i\mathbf{e}$ that i customers obtain service in the system at an arbitrary time moment, $i = \overline{0, 25}$, for the system with the *MAP* having the coefficient of correlation 0.2 described above and the *PH* distribution of the service time of order 4. These values are presented in the second column of Table 3. The third column contains the values of the corresponding probabilities in the case when the arrival flow has the same average arrival rate, but the inter-arrival times are not correlated.

The values of the key performance measures of the system with the correlated arrival process are the following:

- the average number of customers receiving service at an arbitrary moment is $\widehat{N} = 15.042045035003023$,
- the intensity of output of customers that obtained service in the system is $T = 7.902110155717249$,
- the probability of an arbitrary customer loss is $P_{loss} = 0.20977585778138863$,
- the probability of a customer loss at the entrance to the system is $P_{ent-loss} = 0.19123967707217554$,
- the probability of a customer loss due to impatience is $P_{imp-loss} = 0.018536180709213094$.

It is evidently seen from Table 3 that the most probable numbers of customers in the system with the *MAP* having correlation 0.2 are 0 (the system is empty) and N capacity of the system is exhausted. The probability of empty system is 0.19823. The probability of the full system is 0.14356. The reason of this effect is the correlation in the arrival process. Positive correlation of inter-arrival times implies that the time intervals when customers arrive rarely (and the server starves) alternate with periods when a lot of customers arrives and the server's capacity is maximally used.

The values of the probabilities $\pi_i\mathbf{e}$ for the system with the stationary Poisson arrival process with the same average arrival rate given in the third column show that the distribution of the number of customers in service is almost uniform. Probabilities $\pi_i\mathbf{e}$ that i customers obtain service monotonically decrease from 0.040403 for $i = 0$ to 0.034734 for $i = 25$.

The values of the main performance measures of the system with the stationary Poisson arrival process are as follows:

- $\widehat{N} = 12.16929103626954$,
- $T = 9.501722257254754$,
- $P_{loss} = 0.04982777427452467$,
- $P_{ent-loss} = 0.0347337639068962$,
- $P_{imp-loss} = 0.015094010367629774$.

These results again confirm the importance of account of correlation and variance of inter-arrival times. The existence of correlation and high variance causes much higher loss probability and larger average number of customers in the system.

6. Conclusion

The problem of computing the steady-state distributions of the number of customers and the sojourn time distribution for the system with the Markovian arrival process, phase type distribution of the service time, limited processor sharing discipline and impatient customers is solved in this paper. We compared two possible approaches for description of the system behavior by the multi-dimensional Markov chain and illustrated the advantage of the approach which suggests account of the number of customers receiving service at each phase. Advisability of using in computations the graphics processing unit is discussed in brief. The results can be used for managerial decisions relating to organization of multiplexing in various communication networks with account of possible variability of intensities of arrival and service processes. As a possible interesting generalization of the considered model for a future research, we can mention combination of the limited processor sharing system as a model of operation of the cell of the wireless network with the queueing model where the shares of the bandwidth dedicated to service of customers are not equal but depend on the distance from the customer to the base station of the cell, see [8].

Acknowledgments

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

References

[1] Asmussen S. Applied probability and queues. New York: Springer-Verlag; 2003.
[2] Baumann H, Sandmann W. Numerical solution of level dependent quasi-birth-and-death processes. Procedia Comput Sci 2010;1:1561–9.
[3] Baumann H, Sandmann W. Multi-server tandem queue with markovian arrival process, phase-type service times, and finite buffers. Eur J Oper Res 2017;256:187–95.
[4] Buchholz P, Kemper P, Kriege J. Multi-class markovian arrival processes and their parameter fitting. Perform Eval 2010;67:1092–106.
[5] Buchholz P, Kriege J. Fitting correlated arrival and service times and related queueing performance. Queueing Syst 2017;85:337–59.
[6] Chakravarthy SR. The batch markovian arrival process: a review and future work. In: Krishnamoorthy A, editor. Advances in probability theory and stochastic processes: Proc. NJ: Notable Publications; 2001. p. 21–49.
[7] Dudin S, Dudin A, Dudina O, Samouylov K. Analysis of a retrial queue with limited processor sharing operating in the random environment. Lect Notes Comput Sci 2017;10372:38–49.
[8] Dudin S, Kim CS. Analysis of multi-server queue with spatial generation and location-dependent service rate of customers as a cell operation model. IEEE Trans Commun 2017;65:4325–33.
[9] Gaver D, Jacobs P, Latouche G. Finite birth-and-death models in randomly changing environments. Adv Appl Probab 1984;16:715–31.
[10] Ghosh A, Banik AD. An algorithmic analysis of the *BMAP/MSP*/1 generalized processor-sharing queue. Comput Oper. Res 2017;79:1–11.
[11] Graham A. Kronecker products and matrix calculus with applications. Cichester: Ellis Horwood; 1981.
[12] He QM, Alfa AS. Space reduction for a class of multidimensional markov chains: a summary and some applications. INFORMS J Comput 2018;30:10.
[13] Kesten H, Runnenburg JT. Priority in waiting line problems. Amsterdam:

Mathematisch Centrum; 1956.
[14] Kim CS, Dudin S, Taramin O, Baek J. Queueing system $MMAP/PH/n/n+r$ with impatient heterogeneous customers as a model of call center. Appl Math Model 2013;37:958–76.
[15] Kim CS, Mushko VV, Dudin A. Computation of the steady state distribution for multi-server retrial queues with phase-type service process. Ann Oper Res 2012;201:307–23.
[16] Kleinrock L. Queueing systems. computer applications. New York: Wiley; 1976.
[17] Lucantoni D. New results on the single server queue with a batch markovian arrival process. Commun in Statistics-Stochastic Models 1991;7:1–46.
[18] Li QL, Lian Z, Liu L. An RG-factorization approach for a $BMAP/m/1$ generalized processor-sharing queue. Stochastic Models 2005;21:507–30.
[19] Moscholios ID, Vassilakis VG, Logothetis MD, Boucouvalas AC. State-dependent bandwidth sharing policies for wireless multirate loss networks. IEEE Trans Wireless Commun 2017;16:5481–97.
[20] Masuyama H, Takine T. Sojourn time distribution in a $MAP/m/1$ processor-sharing queue. 2003;31:406–12.
[21] Neuts M. Matrix-geometric solutions in stochastic models. Baltimore: The Johns Hopkins University Press; 1981.
[22] Okamura H, Dohi T. Mapfit: an r-based tool for PH/MAP parameter estimation. Lect Notes Comput Sci 2015;9259:105–12.
[23] Pattavina A, Parini A. Modelling voice call interarrival and holding time distributions in mobile networks. Proceedings of the 19th international teletraffic congress (ITC5). 2005. p. 729–38.
[24] Ramaswami V. Independent markov processes in parallel. Comm Statist-Stochastic Models 1985;1:419–32.
[25] Ramaswami V, Lucantoni DM. Algorithms for the multi-server queue with phase-type service. Comm Statist-Stochastic Models 1985;1:393–417.
[26] Samouylov KE, Sopin ES, Gudkova IA. Sojourn time analysis for processor sharing loss queuing system with service interruptions and MAP arrivals. Commun Comput Inf Sci 2016;678:406–17.
[27] Sericola B, Guillemin F, Boyer J. Sojourn times in the m/PH/1 processor sharing queue. Queueing Syst 2005;50:109–30.
[28] Telek M, van Houdt B. Response time distribution of a class of limited processor sharing queues. In proceedings of IFIP WG. 7.3 performance conference. 2017. New York City
[29] van Dantzig D. Chaines de markof dans les ensembles abstraits et applications aux processus avec regions absorbantes et au probleme des boucles. Ann de l'Inst H Poincare 1955;14(fasc. 3):145–99.
[30] Vishnevski VM, Dudin AN. Queueing systems with correlated arrival flows and their applications to modeling telecommunication networks. Autom Remote Control 2017;78:1361-1403.
[31] Yang C, Yao Y, Chen Z, Xia B. Analysis on cache-enabled wireless heterogeneous networks. IEEE Trans Wireless Commun 2016;15:131–45.
[32] Yashkov S. Processor-sharing queues: some progress in analysis. Queueing Syst 1987;2:1–17.
[33] Yashkov S, Yashkova A. Processor sharing: a survey of the mathematical theory. Autom Remote Control 2007;68:1662–731.
[34] Zhen Q, Knessl C. On sojourn times in the finite capacity $m/m/1$ queue with processor sharing. Oper Res Lett 2009;37:447–50.