Journal of Ambient Intelligence and Humanized Computing
https://doi.org/10.1007/s12652-022-04233-w

ORIGINAL RESEARCH

# Priority queueing system with many types of requests and restricted processor sharing

Ciro D'Apice[1] · Alexander Dudin[2] · Sergey Dudin[2] · Rosanna Manzo[3]

Received: 14 May 2022 / Accepted: 22 June 2022


**Abstract**
A priority queueing model with many types of requests and restricted processor sharing is considered. A novel discipline of requests admission and service is proposed. This discipline assumes restriction of the bandwidth (capacity) of the server and the number of requests that can receive service in the system at the same time. This discipline is some kind of realistic hybrid of the traditional discipline of service in a multi-server system and the discipline of the limited processor sharing. The requests of the highest priority can push out from the service the low priority requests. Therefore, the important problem is fitting of the number of requests that can receive service at the same time to the bandwidth of the server. This problem is solved via construction and analysis of a multi-dimensional Markov chain describing operation of the system under any fixed set of the system parameters.

**Keywords** Multi-server priority queueing model · Marked Markov arrival process · Processor sharing · Multi-dimensional Markov chains

## 1 Introduction

Queueing theory is a widely acknowledged mathematical tool for optimal solution of the task of a restricted resource distribution among the competing requests of users. The simplest models assume that service to customers is provided in a certain order sequentially, one-by-one. More general models suggest a possibility of some kind of resource sharing and simultaneous service of several requests at the same time. The two most popular disciplines for managing the simultaneous service of several requests are as follows: **A**—Resource is divided into several parts (called as servers) and each request receiving service uses the assigned to him/her server. The service times of requests are independent. We call the system with such a discipline as a multi-server system; **B**—Resource is jointly used by all requests and the service rate is in inverse ratio to the number of requests receiving service. This discipline is called as processor sharing (PS). For surveys of the research related with this discipline and some its generalizations, see (Yashkov 1987; Yashkov and Yashkova 2007; Altman et al. 2006).

The overwhelming majority of the existing research, starting from the pioneering works by A.K. Erlang, assumes discipline **A**. Queueing models of $GI/PH/N/K$, $BMAP/PH/N/K$ types (in D.G. Kendall's notation) with infinite or finite buffers, losses, retrials and their partial cases are investigated in enough full extend, especially in the case when service time has an exponential distribution. The models with an arbitrary ($G$) distribution of the service time are investigated only approximately or

✉ Rosanna Manzo
rmanzo@unisa.it

Ciro D'Apice
cadapice@unisa.it

Alexander Dudin
dudin@bsu.by

Sergey Dudin
dudin85@mail.ru

[1] Dipartimento di Scienze Aziendali - Management and Innovation Systems, University of Salerno, Via Giovanni Paolo II, 132, Fisciano 84084, Salerno, Italy

[2] Department of Applied Mathematics and Computer Science, Belarusian State University, 4, Nezavisimosti Ave., Minsk 220030, Republic of Belarus

[3] Department of Information Engineering, Electrical Engineering and Applied Mathematics, University of Salerno, Via Giovanni Paolo II, 132, Fisciano 84084, Salerno, Italy

Published online: 13 July 2022

asymptotically. In particular, certain bounds are obtained for some performance characteristics.

An advantage of the discipline **A** is relative easiness of its practical realization. E.g., in a call center, several operators can provide service to users using the separate workstations and communication channels. In information transmission systems, the physical resource, e.g., bandwidth of the channels, can be divided (using various technical schemes like frequency division, time division, code division multiplexing, etc.) into logical channels (servers) each of which is assigned to service of a separate request. The evident disadvantage of discipline **A** is possible under-utilization of the resource. Situations occur when only a few logical channels are busy, while the rest are staying idle.

The PS discipline is free from this disadvantage. The resource (let us call it further as the bandwidth) is always fully used if there are requests for service. However, the essential two disadvantages of PS discipline, besides more difficult technical implementation related to the necessity of dynamic redistribution of the bandwidth, in many concrete applications are as follows:

(i) situations are possible when currently presenting in the system requests do not need in total the whole bandwidth. E.g., if five presenting users need transmission of their information at rate 10 megabits per second (Mbps), they do not need to fully share the available bandwidth of 100 Mbps channel. They will use in total only 50 Mbps;

(ii) there may exist some minimal requirement of the user to the bandwidth assigned to him/her. E.g., the users may require the bandwidth for video on-demand transmission of HD content an MPEG2 transport stream as 12 Mbps and do not agree to use the smaller bandwidth due to poor quality of service. Therefore, the number of simultaneously serviced users in 100 Mbps channel must be less than 9. Thus, the pure PS discipline that assumes that all arriving requests are accepted for service is not applicable for modeling the considered transmission process.

As a tool to overcome the disadvantage (ii), the discipline of limited PS (LPS) was offered in the literature. This discipline suggests that the number of users, which simultaneously use the bandwidth, is limited by some finite number, say $N$. This number is called sometimes as a multiprogramming level, see, e.g., (Nair et al. 2010), or concurrency limit, see, e.g., (Gupta and Zhang 2022). For relevant references, see also, e.g., (Alencar et al. 2021; Telek and Van Houdt 2018; Samouylov et al. 2016; Dudin et al. 2017; Masuyama and Takine 2003; Dudin et al. 2021; Ghosh and Banik 2017; Bocharov et al. 2007; Brugno et al. 2017, 2018; D'Arienzo et al. 2020; Kim et al. 2019).

Contributions of our paper consist of the following.

- We propose a new, hybrid, discipline called as a restricted processor sharing. This discipline combines the positive features of disciplines **A** and **B**. As in both, **A** and LPS, disciplines, we suppose that the maximum number of requests that can receive the service at the same time is an integer number $N$, $N < \infty$. In **A** discipline, $N$ corresponds to the number of servers. In LPS discipline, $N$ corresponds to the concurrency limit. If an arriving request meets $N$ requests in service, in this paper we assume that it is lost. The variants when such a request is queued into the infinite or finite buffer or will make the retrials are left for the future research based on the results of presented below analysis of the system with loss of requests. The derived expressions of the blocks of the generator of the multidimensional Markov chain describing behavior of the system can be used as the bricks for derivation of the form of the generator of the Markov chain that describes the dynamics of the system with buffers and retrials. The hybrid discipline assumes that a required amount of work (amount of information) and a required (nominal) rate of service are associated with each request. The total rate of service of all requests staying in the system is restricted by the parameter $B$ called the bandwidth of the server. If the sum of nominal service rates of all requests staying in the system does not exceed the bandwidth, all requests receive service independently of each other with the nominal rate, as in discipline **A**. If the sum of nominal service rates of requests staying in the system exceeds the bandwidth, all requests receive service at the proportionally reduced rate, as in discipline LPS.
- We suggest that the requests are heterogeneous in respect to their importance and the required bandwidth and the nominal service. There is a finite number $M$ types of requests. Different types of requests have different priorities. One of the types of requests has a preemptive priority over requests of other types. Arrival of such a request when the number of requests obtaining service is equal to $N$ implies the loss of one of the requests having the lowest priority among presenting in the system, if any. To reduce probability of interruption of service of low priority requests, they are not accepted to the system when the number of requests receiving service is less than $N$ but exceeds a certain threshold value. The considered model can have a wide field of applications. The particular case when there are only two types of requests well fits for modelling the system of cognitive radio. Type-1 requests are sent by the primary (licensed) users. Type-2 requests are sent by the cognitive (secondary) users. It is worth to note that the existing in the literature models, see, e.g., (El-Toukhy and Arslan 2019; Goel and

Kulshrestha 2022; Sun et al. 2014b, a; Lee et al. 2022), of cognitive radio systems are the special cases of our model with $M = 2$ and absence of possibility of service of requests with the reduced rate.

- While the overwhelming majority of more or less relevant papers assume that flows of the requests are defined as the stationary Poisson arrival process, here we assume that the arriving heterogeneous flow is described by the Marked Markovian Arrival Process (MMAP) (see, e.g. (He 1996)). This allows to adequately account bursty nature (high variability and dependence of consecutive inter-arrival times) which is the inherent feature of information flows in various modern telecommunication network, contact centers, etc, see, e.g., (Chen et al. 2022) where the information about the real flows traces is presented. It is worth noting that the use of stationary Poisson arrival process as a model of real-life process usually implies too optimistic estimates of the system performance indicators.

The reminder of the paper is as follows. In Sect. 2, the considered mathematical model is described. The Markovian process describing behavior of the model under study is defined and analysed in Sect. 3. Expressions for computation the basic performance indicators of the system are given in Sect. 4. In Sect. 5, an numerical example is presented. Section 6 contains brief conclusion of the paper.

## 2 Mathematical model

We consider a queuing system with a restricted processor sharing discipline. The scheme of the system is shown in Fig. 1.

Incoming to the system requests are divided into $M$ types. The arrival of requests is described by the *MMAP*, see (He 1996). Arrival can occur only at the epochs of transitions of underlying Markov process denoted by $v_t$, $t \geq 0$. This process is a continuous-time Markov chain having a finite state space $\{1, \dots, W\}$. The rates of transitions of this Markov chain are determined by the irreducible generator $D$. The matrix $D$ is split into $M + 1$ matrix summands $D_r$, $m = \overline{0, M}$ : $D = \sum_{m=0}^{M} D_m$. Elements of the matrix $D_m$ determine the rates of transitions of the Markov chain $v_t$, which are accompanied by the generation of the type $m$ request, $m = \overline{1, M}$. Non-diagonal entries in the matrix $D_0$ determine the rate of transitions of the Markov chain $v_t$, which are not accompanied by the generation of a request. The modules of the negative diagonal elements of the matrix $D_0$ determine the rate of the exit of the Markov chain $v_t$ from the corresponding states.

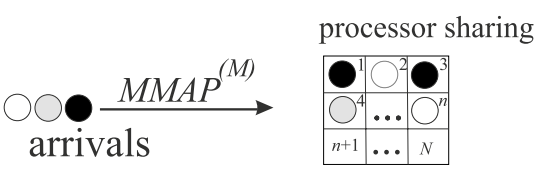


**Fig. 1** System structure

The mean intensity $\lambda_m$, $m = \overline{1, M}$, of arrival of requests of type $m$ is given by $\lambda_m = \boldsymbol{\theta} D_m \mathbf{e}$, where $\boldsymbol{\theta}$ is a row vector of invariant probabilities of the process $v_t$. This vector is defined as the unique solution of the system of linear algebraic equations $\boldsymbol{\theta} D = \mathbf{0}$ with the normalization condition $\boldsymbol{\theta}\mathbf{e} = 1$. Here $\mathbf{e}$ is a column vector of a proper size, consisting of 1s, and $\mathbf{0}$ is a row vector consisting of 0s. The notation $m = \overline{1, M}$ means that the parameter $m$ admits values $1, \dots, M$. The total intensity of requests $\lambda$ is defined as $\lambda = \sum_{m=1}^{M} \lambda_m$.

We interpret the service of requests as the transfer of a certain amount of information. The bandwidth of the server defined as the maximum number of megabits that can be transmitted per unit of time is denoted as $B$. We assume that the bandwidth of the server is used by all requests. The maximum possible number of simultaneously served requests is limited by the parameter $N$. It is assumed that the amount of information to be transmitted to serve a single request of type $m$ has an exponential distribution with rate $\alpha_m$, $m = \overline{1, M}$. The value of $\alpha_m^{-1}$ represents the average data volume of a request of type $m$, $m = \overline{1, M}$. We assume that requests of different types require the different service intensity. Denote by $\hat{\beta}_m$ the bitrate desired for requests of type $m$ (nominal bitrate). Therefore, the nominal service time of a request of type $m$ is $(\hat{\beta}_m \alpha_m)^{-1}$. Accordingly, the nominal service intensity $\beta_m$ of a request of type $m$ is calculated as $\beta_m = \hat{\beta}_m \alpha_m$, $m = \overline{1, M}$. The desired bitrate (nominal service intensity) is provided to any request when there is no shortage of bandwidth of the server, i.e. the sum of the desired bitrates of all requests, which receive service, does not exceed the bandwidth of the server. Otherwise, the bitrates provided to all requests are correspondingly reduced.

We assume that requests have different priorities. Requests of the first type have the highest priority, ..., requests of the type $M$ have the lowest priority. This means the following. First of all, we will assume that requests of type $m$, $m = \overline{2, M}$, are not accepted into the system if the number of already serviced requests is equal to or exceeds the parameter $N_1$. This means that $N - N_1$ places are reserved specifically for servicing requests of the first type. If a type 1 request arrives when the number of requests receiving service is equal to $N$ or the request of type $m$, $m = \overline{2, M}$, arrives when the number of requests receiving service is not less than $N_1$ and there are requests with a lower priority on the service, then the arriving request displaces one of the serviced requests with the lowest priority and starts the service. The displaced request is lost.

## 3 The process describing dynamics of the system

Let $n_t$, $n_t = \overline{0,N}$, be the number of requests on service, and $s_t^{(m)}$ be the number of requests of type $m$ receiving service at moment $t$ such as $s_t^{(m)} = \overline{0,n_t}$, $\sum_{m=1}^{M} s_t^{(m)} = n_t$. Because the bandwidth sharing discipline is applied, *the actual* service intensity of the request is equal to its nominal service rate only if the used at time $t$ bandwidth, which is defined as $\sum_{k=1}^{M} s_t^{(k)} \hat{\beta}_k$, is less than the bandwidth $B$ of the server. Otherwise, the service rate of type $m$ request is cut and equals to $\frac{B}{\sum_{k=1}^{M} s_t^{(k)} \hat{\beta}_k} \beta_m$, $m = \overline{1,M}$.

It is obvious that the $(M+2)$-dimensional random process

$$\zeta_t = \{n_t, v_t, s_t^{(1)}, \dots, s_t^{(M)}\}, \ t \geq 0,$$

where

$$n_t = \overline{0,N}, v_t = \overline{1,W}, \ s_t^{(m)} = \overline{0,n_t}, \ \sum_{m=1}^{M} s_t^{(m)} = n_t,$$

completely describes the behavior of the considered queuing system and is a regular continuous-time Markov chain.

Since this Markov chain is irreducible and has a finite state space, it is known that the limits

$$p(n, v, s^{(1)}, \dots, s^{(M)}) = \lim_{t\to\infty} P\{n_t = n, v_t = v, s_t^{(1)} = s^{(1)}, \dots, s_t^{(M)} = s^{(M)}\}$$

exist for any values of the system parameters. They are called as the stationary probabilities of the system states or steady-state probabilities.

To simplify analysis of the multi-dimensional Markov chain, it is useful to combine the set of states of the process $\zeta_t$ having the value $n$ of the component $n_t$, into so called *level* $n$, $n = \overline{0,N}$. For certainty, we number the states, which belong to the level $n$, in the lexicographic order of the component $v_t$ and the reverse lexicographic order of the components of the $M$-dimensional process $\mathbf{s}_t = (s_t^{(1)}, \dots, s_t^{(M)})$.

In accordance with this enumeration, we combine the stationary probabilities of the states that belong to the level $n$ into the row vectors $\mathbf{p}_n$, $n = \overline{0,N}$. These vectors satisfy the system of linear algebraic equations (balance equations)

$$(\mathbf{p}_0, \mathbf{p}_1, \dots, \mathbf{p}_N)A = \mathbf{0}, \tag{1}$$

where $A$ is the infinitesimal generator of the Markov chain $\zeta_t$ and the normalization condition:

$$(\mathbf{p}_0, \mathbf{p}_1, \dots, \mathbf{p}_N)\mathbf{e} = 1.$$

For solving this system, it is necessary to obtain the generator $A$. On this way, the most difficult particular problem is to describe the transition intensities of the components of the $M$-dimensional process $\mathbf{s}_t$ which determines the current number of each type requests in the system. To compute these intensities, first we need to formally define the process of a request service when the system is not overcrowded and the request permanently receives the nominal required service rate. Analyzing various scenarios, one can make sure that service time of such a request has so-called the generalized phase-type (*GPH*) distribution, see (Dudin et al. 2016). Such a distribution is the generalization of the well-known in the literature phase-type distribution (see (Neuts 1981)) to the case of service of heterogeneous requests. The basic idea of the *GPH* distribution is to avoid the monitoring of the type of each request during its service. It is achieved via the use of different probability vectors for installing the initial state of the underlying process of service of requests of different types and the common sub-generator for description of transitions of the underlying process of service within its state space. For more details about the *GPH* distribution and examples of its applications, see (Dudin et al. 2016).

As an underlying process $s_t$, $t \geq 0$, of service of an arbitrary request we call the continuous-time Markov chain defined as follows. The state space of this chain is the set of integers $\{1, \dots, M\}$. The initial state of the chain $s_t$ at the epoch of a request service beginning is randomly chosen with the probabilities defined as the components of the probability vector $\mathbf{b}_m$ given by

$$\mathbf{b}_m = (\underbrace{0, \dots, 0}_{m-1}, 1, \underbrace{0, \dots, 0}_{M-m}), \ m = \overline{1,M},$$

if this the request is of type $m$. The rates of transition of the Markov chain $s_t$ to the absorbing state are determined by the column vector $-\mathbf{\Omega e}$ where the sub-generator $\mathbf{\Omega}$ is defined by formula $\mathbf{\Omega} = -\mathrm{diag}\{\beta_m, \ m = \overline{1,M}\}$, where $\mathrm{diag}\{\dots\}$ means a diagonal matrix having the diagonal elements specified in parenthesis.

Having defined the service time distribution of a single request, we can describe the intensity of transitions of the multidimensional process $\mathbf{s}_t$. For this purpose, we extend the approach going back to the paper (Ramaswami and Lucantoni 1985). We use the following notation. Conditional that all $n$ requests staying in the system receive service at a nominal (not reduced) rate, let

- the matrix $P_n(\mathbf{b}_m)$ define the transition probabilities of the process $\mathbf{s}_t$ at the service beginning epoch of a new type $m$ request, $n = \overline{0,N-1}$, $m = \overline{1,M}$;
- the matrix $L_n(\boldsymbol{\beta})$, $n = \overline{1,N}$, $\boldsymbol{\beta} = (\beta_1, \beta_2, \dots, \beta_M)$ define the transition intensities of the process $\mathbf{s}_t$ when one of the requests finishes service;

- $T_n = \begin{pmatrix} n+M-1 \\ M-1 \end{pmatrix} = \frac{(n+M-1)!}{n!(M-1)!}$, $n = \overline{1,N}$;
- $E_m^{(n)}$, $m = \overline{2,M}$, if $n = \overline{N_1,N}$, and $m = 1$ if $n = N$, be the square matrices of size $T_n$ whose elements determine the transition probabilities of the process $\mathbf{s}_t$ at epochs of type $m$ request arrival, $m = \overline{2,M}$, when $n$, $n = \overline{N_1,N}$, requests receive service, or a request of the type 1 arrives when $N$ requests are in service, and the arriving request tries to displace from the service a request with a lower priority. Only one element in each row of the matrix $E_m^{(n)}$ is different from zero and equals to 1. To define which entry is equal to 1, we note that each row and column of the matrix $E_m^{(n)}$ correspond to the certain state $\{s_1, s_2, \dots, s_M\}$ of the process $\mathbf{s}_t$, $t \geq 0$. Recall that all states of the process $\mathbf{s}_t$, $t \geq 0$, are numbered in the reverse lexicographic order of the entries $s_t^{(1)}, \dots, s_t^{(M)}$. For example, the first row and column of the matrix $E_m^{(n)}$ correspond to the state of $\{n, 0, 0, \dots, 0\}$, the second row and column correspond to the state of $\{n-1, 1, 0, \dots, 0\}$,⋯, the last row and column correspond to the state of $\{0, 0, 0, \dots, n\}$. In the row of the matrix $E_m^{(n)}$ corresponding to the state $\{s_1, s_2, \dots, s_M\}$, element 1 is placed in the column corresponding to the same state $\{s_1, s_2, \dots, s_M\}$ only if $s_l = 0$ for all $l$, $M \geq l > m$. In this case, the arrived request of type $m$ is lost, since there are no requests of a lower priority in the system. If $s_l > 0$ for some $l$, $M \geq l > m$ and $m^*$ is the maximum of such values $l$, then element 1 is placed in the column corresponding to the state

$$\{s_1, \dots, s_{m-1}, s_m + 1, s_{m+1}, \dots, s_{m*-1}, s_{m*} - 1, 0, \dots, 0\}.$$

In this case, a type $m^*$ request has the lowest priority, and an arriving type $m$ request displaces any type $m^*$ request, which leaves the system (is lost).

A more detailed description of these matrices and the algorithms elaborated to calculate them are presented, for example, in (Kim et al. 2013) and (Kim et al. 2021).

To take into consideration the receiving of *reduced* service rate when the sum of the required by all requests presenting in the system bandwidth is greater than the bandwidth of the server $B$, we need more notation, namely:

- $\mathbf{a}_n = L_n(\hat{\boldsymbol{\beta}})\mathbf{e}$, $n = \overline{1,N}$, where

  $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_M)$;

- $\mathbf{d}_n$ $n = \overline{1,N}$, are the column vectors of dimension $T_n$, whose elements $(\mathbf{d}_n)_i$ are defined as

$$(\mathbf{d}_n)_i = \begin{cases} 1, & \text{if } (\mathbf{a}_n)_i \leq B, \\ \frac{B}{(\mathbf{a}_n)_i}, & \text{otherwise}; \end{cases}$$

- $\mathbf{q}_n$ $n = \overline{1,N}$, are the column vectors of dimension $T_n$, whose elements $(\mathbf{q}_n)_i$ are defined as

$$(\mathbf{q}_n)_i = \begin{cases} 0, & \text{if } (\mathbf{a}_n)_i > B, \\ 1, & \text{otherwise}, \end{cases}$$

- $\mathrm{diag}\{\mathbf{d}_n\}$ is a diagonal matrix with the diagonal elements given by the entries of the vector $\mathbf{d}_n$.

Now we are prepared to present the generator $A$. Since requests enter the system and depart only one at a time, it is clear that the matrix $A$ has the block-tridiagonal structure:

$$A = \begin{pmatrix} A_{0,0} & A_{0,1} & O & \dots & O & O \\ A_{1,0} & A_{1,1} & A_{1,2} & \dots & O & O \\ O & A_{2,1} & A_{2,2} & \dots & O & O \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ O & O & O & \dots & A_{N,N-1} & A_{N,N} \end{pmatrix}.$$

The diagonal elements of the diagonal blocks $A_{n,n}$, $n = \overline{0,N}$, are negative and their modules determine the intensity of the Markov chain $\zeta_t$ departure from the corresponding states. The non-diagonal elements of these blocks are non-negative and determine the transition intensities of the Markov chain inside the level $n$. The elements of the matrices $A_{n,n-1}$, $n = \overline{1,N}$, and $A_{n,n+1}$, $n = \overline{0,N-1}$, are non-negative and determine the transition rates of $\zeta_t$ from level $n$ to the levels $n-1$ and $n+1$, respectively.

**Theorem 1** *The explicit form of the blocks* $A_{n,n'}$, $n, n' = \overline{0,N}$, $\max\{n-1, 0\} \leq n' \leq n+1$, *is as follows*:

$$\begin{aligned}
A_{0,0} =& D_0, \\
A_{n,n} =& D_0 \oplus \mathrm{diag}\{-\mathrm{diag}\{\mathbf{d}_n\}L_n(\boldsymbol{\beta})\mathbf{e}\}, \\
& n = \overline{1, N_1 - 1}, \\
A_{n,n} =& D_0 \oplus \mathrm{diag}\{-\mathrm{diag}\{\mathbf{d}_n\}L_n(\boldsymbol{\beta})\mathbf{e}\} \\
& + \sum_{m=2}^{M} D_m \otimes E_m^{(n)}, \; n = \overline{N_1, N-1}, \\
A_{N,N} =& D_0 \oplus \mathrm{diag}\{-\mathrm{diag}\{\mathbf{d}_N\}L_N(\boldsymbol{\beta})\mathbf{e}\} \\
& + \sum_{m=1}^{M} D_m \otimes E_m^{(N)}, \\
A_{n,n+1} =& \sum_{m=1}^{M} D_m \otimes P_n(\mathbf{b}_m), \; n = \overline{0, N_1 - 1}, \\
A_{n,n+1} =& D_1 \otimes P_n(\mathbf{b}_1), \; n = \overline{N_1, N-1}, \\
A_{n,n-1} =& I_W \otimes \mathrm{diag}\{\mathbf{d}_n\}L_n(\boldsymbol{\beta}), \; n = \overline{1,N},
\end{aligned}$$

*where* $I_W$ *is an identity matrix of size* $W$, $\otimes$ *and* $\oplus$ *denote symbols of Kronecker product and sum of matrices, see, for example*, (*Graham* 2018).

The proof of Theorem 1 is carried out by means of analysis of a Markov chain transitions during an infinitesimal interval and is omitted here. Note that the use of the vector $\mathbf{d}_n$ allows to take into account a decrease of the service rate in case of a shortage of bandwidth.

Chains with a block-tridiagonal structure of the generator are called in the literature as the Level Dependent Quasi-Birth-and-Death processes. The size of system (1) can be large. For solution of such systems, it is recommended to exploit the sparse structure of the generator $A$. E.g., the algorithm from (Baumann and Sandmann 2010) can be used.

## 4 Performance characteristics

Once the vectors $\mathbf{p}_n$, $n = \overline{0,N}$, are calculated, they can be used for computing the values of versatile performance indicators of the analyzed queuing system. Formulas for computation of some performance indicators are presented below.

The mean number of requests in the system is

$$N^{customers} = \sum_{n=1}^{N} n\mathbf{p}_n\mathbf{e}.$$

The rate of the output flow of requests that successfully received service is equal to

$$\mu^{out} = \sum_{n=1}^{N} \mathbf{p}_n(I_W \otimes \mathrm{diag}\{\mathbf{d}_n\}L_n(\boldsymbol{\beta}))\mathbf{e}. \quad (1)$$

The proof of this formula evidently follows from the formula of total probability and equivalent form of formula (1)

$$\mu^{out} = \sum_{n=1}^{N} \mathbf{p}_n A_{n,n-1}\mathbf{e}.$$

Row vectors $\mathbf{p}_n$, $n = \overline{0,N}$, define stationary probabilities of the states of the Markov chain $\zeta_t$ such as the number $n_t$ of the requests in the system is equal to $n$ and the components of the column vectors $A_{n,n-1}\mathbf{e}$ define the rates of successful service completions during the stay of the Markov chain $\zeta_t$ in these states.

The rate of the output flow of type-$m$ requests that received service is equal to

$$\mu_m^{out} = \sum_{n=1}^{N} \mathbf{p}_n(I_W \otimes \mathrm{diag}\{\mathbf{d}_n\}L_n(\boldsymbol{\beta}_m))\mathbf{e},$$

where $\boldsymbol{\beta}_m = \beta_m\mathbf{b}_m$, $m = \overline{1,M}$.

The mean number of type-$m$ requests in the system is

$$N_m^{customers} = \sum_{n=1}^{N} \mathbf{p}_n(I_W \otimes L_n(\mathbf{b}_m))\mathbf{e}, \ m = \overline{1,M}. \quad (2)$$

The proof of this formula is similar to the proof of formula (1). It evidently follows from the formula of total probability with account of the fact that the multiplier $(I_W \otimes L_n(\mathbf{b}_m))\mathbf{e}$ selects only the components of the vector $\mathbf{p}_n$, which account the number of requests of type $m$, and these requests' departure rate is equal to 1. As the result, the sum in the right hand side of (2) defines the mean number of type-$m$ requests in the system.

The probability of an arbitrary request loss at its arrival moment is

$$P^{arrival-loss} = \lambda^{-1}\left(\sum_{n=N_1}^{N}\sum_{m=2}^{M} \mathbf{p}_n(D_m \otimes \tilde{E}_m^{(n)})\mathbf{e} + \mathbf{p}_N(D_1 \otimes \tilde{E}_1^{(N)})\mathbf{e}\right)$$

where $\tilde{E}_m^{(n)}$ is the diagonal matrix having the same diagonal elements as the matrix $E_m^{(n)}$.

The probability of an arbitrary type 1 request loss is

$$P_1^{arrival-loss} = \lambda_1^{-1}\mathbf{p}_N(D_1 \otimes \tilde{E}_1^{(N)})\mathbf{e}.$$

The probability of an arbitrary type $m$ request loss upon arrival is

$$P_m^{arrival-loss} = \lambda_m^{-1}\sum_{n=N_1}^{N} \mathbf{p}_n(D_m \otimes \tilde{E}_m^{(n)})\mathbf{e}, \ m = \overline{2,M}.$$

The probability that at an arbitrary moment there will be a shortage of a bandwidth is equal to

$$P^{sharing} = \sum_{n=1}^{N} \mathbf{p}_n(I_W \otimes \mathrm{diag}\{\mathbf{q}_n\})\mathbf{e}.$$

The probability that all requests at an arbitrary moment receive the required service rate is equal to

$$P^{no-sharing} = 1 - P^{sharing}.$$

Let the square matrix $E_{m,l}^{(n)}$ where $l = \overline{2,M}$, $n = N$, if $m = 1$ and $l = \overline{m+1,M}$, $n = \overline{N_1,N}$ if $m = \overline{2,M-1}$ of size $T_n$ define the transition probabilities of the process $\mathbf{s}_t$, $t \geq 0$, during the moment at which type $m$ request arrives to the system and displaces a type $l$ request when the number of requests receiving service is $n$. Definition of this matrix is similar to definition of the matrix $E_m^{(n)}$ given above. In each row of this matrix only one element can be equal not to zero but to 1. We use the mentioned in definition of the matrix $E_m^{(n)}$ fact that each row and column of the matrix is $E_{m,l}^{(n)}$ correspond to a certain state $\{s_1, s_2, \ldots, s_M\}$ of the process $\mathbf{s}_t$. In the row of the matrix $E_{m,l}^{(n)}$ that corresponds to the state $\{s_1, s_2, \ldots, s_M\}$, element 1 is placed in the column that corresponds to the state $\{s_1, \ldots, s_{m-1}, s_m + 1, s_{m+1}, \ldots, s_{l-1}, s_l - 1, 0, \ldots, 0\}$ only

if $s_m = 0$ for all $m$, $M \geq m > l$, and $s_l > 0$. If this condition is not fulfilled, all entries of this row are null.

The intensity $\lambda^{(m)}_{force-out}$, $m = \overline{2, M}$, of displacement of type $m$ requests is calculated as

$$\lambda_m^{push-out} = \sum_{l=2}^{m-1} \sum_{n=N_1}^{N} \mathbf{p}_n (D_l \otimes E^{(n)}_{l,m})\mathbf{e} + \mathbf{p}_N (D_1 \otimes E^{(N)}_{1,m})\mathbf{e},\ m = \overline{2, M}.$$

The probability $P^{push-loss}$ of an arbitrary request loss due to displacement is equal to

$$P^{push-loss} = \frac{\sum_{m=2}^{M} \lambda_m^{push-out}}{\lambda}.$$

The probability $P_m^{push-loss}$ of an arbitrary request of type $m$, $m = \overline{2, M}$, loss due to displacement is

$$P_m^{push-loss} = \frac{\lambda_m^{push-out}}{\lambda_m}.$$

The probability $P_{1,m}^{push-loss}$ of an arbitrary request of type $m$, $m = \overline{2, M}$, loss due to displacement by type 1 request is equal to

$$P_{1,m}^{push-loss} = \frac{\mathbf{p}_N (D_1 \otimes E^{(N)}_{1,m})\mathbf{e}}{\lambda_m}.$$

The probability $P_{l,m}^{push-loss}$ that an arriving type-$l$, $l = \overline{2, M-1}$, customer pushes out an arbitrary request of type $m$, $m = \overline{l+1, M}$, is equal to

$$P_{l,m}^{push-loss} = \frac{\sum_{n=N_1}^{N} \mathbf{p}_n (D_l \otimes E^{(n)}_{l,m})\mathbf{e}}{\lambda_m}.$$

The loss probability $P_{loss}$ of an arbitrary request is

$$P^{loss} = P^{push-loss} + P^{arrival-loss} = 1 - \frac{\mu^{out}}{\lambda}.$$

## 5 Numerical example

Let us assume that there are three types of requests ($M = 3$). A size of a request is measured in Megabits (Mb). The size of a type $m$ request has the exponential distribution with the rate $\alpha_m$, $m = \overline{1, 3}$. We set $\alpha_1 = 0.025$. Thus, the average size of a type 1 customer is 40 Mb. The nominal bitrate $\hat{\beta}_1$ of type 1 request is 20 Mb per second. Correspondingly, the service rate of type 1 customer in the case of absence of the deficit of bandwidth is $\beta_1 = 0.5$. For requests of type 2 and type 3, $\alpha_2 = \frac{1}{75}$, $\hat{\beta}_2 = 15$ Mbps, $\beta_2 = 0.2$, and $\alpha_3 = \frac{1}{100}$, $\hat{\beta}_3 = 10$ Mbps, $\beta_3 = 0.1$.

We assume that the arrival flow of requests is the *MMAP* defined by matrices

$$D_0 = \begin{pmatrix} -30 & 0 \\ 0 & -3 \end{pmatrix},\ D_1 = \begin{pmatrix} 14.88 & 0.72 \\ 0.018 & 0.606 \end{pmatrix},$$
$$D_2 = \begin{pmatrix} 6.36 & 0.72 \\ 0.036 & 1.74 \end{pmatrix},\ D_3 = \begin{pmatrix} 7.2 & 0.12 \\ 0 & 0.6 \end{pmatrix}.$$

The average total arrival intensity of customers is $\lambda = 3.90335$, the average arrival intensities of type $m$ requests are $\lambda_1 = 1.12506$, $\lambda_2 = 1.95346$, $\lambda_3 = 0.824833$. The coefficient of variation of inter-arrival times is 1.52387, the coefficients of variation of type $m$ requests inter-arrival times are 2.32208, 1.13619, and 1.50726 correspondingly. The coefficient of correlation of two consecutive inter-arrival times is 0.159857, the coefficients of correlation of two consecutive inter-arrival times of type $m$ requests are 0.236901, 0.0466208, and 0.128633, correspondingly.

We fix that the maximum number of requests that can obtain service at the same time as $N = 50$.

In this numerical example, we intend to investigate the impact of the bandwidth of server $B$ and the parameter $N_1$, which defines the acceptance of lower priority requests, on the main performance measures of the system. For this purpose, we vary the bandwidth $B$ in the range [50, 300] with the step 50, and the parameter $N_1$ over interval [1,50] with step 1. The computations were implemented on PC with Intel Core i7-8700 CPU and 16 GB RAM, Wolfram Mathematica 12.1. The run time is about 80 minutes for 300 different pairs $(B, N_1)$ or 16 seconds per one pair on average.

Figure 2 shows the dependence of the average total number $N^{requests}$ of requests and the mean number $N_m^{requests}$, $m = \overline{1, 3}$, of type $m$ requests in the system on the parameters $N_1$ and $B$.

As it is seen from Fig. 2, in the considered case the average total number $N^{requests}$ of requests decreases with the increase of the bandwidth of the server $B$ and increases with the increase of the parameter $N_1$. Under the fixed $N_1$ the decrease of $N^{requests}$ with the increase in bandwidth $B$ stems from the fact that with growth of $B$ the service rates increase and, therefore, the requests faster depart from the system. Under the fixed $B$, the increase of $N^{requests}$ with the increase in $N_1$ occurs due to the fact that increasing of $N_1$ leads to more tolerant acceptance policy. More requests are admitted to the system what potentially can lead to the decrease of the service rates due to the lack of server's bandwidth. The number $N_1^{requests}$ of type 1 requests in the system behaves the same way as the total number $N^{requests}$ of requests in the system. The mean numbers $N_2^{requests}$ and $N_3^{requests}$ of type 2 and type 3 requests also increase with the increase in $N_1$, but behave not monotonically with the growth in bandwidth $B$.

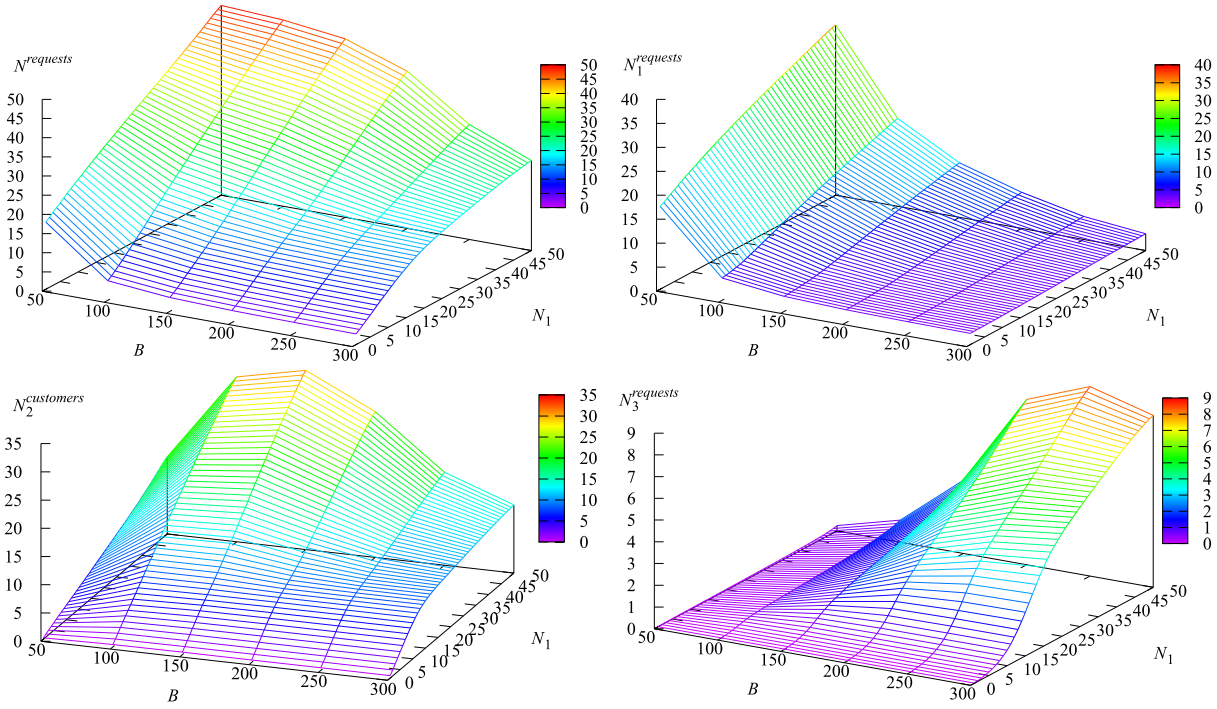


**Fig. 2** Dependence of the mean numbers of requests $N^{requests}$ and $N_m^{requests}$, $m = \overline{1,3}$, on the parameters $N_1$ and $B$

This behavior can be explained as follows. Firstly, when $B$ is small, many type 2 and type 3 requests are rejected at the arrival moment. With the growth in $B$, service rates increase and more requests are accepted to the system what leads to the growth in $N_2^{requests}$ and $N_3^{requests}$. However, with the further growth in $B$ the server becomes less overcrowded what obviously leads to the decrease in the mean number of type 2 and type 3 requests in the system.

Figure 3 illustrates the influence of the parameters $N_1$ and $B$ on the loss probabilities $P^{arrival-loss}$ and $P_m^{arrival-loss}$, $m = \overline{1,3}$.

As it is seen from Fig. 3, with the growth of the bandwidth $B$ the loss probability of any type request decreases because the larger bandwidth implies the bigger average service rates and requests faster leave the server freeing up place for arriving requests. The increase in $N_1$ implies the decrease in the loss probability of $P^{arrival-loss}$, $P_2^{arrival-loss}$, and $P_3^{arrival-loss}$, and decrease in the loss probability of $P_1^{arrival-loss}$. The decrease in the loss probability of $P_1^{arrival-loss}$ despite the preemptive priority over type 2 and type 3 requests can be explained as follows. When $N_1$ increases, evidently more such requests are accepted to the system. The server becomes more loaded and due to sharing the speed of service of type 1 requests decreases, and the situation when an arriving type 1 request meets $N$ type 1 requests obtaining service occurs more often.

The dependence of the probability $P^{no-sharing}$ that all requests at an arbitrary moment obtain required service rate on the parameters $N_1$ and $B$ is presented in Fig. 4. This figure confirms that the probability $P^{no-sharing}$ is large when $B$ is large and $N_1$ is small. Correspondingly, this probability is small when $B$ is small and $N_1$ is large.

These observations, as well as some of dependencies given by Figs. 5, 6, 7 are obvious. However, the behavior of some curves, e.g., figures for $P_3^{push-loss}$ on Fig. 5 and $P_{2,3}^{push-loss}$ on Fig. 6 is quite involved due to complexity of the model. Usefulness of the presented figure consists of giving the exact value of the important performance measures for any fixed values of $B$ and $N_1$. In particular, this allows to solve various optimization problems.

Let us now introduce the cost criterion defined as

$$E(B,N_1) = \sum_{m=1}^{M} (\mathcal{A}_m \mu_m^{out} - \lambda_m \mathcal{B}_m P_m^{arrival-loss}) - \sum_{m=2}^{M} \lambda_m \mathcal{C}_m P_m^{push-loss} - \mathcal{D}B$$

and consider the problem of maximization of this criterion via the proper choice of the parameters $B$ and $N_1$.

Here $\mathcal{A}_m$ is the profit earned by service of one type $m$ request;

$\mathcal{B}_m$ is the charge for loss upon arrival of one type $m$ request;

$\mathcal{C}_m$ is the charge for loss of one type $m$ request due to pushing out;

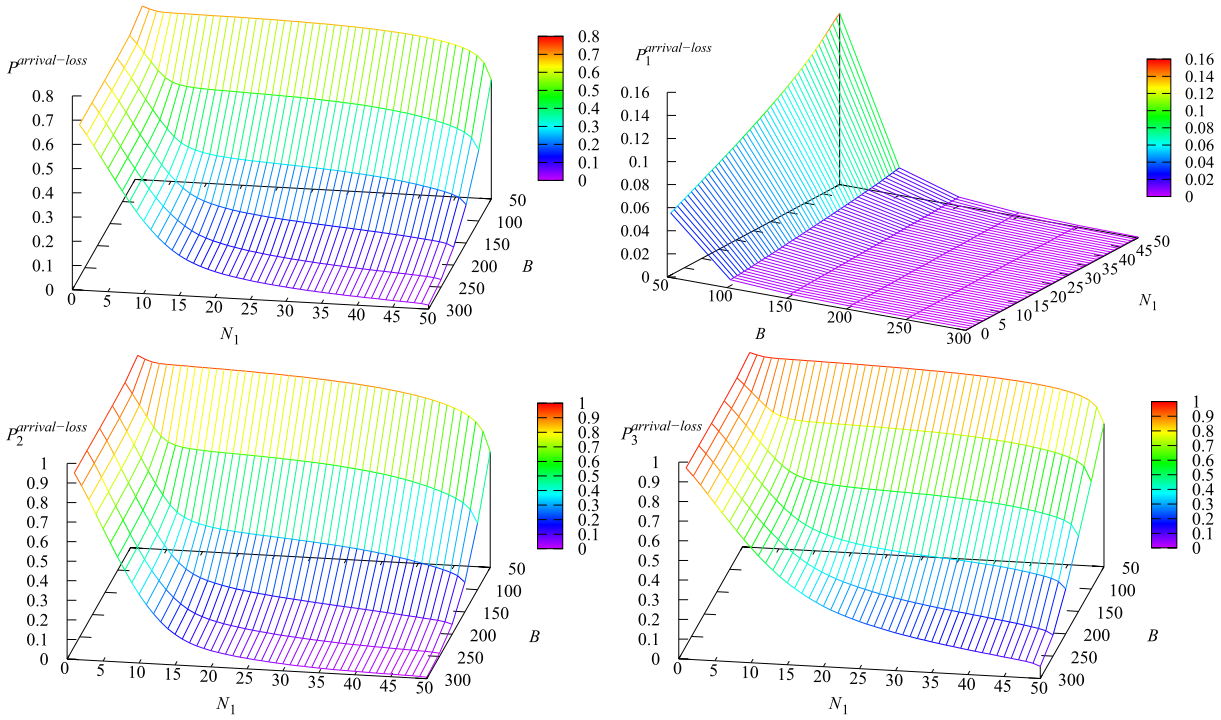


**Fig. 3** Dependence of the probabilities $P^{arrival-loss}$ and $P_m^{arrival-loss}$, $m = \overline{1,3}$, on the parameters $N_1$ and $B$

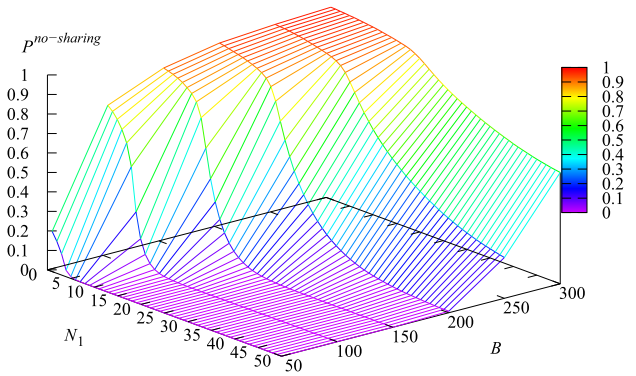


**Fig. 4** Dependence of the probability $P^{no-sharing}$ on the parameters $N_1$ and $B$

$\mathcal{D}$ is the payment per unit time for using one unit of bandwidth.

Let the introduced costs be defined by $\mathcal{A}_1 = 10, \mathcal{A}_2 = 5, \mathcal{A}_3 = 3, \mathcal{B}_1 = 4, \mathcal{B}_2 = 2, \mathcal{B}_3 = 1, \mathcal{C}_2 = 20, \mathcal{C}_3 = 5, \mathcal{D} = 0.05$.

The shape of the function $E(B, N_1)$ is presented on Fig. 8. The optimal value of the cost criterion $E(B, N_1)$ is equal to 7.82733, the optimal values of the bandwidth $B$ and the threshold $N_1$ are equal to 200 and 27, correspondingly.

## 6 Conclusion

In this paper, we introduced and analyzed a novel discipline of simultaneous service of multiple requests. This discipline looks to be realistic for application in real world systems. It assumes restriction on the bandwidth of the server and the number of requests that can receive service at the same time. When the number of requests presenting in the system is relatively small, each of them receives a permanent share of the bandwidth and their service processes are mutually independent, like service in the standard multi-server queueing system. However, when the sum of the bandwidths of the requests admitted to the system exceeds the bandwidth of the server, service to requests is provided at the proportionally reduced rates. Requests are heterogeneous with respect to requirements to the service rates and have different priorities. One of the types of requests has a preemptive priority over the requests of all other types and no restriction in admission until the number of requests presenting in the system reaches the maximum admissible value. The rest of types of requests have more strict restriction in admission and preemptive priorities over each other.

Analysis of the model is performed under realistic suggestion about correlation and possible high variability of inter-arrival times. This is achieved via the assumption that the arrivals occur according to the *MMAP* process which is essentially more general arrival process than the

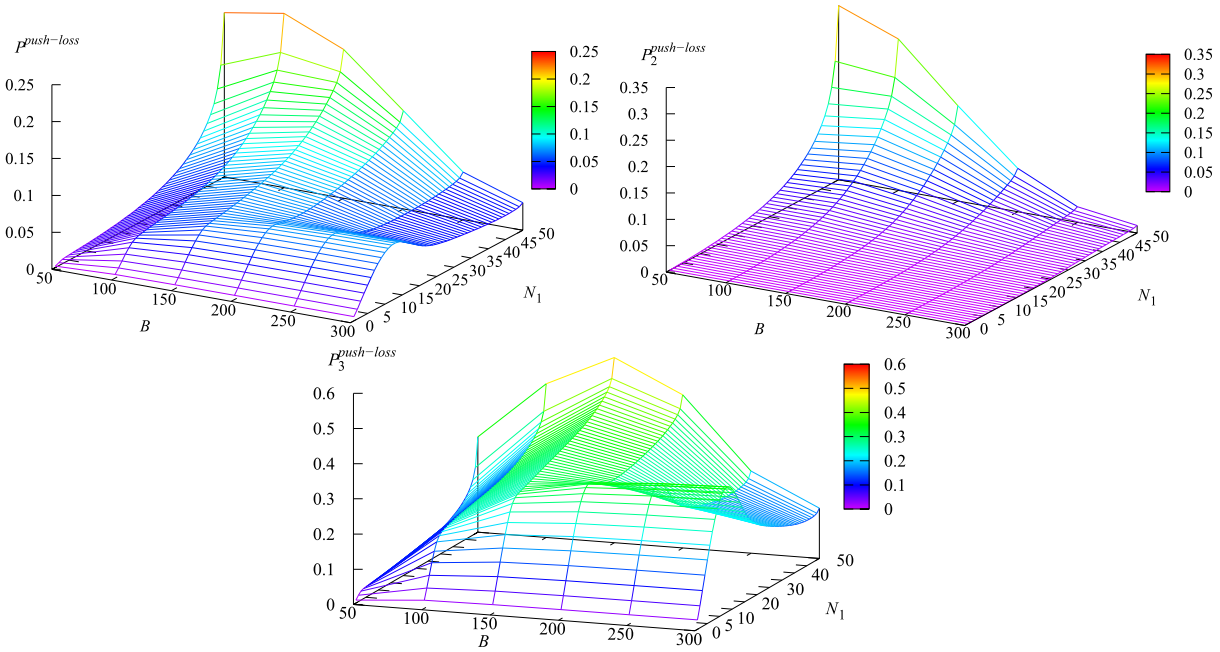


**Fig. 5** Dependence of $P^{push-loss}$ and $P_m^{push-loss}$, $m = 2, 3$ on the parameters $N_1$ and $B$

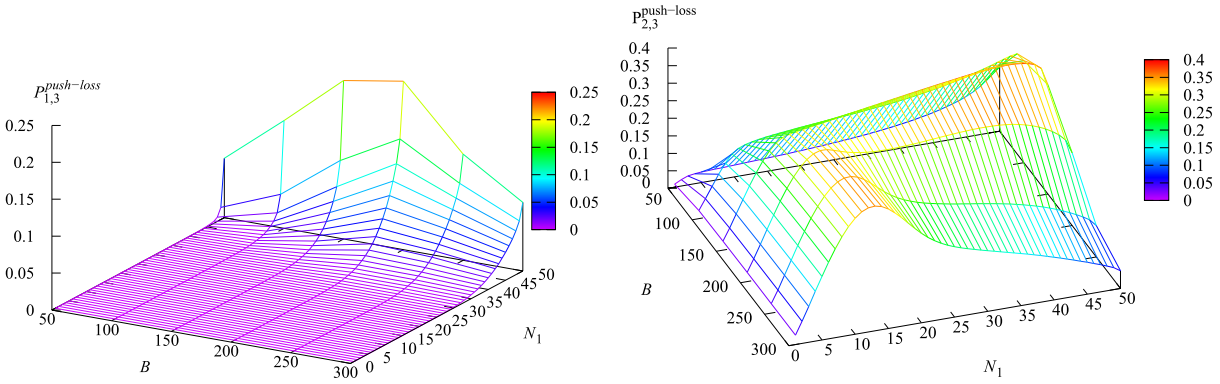


**Fig. 6** Dependence of $P_{1,3}^{push-loss}$ and $P_{2,3}^{push-loss}$ on the parameters $N_1$ and $B$

superposition of the stationary Poisson processes. Feasibility of the proposed method of analysis is illustrated by the numerical example. In particular, the results of solution of the problem of computation of the optimal values of the bandwidth of the server and the number of requests that can receive service simultaneously are presented. Due to application of the technique going back to works by D. Lucantoni and W. Ramaswami, it is possible to implement computations not only for relatively small number of requests receiving service at the same time.

The considered model suggests loss of requests arriving when the number of requests under service has the maximum value. The presented analysis is planned to be extended to the scenarios when storing of such requests in an infinite or finite buffer or repeated attempts to enter the service are possible. In these scenarios, operation of the system can be described by the Markov chain $\tilde{\zeta}_t$ of the form $\tilde{\zeta}_t = \{i_t, \zeta_t\}$ where $i_t$ is the number of requests in the buffer of orbit and $\zeta_t$ is the Markov chain analysed in this paper. If the states of the chain $\tilde{\zeta}_t$ will be enumerated in the direct lexicographic order and the levels of the chain will be defined by the fixed values of the component $i_t$, then the blocks $A_{n,n'}$ of the generator of the Markov chain $\zeta_t$ analysed in this paper will be properly used as the sub-blocks of the blocks of the generator of the Markov chain $\tilde{\zeta}_t$

The case of assigning not equal shares to competing flows of requests, see, e.g., (Chen et al. 2022), can be considered as well. The problem of application of the obtained

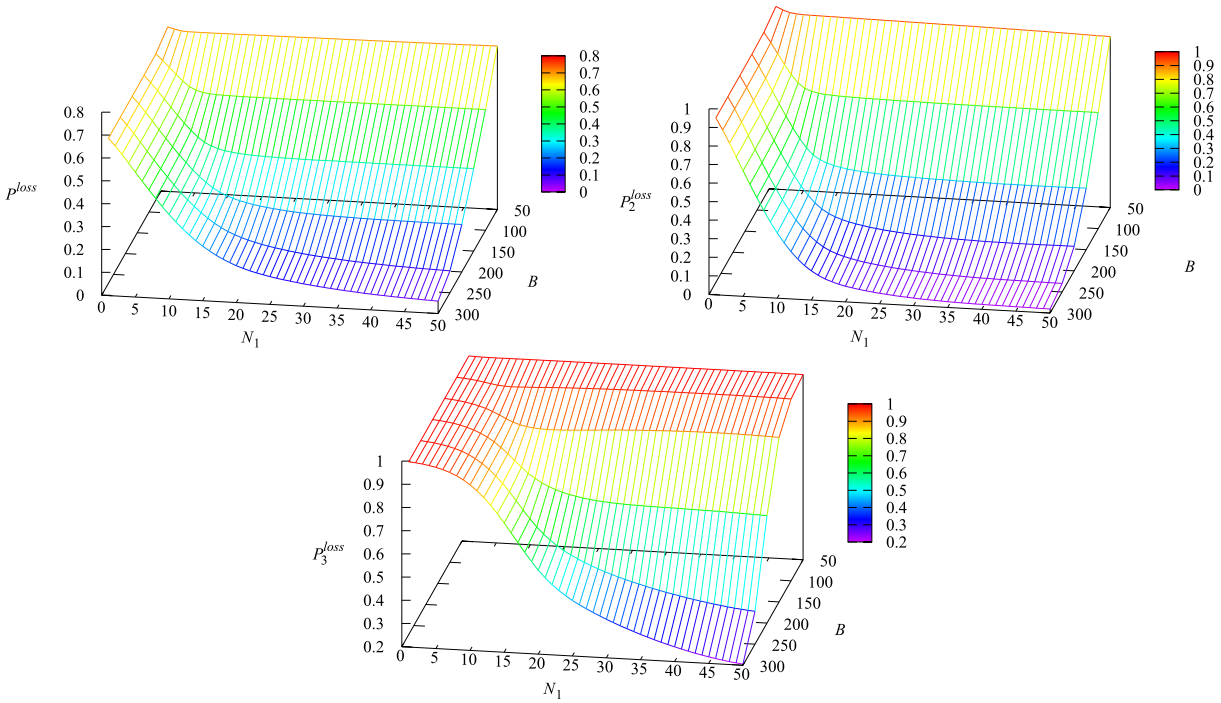


**Fig. 7** Dependence of $P^{loss}$ and $P_m^{loss}$, $m = 2, 3$ on the parameters $N_1$ and $B$

results to analysis of supply systems, see, e.g., (Falco et al. 2017), (Gaeta and Rarità 2013) can be considered.

**Author contributions** The authors declare that all of them have contributed to the realization of the results.

**Funding** Open access funding provided by Università degli Studi di Salerno within the CRUI-CARE Agreement.

**Availability of data and materials** Data sharing not applicable to this article as no datasets were generated or analysed during the current study.

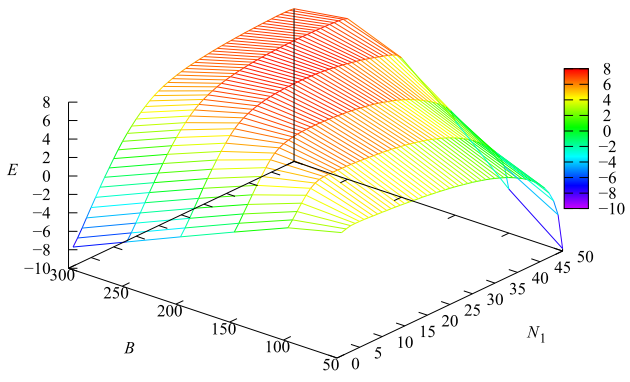


**Fig. 8** Dependence of the cost function $E$ on the parameters $N_1$ and $B$

## Declarations

**Conflict of interest** The authors have no conflict of interest to declare that are relevant to the content of this article.



## References

Alencar M, Yashina M, Tatashev A (2021) Loss queueing systems with limited processor sharing and applications to communication network. In: Paper presented at the 2021 international conference on engineering management of communication and technology, EMCTECH, Vienna, Austria, 20–22 October 2021

Altman E, Avrachenkov K, Ayesta U (2006) A survey on discriminatory processor sharing. Queueing Syst 53(1):53–63

Baumann H, Sandmann W (2010) Numerical solution of level dependent quasi-birth-and-death processes. Procedia Comput Sci 1:1561–1569

Bocharov PP, D'Apice C, Manzo R, Pechinkin AV (2007) Analysis of the multi-server markov queuing system with unlimited buffer and negative customers. Autom Remote Control 68(1):85–94

Brugno A, Dudin AN, Manzo R (2017) Retrial queue with discipline of adaptive permanent pooling. Appl Math Model 50:1–16

Brugno A, Dudin AN, Manzo R (2018) Analysis of a strategy of adaptive group admission of customers to single server retrial system. J Ambient Intell Humaniz Comput 9(1):123–135

Chen G, Xia L, Jiang Z, Peng X, Xu H (2022) A two-class map/ph/1 weighted fair queueing system and its application to telecommunications. J Ambient Intell Humaniz Comput. https://doi.org/10.1007/s12652-022-03857-2

D'Arienzo MP, Dudin AN, Dudin SA, Manzo R (2020) Analysis of a retrial queue with group service of impatient customers. J Ambient Intell Humaniz Comput 11(6):2591–9

Dudin A, Kim CS, Dudina O, Dudin S (2016) Queueing system with generalized phase type service time distribution. Ann Oper Res 239:401–428

Dudin S, Dudin A, Dudina O, Samouylov K (2017) Analysis of a retrial queue with limited processor sharing operating in the random environment. Lect Notes Comput Sci 10372:38–49

Dudin AN, Dudina OS, Dudin SA, Kostyukova OI (2021) Optimization of road design via the use of a queueing model with transit and local users and processor sharing disciplines. Optimization. https://doi.org/10.1080/02331934.2021.2009827

El-Toukhy AT, Arslan H (2019) Enhancing the performance of low priority sus using reserved channels in CRN. IEEE Wirel Commun Lett 9(4):513–517

de Falco M, Mastrandrea N, Rarità L (2017) A queueing networks-based model for supply systems. International conference on optimization and decision science. Springer, New York, pp 375–383

Gaeta M, Rarità L (2013) A stochastic approach for supply systems. In Proceeding of 25th European modeling and simulation symposium, pp 401–409

Ghosh A, Banik A (2017) An algorithmic analysis of the BMAP/MSP/1 generalized processor-sharing queue. Comput Oper Res 79:1–11

Goel S, Kulshrestha R (2022) Queueing based spectrum management in cognitive radio networks with retrial and heterogeneous service classes. J Ambient Intell Humaniz Comput 13:2429–2437

Graham A (2018) Kronecker products and matrix calculus with applications. Courier Dover Publications, New York

Gupta V, Zhang J (2022) Approximations and optimal control for state-dependent limited processor sharing queues. Stochastic Systems 12(2):205–225

He QM (1996) Queues with marked customers. Adv Appl Probab 28:567–587

Kim CS, Dudin SA, Taramin OS, Baek J (2013) Queueing system MAP/PH/N/N+R with impatient heterogeneous customers as a model of call center. Appl Math Model 37:958–976

Kim C, Dudin SA, Dudina OS, Dudin AN (2019) Mathematical models for the operation of a cell with bandwidth sharing and moving users. IEEE Trans Wirel Commun 19(2):744–755

Kim C, Dudin A, Dudin S, Dudina O (2021) Mathematical model of operation of a cell of a mobile communication network with adaptive modulation schemes and handover of mobile users. IEEE Access 9:106933–106946

Lee S, Dudin A, Dudina O, Kim C (2022) Analysis of a priority queueing system with the enhanced fairness of servers scheduling. J Ambient Intell Humaniz Comput. https://doi.org/10.1007/s12652-022-03903-z

Masuyama H, Takine T (2003) Sojourn time distribution in a MAP/M/1 processor-sharing queue. Oper Res Lett 31:406–412

Nair J, Wierman A, Zwart B (2010) Tail-robust scheduling via limited processor sharing. Perform Eval 67(11):978–995

Neuts M (1981) Matrix-geometric solutions in stochastic models. The Johns Hopkins University Press, Baltimore

Ramaswami V, Lucantoni DN (1985) Algorithms for the multi-server queue with phase-type service. Commun Stat Stochastic Models 1:393–417

Samouylov KE, Sopin ES, Gudkova I (2016) Sojourn time analysis for processor sharing loss queuing system with service interruptions and MAP arrivals. Commun Comput Inf Sci 678:406–417

Sun B, Lee MH, Dudin SA, Dudin AN (2014a) Analysis of multiserver queueing system with opportunistic occupation and reservation of servers. Math Prob Eng ID 178108:1–13

Sun B, Lee MH, Dudin SA, Dudin AN (2014b) $MAP + MAP/M_2/N/\infty$ queueing system with absolute priority and reservation of servers. Math Prob Eng ID 813150:1–15

Telek M, Van Houdt B (2018) Response time distribution of a class of limited processor sharing queues. ACM SIGMETRICS Perform Eval Rev 45(3):143–155

Yashkov SF (1987) Processor-sharing queues: some progress in analysis. Queueing Syst 2(1):1–17

Yashkov SF, Yashkova AS (2007) Processor sharing. Autom Remote Control 68:1662–1731