AIMS Mathematics, 9(5): 12144–12169.
DOI: 10.3934/math.2024593
Received: 01 February 2024
Revised: 21 February 2024
Accepted: 05 March 2024
Published: 27 March 2024

http://www.aimspress.com/journal/Math

Research article

Queueing system with batch arrival of heterogeneous orders, flexible limited processor sharing and dynamical change of priorities

Alexander Dudin[1], **Sergey Dudin**[1], **Rosanna Manzo** [2,*]**and Luigi Rarità**[3]

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

[2] Department of Political and Communication Sciences, University of Salerno, Via Giovanni Paolo II, 132, Fisciano 84084, Salerno, Italy

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

* **Correspondence:** Email: rmanzo@unisa.it; Tel: +39-3381838978.

Abstract: A queueing system with the discipline of flexible limited sharing of the server is considered. This discipline assumes the admission, for a simultaneous service, of only a finite number of orders, as well as the use of a reduced service rate when the bandwidth required by the admitted orders is less than the total bandwidth of the server. The orders arrive following a batch-marked Markov arrival process, which is a generalization of the well-known *MAP* (Markov arrival process) to the cases of heterogeneous orders and batch arrivals. The orders of different types have different preemptive priorities. The possibility of an increase or a decrease in order priority during the service is suggested to be an effective mechanism to prevent long processing orders from being pushed out of service by just-arrived higher-priority orders. Under a fixed priority scheme and a mechanism of dynamic change of the priorities, the stationary analysis of this queueing system is implemented by considering a suitable multidimensional continuous-time Markov chain with a generator that has an upper Hessenberg structure. The possibility of the optimal restriction on the number of simultaneously serviced orders is numerically demonstrated.

Keywords: multi-server priority queueing model; batch-marked Markov arrival process; flexible limited processor sharing; multidimensional Markov chains
Mathematics Subject Classification: 60K25, 60K30, 60M20

1. Introduction

Queueing theory is dedicated to effectively solving practically important problems related to sharing and scheduling the use of restricted resources among different potential users (orders) in various areas of human practice. A high diversity of potential applications in telecommunication, manufacturing, transportation, logistics, finance, health care, and other systems has led to the formulation of a great variety of different queueing models and their analysis in the literature since the early 1900s. An effective use of a restricted resource affords the possibility of managing the simultaneous service of many orders. Two basic ways to provide a simultaneous service to a lot of orders are as follows: 1) the division of the resource into a finite number of parts (in the simplest case, equal parts) and the use of each part (server, device, agent, etc.) for the service of a definite order; and 2) the offer of a service to the admitted orders by using the whole resource.

The analysis of the first choice dominates in the existing literature. The subject of this paper is a queueing model that focuses on the second possibility. Different variants of this possible realization could be considered. One popular approach suggests the time division of the resource, namely, the time is divided into short slots in which each admitted order receives full access to the resource itself. When the slot ends, the access is granted to another user while the customer, who has just ended his/her own access, waits for the assignment of another slot. This mechanism is called time-sharing and has been successfully analyzed in terms of so-called polling systems; see, e.g., [1–3].

If the length of a slot in the time-sharing discipline is infinitely small, the processor-sharing (PS) discipline is obtained. The simplest variant of the latter discipline is the egalitarian PS discipline, according to which all orders are always admitted to the system upon arrival and are processed at an equal rate that is inversely proportional to the number of serviced orders. Another type is called discriminatory PS, where all orders are always admitted to the system but the orders can have a different speed of processing; see, e.g., [4]. Extensive reviews of research on queues with the PS discipline can be found in [5, 6].

The essential advantage of the PS discipline is the permanent full use of the resource when there are orders in the systems. This advantage can turn into a disadvantage, as was mentioned in [7]. Besides the technological problems in the implementation of the PS discipline in practical systems, the main two disadvantages are as follows:

(i) The resource can be potentially infinitely divisible, but the orders can have minimum requirements for the rate of service. Thus, dividing the resource into too many parts is an infeasible option;

(ii) the flow of orders can have maximum requirements for the rate of service. A user of the system, which generates the orders at the rate of, say, λ, does not need the service rate μ such that $\mu > \lambda$. Thus, dividing the resource into too small of a number of parts is also an infeasible option.

A reasonable way to mitigate disadvantage (ii) is described in [8] for the so-called mixed service discipline, which considers the differentiation of arriving orders into two classes. Class-1 orders need a permanent service rate. Class-2 orders admit service with a variable rate in the PS mode. The total resource is correspondingly separated into two parts. One part is dedicated to the service of a limiting number of class-1 orders. Another part, as well as the temporarily unused share of the first part, is used for the simultaneous service of an unlimited number of class-2 orders. The analysis, implemented in [8], shows the high efficiency of this mixed service discipline.

Concerning the mitigation of disadvantage (i), the limited *PS* (*LPS*) is recommended for use instead of the classical *PS* discipline. The *LPS* discipline also suggests that all of the orders admitted for service are serviced simultaneously. However, some kind of admission control is implemented in such a way that the number of simultaneously serviced orders is limited by some fixed in-advance integer number. Such a number is called a multiprogramming level; see, e.g., [9], or, for a concurrency limit; see, e.g., [10]. Different scenarios of an order behavior, i.e., when the order arrives while the number of orders receiving service is equal to the limit, have been considered. They include the variants when the arriving order is lost, joins a finite or infinite buffer, or makes repeated attempts to enter the service at a later time. The literature on queues with the *LPS* discipline is already quite extensive; such as that in [7, 11–22].

Most of the mentioned papers deal with queues with the *LPS* discipline and are devoted to systems with homogeneous orders. However, orders in many real-world systems may be heterogeneous and have different requirements for the desired service rate, as well as different values (importance) for the system. In the case of heterogeneous orders, modifications of the *LPS* discipline, similar to the discriminatory *PS* mentioned above or generalized *PS*, such as that in [23], can be applied.

In this paper, we consider a queueing model with heterogeneous orders with various priorities. The priority is provided through a combination of the mechanism that is typical for queues with the discriminatory *PS* discipline (giving different speeds of service for orders of different classes), as well as the one for classical multi-server queues with order acceptance control. We assume admission control via the provision of a preemptive priority. This means that each type of order has a certain priority. If an order arrives when the number of processed orders is less than the maximum value, it is admitted for service regardless of its priority. If it arrives when the number of processed orders has reached its maximum value, then the arriving order pushes out of service a serviced order of the lowest priority among the orders of a lower priority than the arriving order. The pushed-out order is assumed to be lost. If all orders have a priority that is not lower than the arriving order, then this order is lost. Indeed, we assume that a batch arrival of orders is possible. Therefore, many low-priority orders can be simultaneously lost. A more detailed description of such a scenario is given in the next section.

The main contributions of this paper are as follows:

- A new mechanism of priority provisioning in a queueing system with heterogeneous traffic, a fixed bandwidth of the server, and an *LPS* service discipline are proposed and analyzed. This mechanism assumes a combination of providing different nominal service rates to orders of different types and controlling order access and a possibility of interruption of service of low-priority orders because of the preemptive priority discipline.
- The possibility of a temporal provision of service to all orders at a proportionally reduced rate when the sum of the nominal service rates of all servicing orders exceeds the capacity of the server is explored.
- Order impatience, i.e., the possible interruption of unfinished service in the case of a long service duration, is taken into account.
- A dynamic change (increase or decrease) in the order priority during their service is considered, which essentially allows for improvement of the quality of system operations. The increase in priority during the order service represents an effective mechanism to prevent long processing orders from being pushed out of service by just-arrived higher-priority orders. Such pushing out, indeed, means a waste of the resources of the system that have thus far been consumed by

the service of the long-serviced order. In contrast a decrease in priority during the service may be reasonable due to the reduction of the value of a long servicing order, e.g., because of its obsolescence or spoilage. A possibility of a priority change was previously considered only for systems without the use of PS or LPS disciplines; see, e.g., [24–35] and the references therein. In the mentioned papers, the priorities could be changed only during the waiting time of the orders. In our model, customer waiting is not possible and the priorities can be dynamically changed during the service time of the orders. Management based on the priorities during the servicing of orders under the LPS discipline, which reduces the amount of wasted system resources, is more important than the mechanism of picking up of orders for service.

- Unlike most of the mentioned papers, the analysis of the model has been implemented under an essentially more realistic assumption about the arrival flow than the stationary Poisson process. We suggest the batch-marked Markov arrival process ($BMMAP$), which is a generalization of the well-known Markov arrival process (MAP). MAP allows to consider not only the mean arrival rate, as in the case of the stationary Poisson arrival process, but all of the moments of distribution of the inter-arrival times, as well as the possible correlation of these times. Note that the positive correlation deteriorates the performances of a queueing system relative to that of the system with the stationary Poisson arrival process. The $BMMAP$ is a more general process than the MAP for two reasons: arriving orders can have different types, and the arrival of orders can occur, not one by one, but in batches of a random size. If only one-by-one arrivals are possible, the $BMMAP$ turns into the $MMAP$; see [36].

Potential fields of application of priority queueing models include the following ones; see, e.g., [24]. In information transmission networks, signaling information is more important than the routine sent by the users, as time-sensitive information should be transferred more urgently than the elastic, time-insensitive information. The transmission of driving safety information in transportation systems is more urgent than the transmission of infotainment-related information. The handover user, who has arrived at the cell of a mobile network, has to be treated with another priority than the new user by establishing a connection within a given cell. In the emergency healthcare system, the patients can be sorted and treated by injury severity. In food delivery services, the most rapidly deteriorating (perishable) items have to be delivered first. A suitable choice of priorities can essentially increase the revenue of the service provider. However, as the customers wait or processing occurs, the situation can change and the assigned priorities must be dynamically varied. As a convincing example, the treatment of patients in emergency departments is usually mentioned. Upon patient's arrival, the doctors implement a so-called triage, i.e., classification of the patients into several categories according to the severity of the condition and threat to life. The patients are then treated according to their categories. After a certain amount of time, the health condition of any patient can improve or deteriorate and the intensity of his/her treatment can be decreased or must be increased. The decision about the capacity of the emergency department and the required equipment has to take into account the rate of patient arrivals and available statistics about the usual results of the initial triage and possible changes of the health condition of a patient after the triage. The model considered in this paper can be helpful for the optimal design of an emergency department and a policy of patient admissions or redirection to another hospital.

The paper is organized as follows. In Section 2, a detailed description of the constructed queueing model is given. The dynamics of the considered queueing model are presented in Section 3 through

the use of a continuous-time multidimensional Markov Chain (MC). Explicit expressions for the infinitesimal generator of the chain are also obtained. In Section 4, formulas for the key performance characteristics of the system are reported. Some results of the numerical experiments are demonstrated in Section 5. The paper ends with conclusions in Section 6.

2. Mathematical model

We consider a queueing system with a flexible *LPS*, whose scheme of operation is presented in Figure 1. According to the flexible *LPS* scheme, the *LPS* discipline is not used permanently, but only in the case of a shortage of the available server capacity (bandwidth).

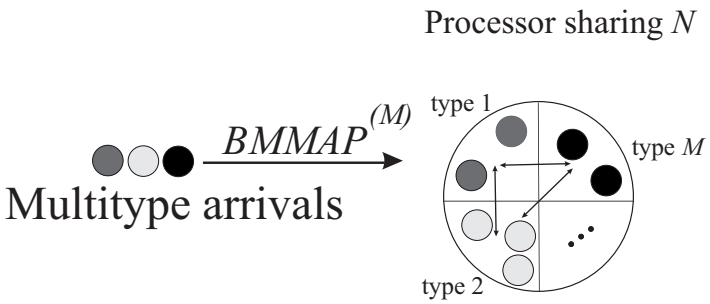

Figure 1. Scheme of operation of the system.

Orders that enter the system are divided into M types. The input flow of orders is described by the *BMMAP*. Order arrivals in the *BMMAP* are defined by the irreducible continuous-time MC ν_t, $t \geq 0$, that has the finite state space $\{1, \ldots, W\}$. The sojourn time of the MC ν_t, $t \geq 0$, in the state ν is exponentially distributed by using the positive parameter λ_ν. After this time expires, the chain jumps to the state ν', where $\nu' \in \{1, 2, \ldots, W\}$, $\nu' \neq \nu$, without the generation of orders with probability $p_0(\nu, \nu')$, and it jumps to the state ν', where $\nu' \in \{1, 2, \ldots, W\}$, with probability $p_m^{(k)}(\nu, \nu')$; then, a batch consisting of k orders of type-m is generated. Here, we assume that the maximum batch size of type-m orders is limited by the parameter K_m, $K_m \geq 1$. Indicate by K the maximum batch size among all types of orders, i.e., $K = \max\{K_m,\ m = \overline{1, M}\}$. Hereinafter, the notation $m = \overline{1, M}$ means that the integer parameter m admits the values in the set $\{1, 2, \ldots, M\}$.

The parameters that define the *BMMAP* can be stored in the square matrices D_0 and $D_r^{(m)}$, $m = \overline{1, M}$, $k = \overline{1, K_m}$, of size W, defined by their entries:

$$(D_0)_{\nu,\nu} = -\lambda_\nu,\ (D_0)_{\nu,\nu'} = \lambda_\nu p_0(\nu, \nu'),\ \nu, \nu' = \overline{1, W},\ \nu \neq \nu,$$

$$(D_m^{(k)})_{\nu,\nu'} = \lambda_\nu p_m^{(k)}(\nu, \nu'),\ \nu, \nu' = \overline{1, W},\ k = \overline{1, K_m}, m = \overline{1, M}.$$

The matrix

$$D(1) = D_0 + \sum_{m=1}^{M} \sum_{k=1}^{K_m} D_m^{(k)}$$

is a generator of the MC ν_t, $t \geq 0$.

Denote by $\boldsymbol{\theta}$ the stationary probability vector of the states of the MC ν_t, $t \geq 0$. This vector is found to be the unique solution to the system given by

$$\boldsymbol{\theta} D(1) = \mathbf{0}, \boldsymbol{\theta}\mathbf{e} = 1.$$

Hereinafter, $\mathbf{0}$ is a zero row vector and $\mathbf{e}$ is the column vector consisting of ones.

The average intensity λ_m of type-m order arrivals is given by

$$\lambda_m = \boldsymbol{\theta} \sum_{k=1}^{K_m} kD_r^{(m)}\mathbf{e}, m = \overline{1, M}.$$

The average intensity λ of order arrivals is given by $\lambda = \sum_{m=1}^{M} \lambda_m$.

For more information about the *BMMAP*, see, e.g., [37].

As we deal with the *LPS* service discipline, the rate of service of an arbitrary order can be changed many times during its service. Therefore, there is a need to explain how the service time of an order of type-m, $m = \overline{1, M}$, is defined. Denote by B the capacity (throughput, bandwidth) of the server. If an order is interpreted as some portion of the information that should be processed by a server or a channel of a telecommunication system, the bandwidth can be measured, e.g., in megabits per second (Mbps). We assume that the size of an order of type-m is random, having an exponential distribution with the parameter α_m; hence, the mean size of an order of type-m is α_m^{-1} megabits. Also, we suppose that a type-m order needs a service rate (or bitrate measured in Mbps) that is equal to $\hat{\beta}_m$. Thus, the mean required service time of the type-m order is equal to $(\hat{\beta}_m\alpha_m)^{-1}$. Correspondingly, the required (nominal) service rate of a type-m order is $\beta_m = \hat{\beta}_m\alpha_m$.

We assume that the maximum number of orders that can be serviced in the system simultaneously is equal to N, which has to depend on either the bandwidth B or the required service rates for different types of orders. An optimal (with respect to some criterion) value of N can be chosen based on the results of the analysis presented below.

In this analysis, first, we fix an arbitrary finite value of N, $N \geq 1$. Under such an assumption, a situation can occur whereby the sum of the required service rates of the orders that receive service is higher than the bandwidth B, especially if the fixed value of N is large. We suppose that such a situation is not extraordinary and, if it occurs, all servicing orders are serviced, not at the required rate, but at a reduced rate, defined in the following way.

Let $s^{(m)}$ be the current number of type-m orders receiving service, $0 \leq s^{(m)} \leq N, m = \overline{1, M}$, $\sum_{m=1}^{M} s^{(m)} = N$ and $\sum_{m=1}^{M} s^{(m)}\hat{\beta}_k > B$. Then, the reduced service rate of a type-m order is defined by

$$\frac{B}{\sum_{k=1}^{M} s^{(k)}\hat{\beta}_k}\beta_m, \ m = \overline{1, M}.$$

The arriving orders have distinct priorities. Precisely, the type-1 orders have the highest priority and the type-M orders have the lowest priority. If an order of any type arrives when the number of orders in the system is less than N, it is admitted for service. If an order of type-m arrives when the number of orders in the system equals N, it is admitted for service only if some orders receiving service are of type-m', where $m' > m$. In this case, an order of the lowest type (among the ones receiving service) terminates the service and is lost. If an order of type-m arrives when the number of orders in the system is equal to N and all orders receiving service are of type $\{1, \ldots, m\}$, then the arriving order is lost.

As the various orders can arrive in batches, we need to distinguish the discipline of individual orders from this batch acceptance. Each order arriving in a batch of size, say, k, is considered to be the rth order in the batch, with the probability $\frac{1}{k}$, $r = \overline{1, k}$, and the orders from one batch are accepted by the system according to this enumeration. The so-called partial admission discipline is considered, and this

suggests that, if the sum of the number of orders in an arriving batch, say, k, and the number i of orders receiving service is higher than N, i.e., $i + k > N$, then only $N - i$ orders are immediately admitted for service. The rest of the $i + k - N$ orders can be individually (according to their numeration in a batch) admitted for service if the appropriate number of low-priority orders are presented in the system. The scenario of low-priority orders that are removed from service is described above. The orders of a batch that were not included in the current servicing round are lost. For example, if a batch consisting of five orders arrives when service is provided to $N - 2$ orders, one of which has a lower priority than the arriving batch, then three orders from the arriving batch are accepted for service, while two orders from this batch and an order of the lower priority are lost.

As the service is received, each order of type-m, $m = \overline{1, M}$, can change (increase or decrease) its priority. We assume that, after an exponentially distributed time with the parameter φ_m, $\varphi_m \geq 0$, any type-m, $m = \overline{1, M}$, order becomes a type-l order with probability $p_{m,l}$, $l = \overline{1, M}$, $l \neq m$, independently of the other orders. Here, $\sum\limits_{l=1,\, l\neq m}^{M} p_{m,l} = 1$, $m = \overline{1, M}$. For instance, if we consider a model whereby orders can only increase their priority, we should put $p_{m,l} = 0$ for $l \geq m$. The intensity φ_m can be equal to zero if type-m orders, $m = \overline{1, M}$, cannot change their priority. Note that an increase or a decrease in the priority of a single order can lead to the occurrence or disappearance of a shortage of bandwidth. This, in turn, can imply the necessity of a service rate reduction for all orders, or a return to the nominal service rate.

The orders admitted for service can be impatient and depart from the system without completing the service, independently of other orders. If, during an exponentially distributed time with parameter γ_m, $m = \overline{1, M}$, a type-m order does not succeed in finishing the service, this order is removed from the system permanently (i.e., it is lost). If, during the service, the order changes its priority and becomes a type-m' order, then the patience time restarts and has an exponential distribution with parameter $\gamma_{m'}$, $m' = \overline{1, M}$. We indicate the set of parameters γ_m, $m = \overline{1, M}$, with $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_M)$.

In what follows, the described queueing system is discussed.

3. Description of the dynamics of the considered queueing system by using a MC

Let n_t, $n_t = \overline{0, N}$, be the number of serviced orders, $s_t^{(m)}$ be the number of serviced orders of type-m, $m = \overline{1, M}$, and $0 \leq s_t^{(m)} \leq n_t$, $\sum\limits_{m=1}^{M} s_t^{(m)} = n_t$, at time t. Since service with a reduced rate is possible, the actual service rate of an order is equal to its nominal service rate β_m if the total used bandwidth at time t, defined as $\sum\limits_{k=1}^{M} s_t^{(k)} \hat{\beta}_k$, is less than the bandwidth B of the server. Otherwise, the actual service intensity is equal to the proportionally reduced nominal service intensity $\frac{B}{\sum\limits_{k=1}^{M} s_t^{(k)} \hat{\beta}_k} \beta_m$, $m = \overline{1, M}$.

It is easy to prove that the process

$$\zeta_t = \{n_t, \nu_t, s_t^{(1)}, \ldots, s_t^{(M)}\},\ n_t = \overline{0, N}, \nu_t = \overline{1, W},\ s_t^{(m)} = \overline{0, n_t},\ m = \overline{1, M},\ \sum_{m=1}^{M} s_t^{(m)} = n_t,\ t \geq 0,$$

comprehensively describes the behavior of the queueing system under consideration and is a continuous-time MC.

Since this MC is irreducible and has a finite state space, the stationary probabilities of the state of the system

$$\boldsymbol{\pi}(n, \nu, s^{(1)}, \ldots, s^{(M)}) = \lim_{t\to\infty} P\{n_t = n, \nu_t = \nu, s_t^{(1)} = s^{(1)}, \ldots, s_t^{(M)} = s^{(M)}\}$$

exist for any values of the system parameters.

We enumerate the states of the process ζ_t in the direct lexicographic order of the components (n_t, ν_t) and the reverse lexicographic order of the components of the vector process $\mathbf{s}_t^{n_t}$, where

$$\mathbf{s}_t^{n_t} = \{(s_t^{(1)}, \ldots, s_t^{(M)}),\ 0 \le s_t^{(m)} \le n_t,\ m = \overline{1, M},\ \sum_{m=1}^{M} s_t^{(m)} = n_t\}.$$

Finally, we indicate the set of process states by ζ_t, having the value n of the component n_t as level n of the MC ζ_t, $n = \overline{0, N}$. Following the introduced enumeration, we construct the row vectors $\boldsymbol{\pi}_n$, $n = \overline{0, N}$, for the stationary probabilities of the states belonging to the level n.

It is well known that the vectors of stationary probabilities $\boldsymbol{\pi}_n$, $n = \overline{0, N}$, satisfy the conditions of the following system of linear algebraic equations (equilibrium equations or Chapman-Kolmogorov system):

$$(\boldsymbol{\pi}_0, \boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_N)G = \mathbf{0},\ \ (\boldsymbol{\pi}_0, \boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_N)\mathbf{e} = 1 \tag{1}$$

where G is the generator of the MC ζ_t.

To calculate these vectors, we first need to write down the explicit form of the generator G. The most difficult task here is to analyze the transition intensities of the components of the multidimensional process $\mathbf{s}_t^{n_t}$, which determines the number of orders of each type in the system. To this end, we implement four steps.

In Step 1, we describe the process of servicing a single order in the system when the system is not overloaded and the order receives the nominal required service intensity regardless of its type, and when there are also no removals of the order from the service. In Step 2, we analyze the process of $\mathbf{s}_t^n$ that describes the simultaneous service of n orders, provided that these orders receive the nominal (not reduced) service rate and there is no removal of the orders that are out of the service. In Step 3, we consider the chance of removing an order from the service because of the arriving orders with a higher priority. Finally, in Step 4, we account for the possibility of order service at a reduced rate.

Step 1. To describe the process of servicing a single order when the system is not overloaded and the order receives the nominal required service intensity regardless of its type, and when there are no removals of orders from the service, we analyze various scenarios of an order service. It can be verified that the service time has a distribution that is an extension of the so-called generalized phase-type distribution introduced in [38].

As the underlying process of service for an arbitrary order, consider the continuous-time MC s_t, $t \ge 0$, with the space of transient states $\{1, \ldots, M\}$. The initial state of the MC s_t at the beginning of service is determined by by the probability vector $\mathbf{b}_m$, that is,

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

if the incoming order has the type-m. The transitions of the underlying process within the space of the

transient states are defined by the sub-generator given by

$$S = \begin{pmatrix} -\beta_1 - \varphi_1 - \gamma_1 & p_{1,2}\varphi_1 & p_{1,3}\varphi_1 & \cdots & p_{1,M-1}\varphi_1 & p_{1,M}\varphi_1 \\ p_{2,1}\varphi_2 & -\beta_2 - \varphi_2 - \gamma_2 & p_{2,3}\varphi_2 & \cdots & p_{2,M-1}\varphi_2 & p_{2,M}\varphi_2 \\ p_{3,1}\varphi_3 & p_{3,2}\varphi_3 & -\beta_3 - \varphi_3 - \gamma_3 & \cdots & p_{3,M-1}\varphi_3 & p_{3,M}\varphi_3 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ p_{M,1}\varphi_M & p_{M,2}\varphi_M & p_{M,3}\varphi_M & \cdots & p_{M,M-1}\varphi_M & -\beta_M - \varphi_M - \gamma_M \end{pmatrix}.$$

Step 2. We analyze the process of $\mathbf{s}_t^n$ that describes the simultaneous service of n orders, provided that these orders receive the nominal (not reduced) service rate and there is no removal of orders from the service.

Assume the following:

- Matrix $Y_n = Y_n(\Phi)$, $n = \overline{1, N}$, contains the transition intensities of this process when some order changes the priority. Here, the matrix Φ defines the intensities associated with increasing and decreasing the priority. It is given by the following formula:

$$\Phi = S + \text{diag}\{\beta_m + \varphi_m + \gamma_m,\ m = \overline{1, M}\}$$

  where diag{...} is a diagonal matrix whose diagonal elements are specified by the elements or a vector given in the brackets;
- Matrix $P_n(\mathbf{b}_m)$ defines the transition probabilities of the process $\mathbf{s}_t^n$ at the moment that the servicing of a new type-m order, $n = \overline{0, N-1}$, $m = \overline{1, M}$, begins;
- Matrix $L_n(\boldsymbol{\beta})$, $n = \overline{1, N}$, $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_M)$, sets the transition intensities of the process $\mathbf{s}_t^n$ at the end of the service of one of the orders;
- Matrix $\Gamma_n(\boldsymbol{\gamma})$ describes the intensities of transition of the process $\mathbf{s}_t^n$ when one of n orders departs from the system due to impatience. A detailed description of these matrices and the used algorithms for their calculation are presented, for example, in [39–41].

Note that Y_n, $n = \overline{1, N}$, denotes square matrices of size T_n, the matrices denoted by $P_n(\mathbf{b}_m)$, $n = \overline{0, N-1}$, have size $T_n \times T_{n+1}$, and the matrices $\Gamma_n(\boldsymbol{\gamma})$ and $L_n(\boldsymbol{\beta})$, $n = \overline{1, N}$, have size $T_n \times T_{n-1}$, where $T_n = \binom{n+M-1}{M-1}$, $n = \overline{1, N}$.

Step 3. Due to the preemptive priority of some types of orders, it is necessary to account for the event of removing an order from the service because of arriving orders with a higher priority. Because the orders arrive in batches, we assumed above that they are randomly numbered. We suggest that system be designed such that the orders attemp to push out from service the ones with lower priorities, and, if it is necessary and possible, one by one. This means that the order with the number 1 makes the first attempt, the order with the number 2 makes the second attempt, and the order with the last number makes the last attempt. Note that these attempts are not implemented sequentially, but instantaneously, at the same moment. To account for transitions of the process $\mathbf{s}_t^N$ when a single order causes the system to remove a low-priority order, we introduce the following matrices.

- E_m, $m = \overline{1, M-1}$, denotes square matrices of size T_N whose elements determine the transition probabilities of the process $\mathbf{s}_t^N$ when an order of type-m, $m = \overline{1, M-1}$, arrives at the system; there is no free space for it, and the incoming order tries to remove an order with lower priority

from the service. Each row and column of E_m corresponds to some states $\{s_1, s_2, \ldots, s_M\}$ and $\{s'_1, s'_2, \ldots, s'_M\}$ of the process $\mathbf{s}_t^N$, $t \geq 0$, numbered in the reverse lexicographic order. The arrival of a high-priority order that pushes out of service a low-priority order, has to imply, with probability 1, the transition of the vector process $\mathbf{s}_t^N$ into another fixed state. Therefore, all elements in each row of the matrix E_m are equal to zero, except for one element that is equal to 1. In the row of matrix E_m corresponding to the state $\{s_1, s_2, \ldots, s_M\}$, element 1 is located in the column corresponding to the same state $\{s_1, s_2, \ldots, s_M\}$ only if $s_l = 0$ for all l, $M \geq l > m$. In this case, the received order of type-m is lost since orders with a lower priority are absent in the service. If $s_l > 0$ for some l, $M \geq l > m$, and j^* is the maximum of such values l, then element 1 is located in the column corresponding to the state given by

$$\{s_1, \ldots, s_{m-1}, s_m + 1, s_{m+1}, \ldots, s_{j^*-1}, s_{j^*} - 1, 0, \ldots, 0\}.$$

In this case, an order of type-j^* has the lowest priority among the ones serviced in the system, and an incoming order of type-m pushes one order of type-j^* out of service. This order is removed from the system (is lost).

Step 4. To consider the important fact that all orders receive the reduced service rate when the total required bandwidth is higher than the bandwidth B of the server, we also use the following additional notations:

- $\mathbf{a}_n = L_n(\hat{\boldsymbol{\beta}})\mathbf{e}$, $n = \overline{1, N}$, where
$$\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_M).$$
The components of the vector $\mathbf{a}_n$ define the sum of all desired service rates of the serviced orders under the corresponding states of the process $\mathbf{s}_t^n$;
- $\mathbf{d}_n$, $n = \overline{1, N}$, denotes column vectors of size T_n, and elements denoted by $(\mathbf{d}_n)_l$, $l = \overline{1, T_n}$, are defined as
$$(\mathbf{d}_n)_l = \begin{cases} 1, & \text{if } (\mathbf{a}_n)_l \leq B, \\ \frac{B}{(\mathbf{a}_n)_l}, & \text{otherwise.} \end{cases}$$
The components of the vector $\mathbf{d}_n$ define the reduction factors for desired service rates under the corresponding states of the process $\mathbf{s}_t^n$;
- $\Delta_n = -\text{diag}\{Y_n\mathbf{e} + \Gamma_n(\boldsymbol{\gamma})\mathbf{e} + \text{diag}\{\mathbf{d}_n\}L_n(\boldsymbol{\beta})\mathbf{e}\}$, $n = \overline{1, N}$.

After the implementation of Steps 1–4, we are ready to write down the explicit form of the generator G of the MC ζ_t. Since orders can enter the system in batches and leave the system only one at a time, it is obvious that this generator is a block upper Hessenberg matrix of the following form:

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

The diagonal elements of the blocks $G_{n,n}$, $n = \overline{0, N}$, are negative and their modules determine the intensity of the exit from the corresponding states of the MC. The non-diagonal elements of these

blocks are non-negative and determine the transition intensities of *MC* ζ_t within the level n. The block elements $G_{n,n-1}$, $n = \overline{1,N}$, and $G_{n,n+l}$, $n = \overline{0,N-1}$, $l = \overline{1,N-n}$, are non-negative and determine the intensities of the *MC* ζ_t transitions from the level n to the levels $n-1$ and $n+l$, $l = \overline{1,N-n}$, respectively.

Theorem 1. *The non-zero blocks of the generator G have the following form:*

$$G_{0,0} = D_0,$$

$$G_{n,n} = D_0 \oplus (Y_n + \Delta_n),\ n = \overline{1,N-1}, \tag{2}$$

$$G_{N,N} = D_0 \oplus (Y_N + \Delta_N) + \sum_{m=1}^{M} \sum_{k=1}^{K_m} D_m^{(k)} \otimes (E_m)^k, \tag{3}$$

$$G_{n,n+k} = \sum_{m=1}^{M} \delta_{k \le K_m} D_m^{(k)} \otimes \prod_{l=n}^{n+k-1} P_l(\mathbf{b}_m),\ n = \overline{0, \min\{N-1, K\}},\ k = \overline{1, \min\{N-n, K\}}, \tag{4}$$

$$G_{n,N} = \sum_{m=1}^{M} \Bigg(\sum_{k=N-n+1}^{K_m} D_m^{(k)} \otimes \prod_{l=n}^{N-1} P_l(\mathbf{b}_m)(E_m)^{k-(N-n)} + D_m^{(N-n)} \otimes \prod_{l=n}^{N-1} P_l(\mathbf{b}_m) \Bigg), \tag{5}$$

$$G_{n,n-1} = I_W \otimes \Big(\mathrm{diag}\{\mathbf{d}_n\} L_n(\boldsymbol{\beta}) + \Gamma_n(\boldsymbol{\gamma}) \Big),\ n = \overline{1,N}, \tag{6}$$

where I is the identity matrix of the corresponding dimension, δ_A is the indicator of an event A, and $\otimes$ and $\oplus$ denote Kronecker products and the sums of matrices, respectively; see [42].

Proof. The proof is carried out by analyzing possible transitions of the *MC* ζ_t on an interval of infinitesimal length.

As already remarked, the diagonal elements of the blocks $G_{n,n}$, $n = \overline{0,N}$, are negative and their modules determine the intensity of the exit from the corresponding states of the *MC*. The non-diagonal elements of the blocks $G_{n,n}$, $n = \overline{0,N}$, are non-negative and determine the transition intensities among the respective states of the *MC*.

When $n = 0$, no service is provided and all possible transitions of the *MC* ζ_t (and exits from its states) can occur only due to the transitions (or exits) of the underlying *MC* of arrivals ν_t. Thus, $G_{0,0} = D_0$.

Consider the case that $n = \overline{1,N-1}$: along with the transitions of the underlying *MC* of arrivals ν_t, transitions caused by changes of the priority of the orders, described by the matrix Y_n, are possible. The exits from the states, along with the exits of the underlying *MC* of arrivals ν_t, can occur due to the exits caused by the change of a priority, changes of the vector for the underlying process of service (at the required or reduced rate), or order removal because of impatience. The rates of these exits are given by the diagonal entries of the diagonal matrix Δ_n. Therefore, we obtain that $G_{n,n} = D_0 \otimes I_{T_n} + I_W \otimes (Y_n + \Delta_n)$. Taking into account the definition of the Kronecker sum of matrices (see [42]), we obtain formula (2).

Focus on the case that $n = N$: in addition to the already explained summand $D_0 \oplus (Y_N + \Delta_N)$, we obtain the summand $\sum_{m=1}^{M} \sum_{k=1}^{K_m} D_m^{(k)} \otimes (E_m)^k$. It corresponds to the possible transitions of the *MC* ζ_t caused by the arrival of batches of the size k of type-m orders, $k = \overline{1,K_m}$, $m = \overline{1,M}$. As the number of orders in the system is already equal to the limiting value N, the arrived orders try to push out of service

low-priority orders, if any exist. The matrix $(E_m)^k$ describes transitions of the vector for the underlying process $\mathbf{s}_t^N$ of a service that occurred as the result of k sequential trials. As a result, we get formula (3).

The matrix $G_{n,n+k}$, $k = \overline{1, \min\{N-n, K\}}$, $n = \overline{0, \min\{N-1, K\}}$, describes the rates of transition that lead to an increase of the number of orders in the system from n to $n+k$. Such an increase can occur if the batch of k orders of any type-m such that $k \leq K_m$ arrives. The service of each of these orders begins, and the initial states of the underlying processes of service are installed. It is easy to show that probabilities of the transitions of the vector for the underlying process of service at this moment are equal to the entries of the matrix $\prod_{l=n}^{n+k-1} P_l(\mathbf{b}_m)$. From these considerations, we have formula (4).

Formula (5) is explained similarly. Only here, the number of orders in the system after transition becomes equal to N. If an arriving batch of orders of any type consists of exactly $N-n$ orders, we have the same situation as in the explanation of formula (4). But, if the size k of the type-m order batch is more than $N-n$, the service begins only on $N-n$ servers, while the other $k-(N-n)$ orders of this batch try to push out of service low-priority orders, if any exist. This explains the multiplier $(E_m)^{k-(N-n)}$ in the first summand in formula (5).

Formula (6) describes the transition rates of the *MC* ζ_t when the number of orders in the system decreases from n to $n-1$. Such transitions can be induced by an order loss due to impatience or the service completion of one order. Here, the use of the vector $\mathbf{d}_n$ indicates the decrease in service intensity in the case of a bandwidth shortage.

The theorem is proven. □

The number of equations in the system (1) with a block upper Hessenberg structure of the generator G can be large, and the solution for this system on a computer requires highly complex computations. To solve this system, it is recommended to use an algorithm that effectively accounts for the sparse structure of the generator. In particular, we recommend the numerically stable algorithm developed in [24].

4. System performance characteristics

After the computation of the vectors denoted by $\boldsymbol{\pi}_n$, $n = \overline{0, N}$, we can determine various useful performance indicators of the queueing system under consideration.

The average number of orders in the system is given by

$$N_{orders} = \sum_{n=1}^{N} n\boldsymbol{\pi}_n \mathbf{e}.$$

The average number $N_{orders}^{(m)}$ of type-m orders in the system, where $m = \overline{1, M}$, is defined as

$$N_{orders}^{(m)} = \sum_{n=1}^{N} \boldsymbol{\pi}_n (I_W \otimes L_n(\mathbf{b}_m))\mathbf{e}.$$

Here, the matrix $L_n(\mathbf{b}_m)$ is computed by using the same algorithm as that for the matrix $L_n(\boldsymbol{\beta})$, $n = \overline{1, N}$, replacing the vector $\boldsymbol{\beta}$ of service rates by the stochastic vector $\mathbf{b}_m$.

The intensity of the output flow of successfully serviced orders is equal to

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

The intensity of the output flow $\mu_{out}^{(m)}$ of successfully serviced orders of type-m, $m = \overline{1, M}$, equals

$$\mu_{out}^{(m)} = \sum_{n=1}^{N} \boldsymbol{\pi}_n (I_W \otimes \mathrm{diag}\{\mathbf{d}_n\} L_n(\tilde{\boldsymbol{\beta}}_m))\mathbf{e}.$$

Here, the vector $\tilde{\boldsymbol{\beta}}_m$ of size M has all zero components, except for the m-th one that is equal to β_m, $m = \overline{1, M}$.

The rate μ_{imp}, i.e., the rate at which orders leave the system due to impatience, is equal to

$$\mu_{imp} = \sum_{n=1}^{N} \boldsymbol{\pi}_n (I_W \otimes \Gamma_n(\boldsymbol{\gamma}))\mathbf{e}.$$

The departure rate $\mu_{imp}^{(m)}$ of type-m orders m, $m = \overline{1, M}$, due to impatience is computed as follows:

$$\mu_{imp}^{(m)} = \sum_{n=1}^{N} \boldsymbol{\pi}_n (I_W \otimes \Gamma_n(\tilde{\boldsymbol{\gamma}}_m))\mathbf{e}.$$

Here, the vector $\tilde{\boldsymbol{\gamma}}_m$ of size M has all zero components, except for the m-th one that is equal to γ_m, $m = \overline{1, M}$.

The probability of losing an arbitrary order is given by

$$P_{loss} = 1 - \lambda^{-1}\mu^{out}.$$

The probability of losing an arbitrary order due to impatience is given by

$$P_{imp} = \lambda^{-1}\mu^{imp}.$$

The probability of losing an arbitrary order because of the arrival of a batch of orders for which there is not enough space is equal to

$$P_{arrival-loss} = \lambda^{-1}\Bigg(\sum_{n=0}^{N}\sum_{m=1}^{M}\sum_{k=N-n+1}^{K_m}(k-(N-n))\boldsymbol{\pi}_n(D_m^{(k)} \otimes I_{T_n})\mathbf{e}\Bigg) = P_{loss} - P_{imp}.$$

It is evident that $k - (N - n)$ orders are lost if a batch of k orders of any type arrives when the number of processed orders equals n, where $k > N - n$. Using the formula of the total probability, the rate of flow of the lost orders is $\Bigg(\sum\limits_{n=0}^{N}\sum\limits_{m=1}^{M}\sum\limits_{k=N-n+1}^{K_m}(k-(N-n))\boldsymbol{\pi}_n(D_m^{(k)} \otimes I_{T_n})\mathbf{e}\Bigg)$. Dividing the rate of flow of the lost orders by the rate of arriving orders, we get the formula for $P_{arrival-loss}$.

Assume that $R_{m,l}$ is the diagonal matrix of size T_N, whose diagonal entries determine the number of times the corresponding entry of the identity matrix of size T_N is shifted when this matrix is sequentially multiplied l times by the matrix E_m, $l = \overline{1, K_m}$, $m = \overline{1, M}$.

The probability of losing an arbitrary order because of the pushing out of service by higher-priority orders is calculated as follows:

$$P_{push-loss} = \lambda^{-1}\Big(\sum_{n=0}^{N-1}\sum_{m=1}^{M-1}\sum_{k=N-n+1}^{K_m}\boldsymbol{\pi}_n(D_m^{(k)}\otimes\prod_{l=n}^{N-1}P_l(\mathbf{b}_m))(I_W\otimes R_{m,k-(N-n)})\mathbf{e} + \sum_{m=1}^{M-1}\sum_{k=1}^{K_m}\boldsymbol{\pi}_N(D_m^{(k)}\otimes I_{T_N})(I_W\otimes R_{m,k})\mathbf{e}\Big).$$

Observe that each shift occurs only when some low-priority order is pushed out of service. Therefore, the expression in the brackets in the given formula for $P_{push-loss}$ represents the rate of orders being pushed out of service, as calculated by using the formula for total probability. Dividing this rate by the order arrival rate λ, we obtain the formula for $P_{push-loss}$.

The probability of losing an arbitrary arriving order upon entry to the system due to the lack of space and the inability to push out a low-priority order from service is calculated as follows:

$$P_{ent-loss} = \lambda^{-1}\Big(\sum_{n=0}^{N-1}\sum_{m=1}^{M}\sum_{k=N-n+1}^{K_m}\boldsymbol{\pi}_n(D_m^{(k)}\otimes\prod_{l=n}^{N-1}P_l(\mathbf{b}_m))(I_W\otimes((k-(N-n))I_{T_N}-R_{m,k-(N-n)})\mathbf{e} + \sum_{m=1}^{M}\sum_{k=1}^{K_m}\boldsymbol{\pi}_N(D_m^{(k)}\otimes I_{T_N})(kI_{T_N}-R_{m,k})\mathbf{e}\Big) = P_{arrival-loss}-P_{push-loss}.$$

Note that the definitions of the loss probabilities $P_{arrival-loss}$ and $P_{ent-loss}$ are similar. The difference between them is as follows. The first one is the probability that, at an arbitrary arrival moment, an order loss happens. The lost order may either be from a just-arrived batch or a pushed-out-of-service order. The second probability considers the possibility of the loss of an arriving order.

The existence of two different formulas for the calculation of the probabilities $P_{arrival-loss}$ and $P_{ent-loss}$ is helpful to control the accuracy of the computation of the probability vectors $\boldsymbol{\pi}_n,\ n=\overline{0,N}$.

The probability of losing an arbitrary order of type-m upon arrival to the system due to the lack of space and the inability to knock out a low-priority order from service is defined as

$$P_{ent-loss}^{(m)} = \lambda_m^{-1}\Big(\sum_{n=0}^{N-1}\sum_{k=N-n+1}^{K_m}\boldsymbol{\pi}_n(D_m^{(k)}\otimes\prod_{l=n}^{N-1}P_l(\mathbf{b}_m))(I_W\otimes((k-(N-n))I_{T_N}-R_{m,k-(N-n)})\mathbf{e} + \sum_{k=1}^{K_m}\boldsymbol{\pi}_N(D_m^{(k)}\otimes I_{T_N})(kI_{T_N}-R_{m,k})\mathbf{e}\Big).$$

The probability that, at an arbitrary moment, a bandwidth shortage occurs in the system is equal to

$$P_{sharing} = \sum_{n=1}^{N}\boldsymbol{\pi}_n(I_W\otimes\mathrm{diag}\{\mathbf{q}_n\})\mathbf{e}$$

where $\mathbf{q}_n\ n=\overline{1,N}$, denotes column vectors of size T_n, and elements denoted by $(\mathbf{q}_n)_l$, $l=\overline{1,T_n}$, are defined as follows:

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

Correspondingly, the probability that, at an arbitrary moment, no bandwidth shortage occurs in the system is equal to $P_{no-sharing} = 1 - P_{sharing}$.

The average intensity $\hat{\lambda}_{to}^{(m)}$ of the type-l, $l = \overline{1, M}$, $m \neq l$, order transformation to the type-m, $m = \overline{1, M}$, orders is computed as follows:

$$\hat{\lambda}_{to}^{(m)} = \sum_{l=1,\, l\neq m}^{M} p_{l,m}\varphi_l N_{orders}^{(l)}.$$

The average intensity $\tilde{\lambda}_{from}^{(m)}$ of the type-m, $m = \overline{1, M}$, order transformation to the other types of orders is given by

$$\tilde{\lambda}_{from}^{(m)} = \sum_{l=1,\, l\neq m}^{M} p_{m,l}\varphi_m N_{orders}^{(m)}.$$

The probability of the loss of an arbitrary type-m order due to impatience, i.e., $P_{imp-loss}^{(m)}$, $m = \overline{1, M}$, is given by

$$P_{imp-loss}^{(m)} = \frac{\mu_{imp}^{(m)}}{\lambda_m + \hat{\lambda}_{to}^{(m)} - \tilde{\lambda}_{from}^{(m)}}.$$

5. Numerical examples

It is intuitively evident that the bandwidth, B, of the server and the number of orders N, that can be admitted for simultaneous service have a deep effect on the performances of the system. The goal of the presented numerical results is to quantitatively highlight the impact of B and N on the main features of the system.

Assume that the number of types of arriving orders, M, is 3. The arriving $BMMAP$ is defined by matrices of size 2:

$$D_0 = \text{diag}\{-0.529945028, -0.5425546\},$$

$$D_1^{(1)} = \begin{pmatrix} 0.00643918 & 0.0070831 \\ 0.00611722 & 0.00643918 \end{pmatrix}, D_1^{(2)} = \begin{pmatrix} 0.0128784 & 0.0099807 \\ 0.0122344 & 0.0135223 \end{pmatrix},$$

$$D_1^{(3)} = \begin{pmatrix} 0.0177077 & 0.0170638 \\ 0.0164199 & 0.0189956 \end{pmatrix}, D_1^{(4)} = \begin{pmatrix} 0.0193175 & 0.0196395 \\ 0.0199615 & 0.0189956 \end{pmatrix},$$

$$D_2^{(1)} = \begin{pmatrix} 0.0325179 & 0.0647138 \\ 0.0972316 & 0.0962657 \end{pmatrix}, D_2^{(2)} = \begin{pmatrix} 0.0482939 & 0.112686 \\ 0.0486158 & 0.112364 \end{pmatrix},$$

$$D_3^{(1)} = \begin{pmatrix} 0.0482939 & 0.0321959 \\ 0.001235 & 0.009765 \end{pmatrix}, D_3^{(2)} = \begin{pmatrix} 0.0643918 & 0.000643918 \\ 0.024563 & 0.0076329 \end{pmatrix},$$

$$D_3^{(3)} = \begin{pmatrix} 0.011532 & 0.004566 \\ 0.0197839 & 0.012412 \end{pmatrix}.$$

The mean arrival rates of orders of different types are as follows:

$$\lambda_1 = 0.0321031,\ \lambda_2 = 0.0466686,\ \lambda_3 = 0.0212283.$$

The mean arrival rates of the batches of order of different types are as follows:

$$\lambda_1^{batch} = 0.0110805,\ \lambda_2^{batch} = 0.0306644,\ \lambda_3^{batch} = 0.0115951.$$

The total mean arrival rate is $\lambda = 0.1$.

The nominal service rates are fixed to be as follows:

$$\beta_1 = 0.05,\ \beta_2 = \frac{1}{35},\ \beta_3 = \frac{1}{80}.$$

The parameters of the exponential distribution of the duration of time before the change of priority are as follows:

$$\varphi_1 = 0,\ \varphi_2 = 0.03,\ \varphi_3 = 0.03.$$

The matrix $P = \|p_{m,l}\|_{m,l=\overline{1,3}}$, which defines the probabilities of the changes in the type of orders, is given by

$$P = \begin{pmatrix} 0 & 0.2 & 0.8 \\ 1 & 0 & 0 \\ 0.5 & 0.5 & 0 \end{pmatrix}.$$

Because $\varphi_1 = 0$, type-1 orders cannot change the priority. Thus, the probabilities in the first row of the matrix P do not have any effect.

The impatience rates were chosen as

$$\gamma_1 = 0.003,\ \gamma_2 = 0.002,\ \gamma_3 = 0.001.$$

Let us vary the bandwidth of the server, B, in the range $[50, 500]$ with step equal to 50 and the limit of simultaneously serviced orders, N, in the range $[1, 30]$ with one step.

Figures 2–5 show the dependencies of the average number of orders N_{orders} and the average number of type-m orders, $m = \overline{1,3}$, $N_{orders}^{(m)}$ on the parameters N and B. We notice that, when the bandwidth B is sufficiently large, say, larger than 200, the effect of the parameter N is not very essential. The shown average numbers, more or less, essentially increase only for small values of N. With the subsequent increase of N, the average numbers of orders become constant. It is clear because, for large B, the service rate is high and the probability that the maximum number of orders, N, receives the service is small for large N.

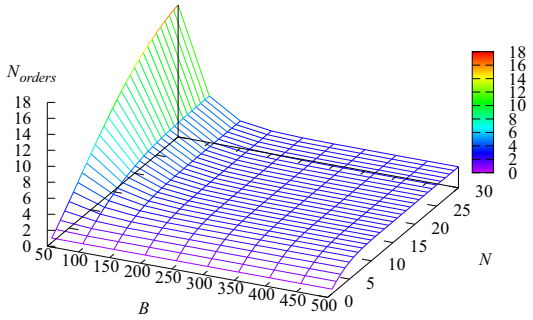

Figure 2. The dependence of the average number of orders N_{orders} in the system on the parameters N and B.

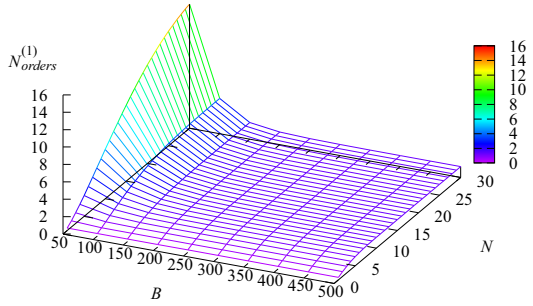

Figure 3. The dependence of the average number of type-1 orders $N^{(1)}_{orders}$ in the system on the parameters N and B.

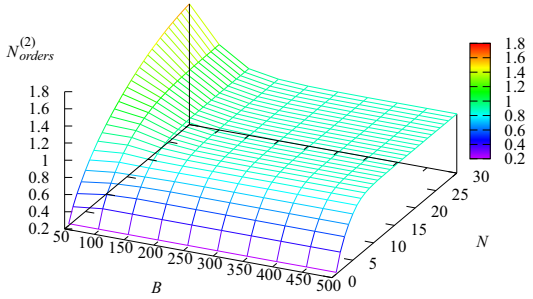

Figure 4. The dependence of the average number of type-2 orders $N^{(2)}_{orders}$ in the system on the parameters N and B.

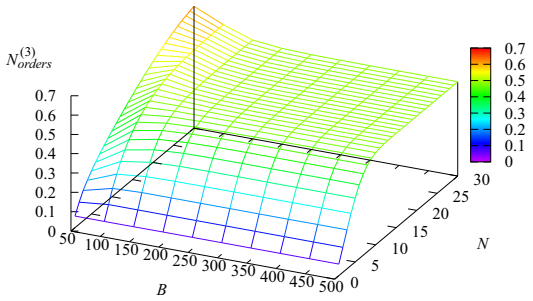

Figure 5. The dependence of the average number of type-3 orders $N^{(3)}_{orders}$ in the system on the parameters N and B.

But, when the bandwidth B is relatively small, the influence of the parameter N drastically grows. An increase of N implies a sharp increase of N_{orders} and $N^{(m)}_{orders}$, where $m = \overline{1,3}$. These figures can be helpful when choosing the minimum value of B and the maximum value of N for which the mean number of service orders does not exceed a predetermined number.

Figures 6–9 illustrate the dependencies of the loss probabilities of the orders entering the system $P_{ent-loss}$ and $P^{(m)}_{ent-loss}, m = \overline{1,3}$, on the parameters N and B. These probabilities are very large when N is small. The dependence on B is weaker, but such probabilities can increase when the value of B is small.

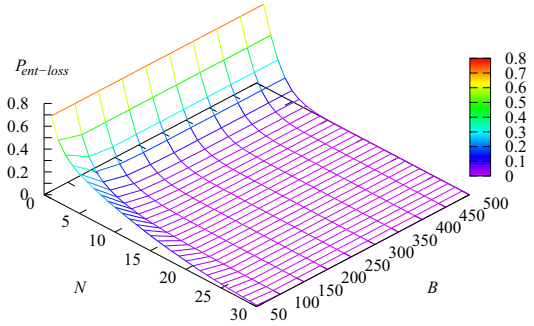

Figure 6. The dependence of the probability $P_{ent-loss}$ on the parameters N and B.

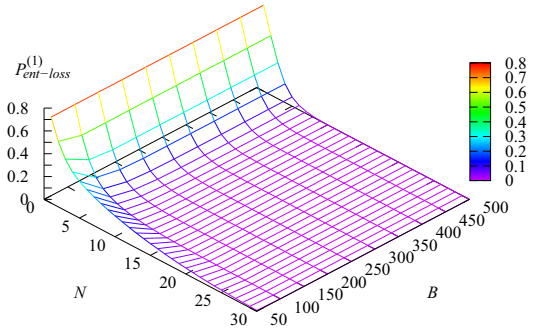

Figure 7. The dependence of the probability $P^{(1)}_{ent-loss}$ on the parameters N and B.

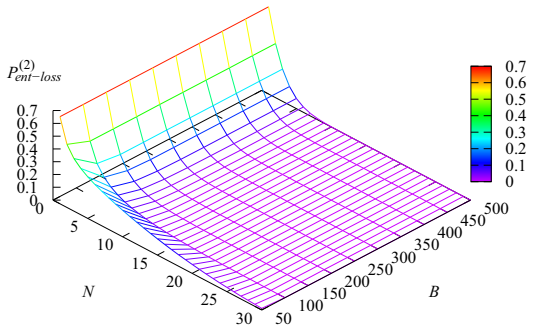

Figure 8. The dependence of the probability $P^{(2)}_{ent-loss}$ on the parameters N and B.

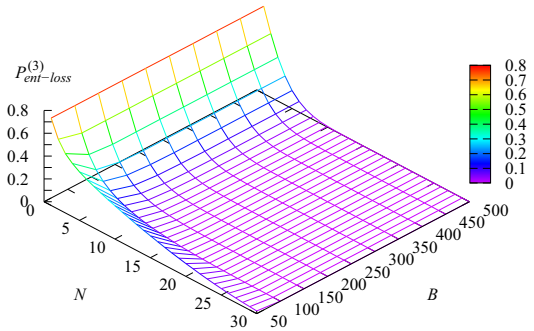

Figure 9. The dependence of the probability $P^{(3)}_{ent-loss}$ on the parameters N and B.

Figure 10 illustrates the dependence of the loss probability $P_{arrival-loss}$ on the parameters N and B. This probability is large when N is small and quickly decreases with the increase of N. For large N, it is not negligible when B is small, and it becomes very small with the increase of B. Notice that the surface in this figure looks similar to the one in Figure 6 for the probability $P_{ent-loss}$. However, these surfaces are different.

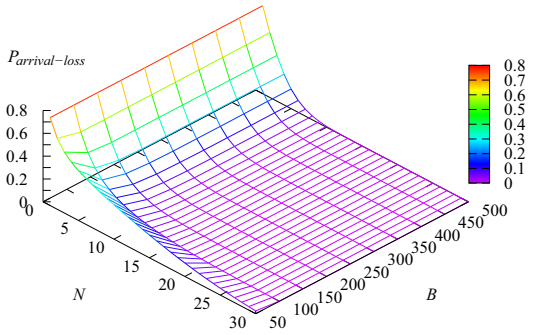

Figure 10. The dependence of the probability $P_{arrival-loss}$ on the parameters N and B.

Figure 11 shows the dependence of the probability of an arbitrary order loss due to impatience $P_{imp-loss}$ on the parameters N and B. As it may be anticipated, this probability sharply increases when the bandwidth B is small but a large number of orders can be admitted to service. Due to the small bandwidth, many orders receive a reduced service rate. Thus, their service time is long, and their loss due to impatience becomes very likely. The data in this figure may help to match the limit N with the bandwidth W to avoid numerous losses of orders as caused by excessively mild restrictions on the admission of arriving orders.

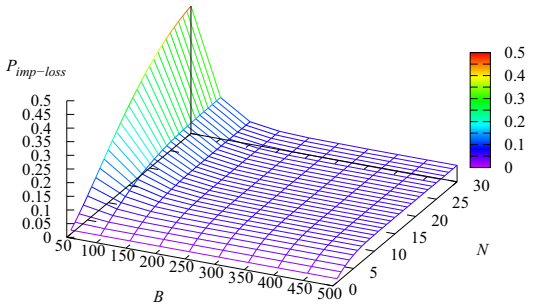

Figure 11. The dependence of the probability $P_{imp-loss}$ on the parameters N and B.

The above figures are quite easily tractable. Less obvious is the surface, which describes the behavior of the probability of removal from service of an admitted low-priority order, $P_{push-loss}$; see Figure 12. The smallest value of this probability, for any B, is achieved for $N = 1$. This is clear because only one order can be accepted for service, and, regarding its removal from service, it is mandatory that no type-1 order is in service and the arriving order is not rejected upon arrival and has higher priority. This event has a relatively low probability. When N increases, the probability of this event increases because (i) fewer orders are rejected upon arrival and more low-priority orders are admitted for service and become the potential targets of being pushed out, and (ii) more high-priority orders can arrive and interrupt the service of low-priority orders. However, with a further increase in N, the probability of rejection of an arbitrary order, including low-priority orders, decreases. Thus, the

number of potential targets increases, which decreases the probability of existing order being pushed out, namely, an arbitrarily considered order. These intuitive considerations can help us to understand the observed effect. But, the concrete value of the number N after which the probability $P_{push-loss}$ starts decreasing can be found only by using the results of the above algorithmic analysis. There is also a natural effect in Figure 12 that indicates that the probability $P_{push-loss}$ is essential when the bandwidth is small.

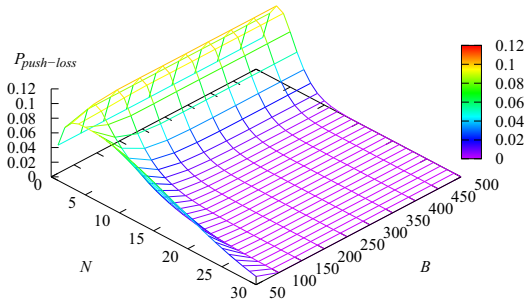

Figure 12. The dependence of the probability $P_{push-loss}$ on the parameters N and B.

Figure 13 illustrates the behavior of the probability P_{loss} of losing an arbitrary order (due to any reason). Because $P_{loss} = P_{arrival-loss} + P_{imp-loss}$, the form of the surface in Figure 13 is predefined by one of the surfaces in Figures 10 and 11.

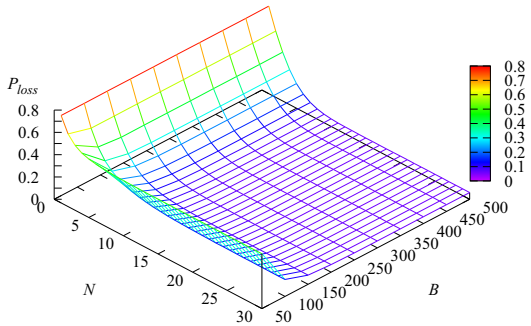

Figure 13. The dependence of the probability of losing an arbitrary order P_{loss} on the parameters N and B.

Figure 14 shows the dependence of the probability $P_{no-sharing}$ on the parameters N and B. This probability approaches the value 1 when B increases. When B is small, this probability quickly decreases when N increases, and the situation that the bandwidth B is sufficient for the service of all orders becomes rare. Under any fixed value of B, this figure can help one to answer the following managerial question: How many orders can be admitted to the system to guarantee that an arbitrary order will receive the required, not the reduced, service rate with a probability higher than some value? For example, this value can be fixed in a service-level agreement between customers generating the orders and the service provider.

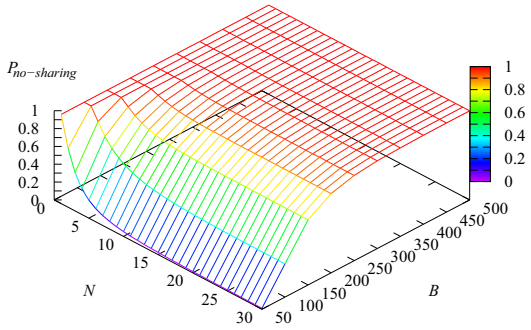

Figure 14. The dependence of the probability $P_{no-sharing}$ on the parameters N and B.

Having proved the impact of the parameters N and B on the system performance measures, we can solve various problems of unconditional or conditional (with restrictions on the values of certain performance measures) optimization of the system operation. As an example, we fix the following criterion for the quality of the system's operation:

$$E = E(B, N) = a\mu_{out} - c_1\lambda P_{ent-loss} - c_2\lambda P_{push-loss} - c_3\lambda P_{imp-loss} - d_1 B - d_2 N.$$

The value E represents the revenue that the service provider gains per unit of time. Here, a is the revenue earned through the service of one order, and c_k, $k = 1, 2, 3$, denotes the charges because of the loss of one order because of admission control as a result of the order removal caused by higher-priority order arrival or an excessively long servicing (due to order impatience). The coefficient d_1 is the cost of the use of a unit of bandwidth during a unit of time. The coefficient d_2 is the cost of the possible maintenance of one order in service (i.e., the cost of a used multiplexer per order) during a unit of time.

Consider the following values of the cost coefficients:

$$a = 5, c_1 = 1, c_2 = 3, c_3 = 4, d_1 = 0.0005, d_2 = 0.001.$$

We fixed the values c_k, $k = 1, 2, 3$, to be $c_1 < c_2$ and $c_1 < c_3$ because it seems better not to admit an order into the system at all (and not to waste the system's resources for its processing) than to admit it upon arrival, spend system resources, but then obtain no revenue and disappoint the user who has generated this order. We also assumed $c_2 < c_3$ to increase the value of type-3 orders. Due to being the lowest priority, type-3 orders are more often pushed out of service and have the longest service time (with the lowest impatience rate).

The 3-D graph that shows the dependence of the revenue criterion E on the parameters N and B is in Figure 15. Table 1 contains additional information about the optimal values N^* of N and E^* of the criterion E for several values of the available bandwidth B.

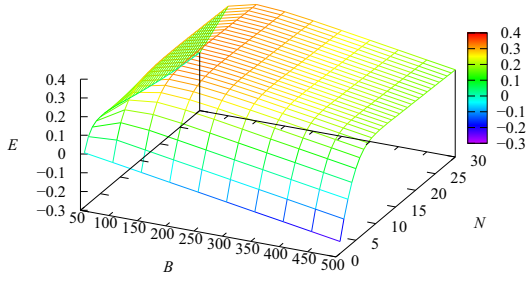

Figure 15. The dependence of the revenue criterion E on the parameters N and B.

Table 1. The optimal values E^* and N^* for different values of the bandwidth B.

B	50	100	150	200	250	300	350	400	450	500
E^*	0.111	0.2962	0.3344	0.3247	0.3045	0.2814	0.2571	0.2324	0.2075	0.1825
N^*	6	15	14	13	13	12	12	12	12	12

Thus, the optimal value $E^*(B^*, N^*)$ is 0.33438. It is achieved when the server bandwidth equals 150 and the limiting value N is equal to 14. It is worth noting that the optimal value of the criterion achieved when $B = 150$ is essentially higher than the values of this criterion when B is small ($E = 0.110987$ for $B = 50$) and when B is large ($E = 0.182535$ for $B = 500$). Thus, the obtained results can be important for managerial aims.

As it was mentioned above, a variety of different possible optimization problems can be solved with the use of the results presented above.

6. Conclusions

We introduced and explored a new discipline for the simultaneous service of heterogeneous orders with different service requirements and importance. The proposed approach is realistic for applications in a lot of real-world systems. It is assumed that there is a limit N on the number of orders that can be processed in the system at the same time. When the number of orders in the system is not high, they receive a desired constant share of bandwidth (and service rate), depending on their types. Their service processes are mutually independent, as in a conventional multi-server queueing system. However, when the sum of the bandwidths of the orders admitted to the system exceeds the total available bandwidth, the orders are serviced at proportionally reduced rates. An order that arrives when the number of serviced orders is equal to the limit pushes out of service an order of the lowest priority that is currently receiving service, if any exist. The order, whose service is interrupted, is lost. The patience time of any order is restricted, and the order can be removed from the system without full service. Orders can change their priority during the service.

The choice of the optimal value of N is a non-trivial and challenging problem. If N is too small, many orders are rejected upon arrival and the system may be underutilized. If N is too large, the service can become too slow and many orders cannot receive full servicing due to either being pushed out by the arrival of an order of higher priority or impatience. In both cases, due to the loss of orders, the revenue obtained by the service provider is low. Therefore, the problem of the optimal choice of N arises and has been addressed in this paper.

The model analysis was done under the realistic assumption that the orders arrive according to a $BMMAP$ process, which is a generalization of the known MAP process to the cases of heterogeneous orders and their batch arrival. The feasibility of the proposed algorithmic analysis has been shown through the use of a numerical example. In particular, we have presented the result of computing the optimal value of the limit of the number of orders that can receive service simultaneously.

The considered model assumes a loss of orders that occurs when the number of serviced orders is at its maximum. In the future, presented analysis will extended to scenarios in which there is also a restriction on the total bandwidth of the orders admitted for service. It is intuitively clear that, due to the random nature of the flow of arriving orders, the limit defining this restriction can be higher than the server bandwidth. A large excess of the limit over the server bandwidth can ensure an essential

increase in the provider's revenue, but it worsen the quality of the user's service. Thus, some reasonable trade-off should be found through the use of adequate mathematical modeling of the system. Results can be extended to cases in which the rejected or pushed-out order can be stored in a buffer of infinite or finite capacity or can attempt service entry after a random amount of time.

The paper focused on the performance evaluation of the considered model under fixed values of its parameters, including the priorities of the different types of orders, the time until the possible change of each priority, and the probabilities of different changes of a priority. Based on this information, the problem of the optimal choice of all or some of these parameters can be solved. This is a possible issue for future research activities.

Use of AI tools declaration

The authors declare that they have not used artificial intelligence tools in the creation of this article.

Conflict of interest

All authors declare no conflicts of interest that may influence the publication of this paper.

References

1. V. M. Vishnevskii, O. V. Semenova, Mathematical methods to study the polling systems, *Autom. Remote. Control*, **67** (2006), 173–220. http://dx.doi.org/10.1134/S0005117906020019
2. S. Borst, O. Boxma, Polling: past, present, and perspective, *Top*, **26** (2018), 335–369. http://dx.doi.org/10.1007/s11750-018-0484-5
3. V. Vishnevsky, O. Semenova, Polling systems and their application to telecommunication networks, *Mathematics*, **9** (2021), 117. http://dx.doi.org/10.3390/math9020117
4. E. Altman, K. Avrachenkov, U. Ayesta, A survey on discriminatory processor sharing, *Queueing Syst.*, **53** (2006), 53–63. http://dx.doi.org/10.1007/s11134-006-7586-8
5. S. F. Yashkov, Processor-sharing queues: some progress in analysis, *Queueing Syst.*, **2** (1987), 1–17. http://dx.doi.org/10.1007/BF01182931
6. S. F. Yashkov, A. S. Yashkova, Processor sharing: a survey of the mathematical theory, *Autom. Remote. Control*, **68** (2007), 1662–1731. http://dx.doi.org/10.1134/S0005117907090202
7. C. D'Apice, A. Dudin, S. Dudin, R. Manzo, Priority queueing system with many types of requests and restricted processor sharing, *J. Ambient Intell. Humaniz. Comput.*, **14** (2023), 12651–12662. http://dx.doi.org/10.1007/s12652-022-04233-w
8. A. N. Dudin, S. A. Dudin, O. S. Dudina, Analysis of a queueing system with mixed service discipline, *Methodol. Comput. Appl. Probab.*, **25** (2023), 1–19. http://dx.doi.org/10.1007/s11009-023-10042-1
9. J. Nair, A. Wierman, B. Zwart, Tail-robust scheduling via limited processor sharing, *Perform. Evaluation*, **67** (2010), 978–995. http://dx.doi.org/10.1016/j.peva.2010.08.012

10. V. Gupta, J. Zhang, Approximations and optimal control for state-dependent limited processor sharing queues, *Stoch. Syst.*, **12** (2022), 205–225. http://dx.doi.org/10.1287/stsy.2021.0087

11. M. Alencar, M. Yashina, A. Tatashev, Loss queueing systems with limited processor sharing and applications to communication networks, In: *2021 International Conference on Engineering Management of Communication and Technology (EMCTECH)*, 2021, 1–5. http://dx.doi.org/10.1109/EMCTECH53459.2021.9618978

12. M. Yashina, A. Tatashev, M. S. de Alencar, Loss probability in priority limited processing queueing system, *Math. Meth. Appl. Sci.*, **48** (2023), 13279–13288. http://dx.doi.org/10.1002/mma.9249

13. M. Telek, B. Van Houdt, Response time distribution of a class of limited processor sharing queues, *ACM SIGMETRICS Perform. Eval. Rev.*, **45** (2018), 143–155. http://dx.doi.org/10.1145/3199524.3199548

14. K. E. Samouylov, E. S. Sopin, I. A. Gudkova, Sojourn time analysis for processor sharing loss queuing system with service interruptions and MAP arrivals, *Commun. Comput. Inf. Sci.*, **678** (2016), 406–417. http://dx.doi.org/10.1007/978-3-319-51917-3-36

15. S. Dudin, A. Dudin, O. Dudina, K. Samouylov, Analysis of a retrial queue with limited processor sharing operating in the random environment, *Lect. Notes Comput. Sci.*, **10372** (2017), 38–49. http://dx.doi.org/10.1007/978-3-319-61382-6-4

16. A. N. Dudin, S. A. Dudin, O. S. Dudina, K. E. Samouylov, Analysis of queueing model with processor sharing discipline and customers impatience, *Oper. Res. Perspect.*, **5** (2018), 245–255. http://dx.doi.org/10.1016/j.orp.2018.08.003

17. H. Masuyama, T. Takine, Sojourn time distribution in a $MAP/M/1$ processor-sharing queue, *Oper. Res. Lett.*, **31** (2003), 406–412. http://dx.doi.org/10.1016/S0167-6377(03)00028-2

18. A. N. Dudin, O. S. Dudina, S. A. Dudin, O. I. Kostyukova, Optimization of road design via the use of a queueing model with transit and local users and processor sharing discipline, *Optimization*, **71** (2022), 3147–3164. http://dx.doi.org/10.1080/02331934.2021.2009827

19. A. Ghosh, A. D. Banik, An algorithmic analysis of the $BMAP/MSP/1$ generalized processor-sharing queue, *Comput. Oper. Res.*, **79** (2017), 1–11. http://dx.doi.org/10.1016/j.cor.2016.10.001

20. M. Nuyens, W. V. D. Weij, Monotonicity in the limited processor-sharing queue, *Stoch. Models*, **25** (2009), 408–419. http://dx.doi.org/10.1080/15326340903088545

21. J. Zhang, B. Zwart, Steady state approximations of limited processor-sharing queues in heavy traffic, *Queueing Syst.*, **60** (2008), 227–246. http://dx.doi.org/10.1007/s11134-008-9095-4

22. I. D. Moscholios, V. G. Vassilakis, M. D. Logothetis, A. C. Boucouvalas, State-dependent bandwidth sharing policies for wireless multirate loss networks, *IEEE Trans. Wirel. Commun.*, **16** (2017), 5481–5497. http://dx.doi.org/10.1109/TWC.2017.2712153

23. S. Borst, M. Mandjes, M. Van Uitert, Generalized processor sharing queues with heterogeneous traffic classes, *Adv. Appl. Prob.*, **35** (2003), 806–845. http://dx.doi.org/10.1239/aap/1059486830

24. A. Dudin, O. Dudina, S. Dudin, K. Samouylov, Analysis of single-server multi-class queue with unreliable service, batch correlated arrivals, customers impatience, and dynamical change of priorities, *Mathematics*, **9** (2021), 1257. http://dx.doi.org/10.3390/math9111257

25. V. Klimenok, A. Dudin, O. Dudina, I. Kochetkova, Queuing system with two types of customers and dynamic change of a priority, *Mathematics*, **8** (2020), 824. http://dx.doi.org/10.3390/MATH8050824

26. P. Cao, J. Xie, Optimal control of a multiclass queueing system when customers can change types, *Queueing Syst.*, **82** (2016), 285–313. http://dx.doi.org/10.1007/s11134-015-9466-6

27. Q. M. He, J. Xie, X. Zhao, Priority queue with customer upgrades, *Nav. Res. Logist.*, **59** (2012), 362–375. http://dx.doi.org/10.1002/nav.21494

28. J. Xie, P. Cao, B. Huang, M. E. H. Ong, Determining the conditions for reverse triage in emergency medical services using queuing theory, *Int. J. Prod. Res.*, **54** (2012), 3347–3364. http://dx.doi.org/10.1080/00207543.2015.1109718

29. V. A. Fajardo, S. Drekic, Waiting time distributions in the preemptive accumulating priority queue, *Methodol. Comput. Appl. Probab.*, **19** (2017), 255–284. http://dx.doi.org/10.1007/s11009-015-9476-1

30. M. Mojalal, D. A. Stanford, R. J. Caron, The lower-class waiting time distribution in the delayed accumulating priority queue, *INFOR Inf. Syst. Oper. Res.*, **58** (2020), 60–86. http://dx.doi.org/10.1080/03155986.2019.1624473

31. K. C. Sharma, G. C. Sharma, A delay dependent queue without preemption with general linearly increasing priority function, *J. Oper. Res. Soc.*, **45** (1994), 948–953. http://dx.doi.org/10.2307/2584019

32. D. A. Stanford, P. Taylor, I. Ziedins, Waiting time distributions in the accumulating priority queue, *Queueing Syst.*, **77** (2014), 297–330. http://dx.doi.org/10.1007/s11134-013-9382-6

33. O. Xie, Q. M. He, X. Zhao, Stability of a priority queueing system with customer transfers, *Oper. Res. Lett.*, **36** (2008), 705–709. http://dx.doi.org/10.1016/j.orl.2008.06.007

34. J. Xie, T. Zhu, A. K. Chao, S. Wang, Performance analysis of service systems with priority upgrades, *Ann. Oper. Res.*, **253** (2017), 683–705. http://dx.doi.org/10.1007/s10479-016-2370-6

35. M. Cildoz, A. Ibarra, F. Mallor, Accumulating priority queues versus pure priority queues for managing patients in emergency departments, *Oper. Res. Health Care*, **23** (2019), 100224. http://dx.doi.org/10.1016/j.orhc.2019.100224

36. Q. M. He, Queues with marked customers, *Adv. Appl. Prob.*, **28** (1996), 567–587. http://dx.doi.org/10.2307/1428072

37. A. N. Dudin, V. I. Klimenok, V. M. Vishnevsky, *The Theory of Queuing Systems with Correlated Flows*, Berlin: Springer Cham, 2020. http://dx.doi.org/10.1007/978-3-030-32072-0

38. A. Dudin, C. S. Kim, O. Dudina, S. Dudin, Multi-server queueing system with generalized phase type service time distribution, *Ann. Oper. Res.*, **239** (2016), 401–428. http://dx.doi.org/10.1007/s10479-014-1626-2

39. C. S. Kim, S. A. Dudin, O. S. Taramin, J. Baek, Queueing system $MAP/PH/N/N + R$ with impatient heterogeneous customers as a model of call center, *Appl. Math. Model.*, **37** (2013), 958–976. http://dx.doi.org/10.1016/j.apm.2012.03.021

40. C. Kim, A. Dudin, S. Dudin, O. Dudina, Mathematical model of operation of a cell of a mobile communication network with adaptive modulation schemes and handover of mobile users, *IEEE Access*, **9** (2021), 106933–106946. http://dx.doi.org/10.1109/ACCESS.2021.3100561

41. S. Lee, S. Dudin, O. Dudina, C. Kim, V. Klimenok, A priority queue with many customer types, correlated arrivals and changing priorities, *Mathematics*, **8** (2020), 1292. http://dx.doi.org/10.3390/MATH8081292

42. A. Graham, *Kronecker Products and Matrix Calculus with Applications*, New York: Courier Dover Publications, 2018.

AIMS Press