mathematics

MDPI

Article

Randomized Threshold Strategy for Providing Flexible Priority in Multi-Server Queueing System with a Marked Markov Arrival Process and Phase-Type Distribution of Service Time

A. N. Dudin *, S. A. Dudin and O. S. Dudina

Department of Applied Mathematics and Computer Science, Belarusian State University, 4, Nezavisimosti Ave., 220030 Minsk, Belarus; dudins@bsu.by (S.A.D.); dudina@bsu.by (O.S.D.)
* Correspondence: dudin@bsu.by

Abstract: In this paper, we analyze a multi-server queueing system with a marked Markov arrival process of two types of customers and a phase-type distribution of service time depending on the type of customer. Customers of both types are assumed to be impatient and renege from the buffers after an exponentially distributed number of times. The strategy of flexible provisioning of priorities is analyzed. It assumes a randomized choice of the customers from the buffers, with probabilities dependent on the relation between the number of customers in a priority finite buffer and the fixed threshold value. To simplify the construction of the underlying Markov chain and the derivation of the explicit form of its generator, we use the so-called generalized phase-type distribution. It is shown that the created Markov chain fits the category of asymptotically quasi-Toeplitz Markov chains. Using this fact, we show that the considered Markov chain is ergodic for any value of the system parameters and compute its stationary distribution. Expressions for key performance measures are presented. Numerical results that show how the parameters of the control strategy affect the system's performance measurements are given. It is shown that the results can be used for managerial purposes and that it is crucial to take correlation in the arrival process into account.

Keywords: priority queueing system; *MMAP*; impatience; optimization; generalized phase-type distribution

MSC: 60K25; 60K30; 68M20; 90B22

Citation: Dudin, A.N.; Dudin, S.A.; Dudina, O.S. Randomized Threshold Strategy for Providing Flexible Priority in Multi-Server Queueing System with a Marked Markov Arrival Process and Phase-Type Distribution of Service Time. *Mathematics* **2023**, *11*, 2669. https://doi.org/10.3390/math11122669

Academic Editors: Vladimir Rykov and Dmitry Efrosinin

Received: 5 May 2023
Revised: 8 June 2023
Accepted: 11 June 2023
Published: 12 June 2023

1. Introduction

Queueing theory is a rapidly developing branch of applied probability that helps solve problems related to the sharing and scheduling of finite resources that have to be used by competing users. Systems with heterogeneous customers have been investigated less in the existing queueing literature. The varying requirements of several types of customers for the required system resource, indicators of the quality of service, and their different importance and value all lead to their needing different treatments. In particular, a powerful tool for the enhancement of the operation of such systems is the assignment of different priorities in accessing the system resources for different types of customers. The theory of priority queues is pretty well developed in the case of single-server queues (see, e.g., [1–5]). However, the operation of many real-world systems can be adequately described only by multi-server priority queueing systems. The problem of optimal management in such systems is much more complicated than in single-server systems due to the larger dimensionality of the stochastic process that describes the dynamics of the system. However, the importance of consideration of such priority queueing systems made them a popular subject of research; see, e.g., the papers [6–16] published within the last two years.

There are a lot of various priority schemes (preemptive and non-preemptive, alternating, static and dynamic, changing and accumulating, etc.) known in the literature;

some references can be found, e.g., in [17]. All these schemes have their own advantages and disadvantages. For example, the static priorities assuming strict adherence to the established rules of picking up from the queue are more easily realized in real systems. The preemptive priorities suggesting service interruption in the case of the arrival of a customer who has a higher priority than the one receiving service are fine for high-priority customers but are discriminatory with respect to low-priority customers. The dynamic priorities (see, e.g., [18,19],) give more flexibility but require efforts to monitor the system states. Therefore, to better manage the operation of different real-world systems, different new priority schemes are generated in the literature. For example, the scheme proposed in [17] suggests the maintenance of auxiliary finite input buffers with different rates of customer transfer to the main buffer. A proper choice of the capacity of the buffers and the corresponding rates allows one to reach any desired degree of non-preemptive priority for one class of customers over another. Another flexible scheme for providing non-preemptive priority was offered in [20]. That scheme assumes the randomized choice of the next customer for service, with probabilities of choice depending on the length of the queue of the priority customers. The scheme was applied and analyzed for the single-server queue. In this paper, we extend the analysis to the case of a multi-server queue. It is worth noting that there exist works (see, e.g., [21]) where a priority to certain flows, e.g., flows of handover users in cellular mobile networks or primary users in cognitive radio systems, is provided via different kinds of schemes of reservation of some servers of the multi-server queueing system for service of priority users. Here, we do not touch on such a possibility. Priority is provided via the proper strategy of customers picking up from the queues.

It is already well known that flows of customers receiving service in many telecommunication systems, contact centers, hospitals, etc., exhibit a so-called bursty nature. The instantaneous arrival rate may significantly fluctuate during system operation. This makes the very popular-in-the-literature, stationary Poisson model of the arrival process a poor descriptor of real flows. The use of this model is not suitable for reliable prediction of the main performance measures of the systems. It may lead to huge errors in the estimation of the required system resources sufficient for providing the desired quality of service. Therefore, in this paper, we assume that the arrival flows of two types of customers receiving service in the system are defined by the much more general Marked Markov Arrival Process ($MMAP$), see, e.g., [22–24]. The $MMAP$ is an extension of the Markov Arrival Process (MAP) for accounting for the possibility of heterogeneous customer arrivals. In contrast to the stationary Poisson arrival process, the MAP is suitable for modeling flows with a non-zero coefficient of correlation and high variation in inter-arrival times. For more information about the basic properties, particular cases, and possible applications, as well as about the analysis of the systems with the MAP, see, e.g., [24–44]. The multi-server priority queues with $MMAP$ arrival process have been analyzed, e.g., in [9,13,16,17,20,45–47].

The model considered in this paper assumes a so-called phase-type (PH) distribution of service times of both types (see, e.g., [24,48–52]). This distribution essentially generalizes the well-known exponential distribution but still allows us to obtain nice analytical results. It can be applied to approximate any non-negative random variable distribution. In particular, it allows taking into account the variance of a random variable while the exponential distribution has a coefficient of variation equal to 1. However, the use of this distribution instead of the exponential distribution for the description of, e.g., service time, leads to the necessity of extending the dimension of the state space of the Markovian process describing the dynamics of a queueing system. It is necessary to keep track of the state of the underlying process (phase) of the PH distribution in each busy server. If the considered queueing system has many servers, it essentially complicates the analysis of the system. If the number of identical servers is equal to N while the number of phases is equal to M, then the direct way of accounting for the phase of the underlying processes of service on all servers leads to the necessity of operating with matrices of size M^N describing the dynamics of the service process. When at least one of the numbers N and M is relatively large (while the number N of servers in real systems can be pretty large), operating with

matrices of such a size can be difficult. This is why in this paper, we use another way to monitor the service process. This way assumes an account of the number of servers providing service at each of M possible phases. In the case when M is relatively small, the state space of the service processes becomes much smaller, namely, $\binom{N+M-1}{M-1} = \frac{(N+M-1)!}{N!(M-1)!}$. For example, in many cases, the value of M equal to 2 allows us to have a good catch on the variance of service time. For $M = 2$, the size of the matrices that have to be treated is equal to $N + 1$, which may be much less than M^N. This way of describing the service process in multi-server queues was offered in [53,54].

In those papers, the recursive formulas for the computation of the matrices defining transition rates of the multi-dimensional Markov process, which describes the service process in many servers without a service completion and at the moments of new service beginning or ending, are presented. Formulas for computation of these matrices for the value n of the number of currently busy servers (among N available servers), $n = \overline{1, N}$, depend on both n and N. This is not convenient, especially when it is required to make computations for several values of N, e.g., in the process of computing the minimum required number N of servers to guarantee the fixed values of the service quality indicators. In [55], the formulas derived in [53,54] are modified to eliminate the use of the value of N in recursive computations.

In the considered model, it is assumed that the parameters of PH distribution of the service time of customers of different types may be different. To avoid the necessity of separately monitoring the number of each type of customer in service and make the analysis easier, we use the notation of the generalized phase-type distribution introduced in [56,57].

We take into consideration the possibility of customers departing from the buffers due to impatience. Many real-world systems are inherently characterized by the impatience of their customers. For example, customers (information units) in telecommunication networks can depart from the waiting area due to information obsolescence, users' departure from the service zone, etc. In contact centers, impatience is explained by the psychology of users. In retail networks and health systems (such as blood banks or organ transplantation hospitals), impatience is explained by the restricted term of the suitability of the items required for service. Recent lists of related research on queues with impatient customers can be found, e.g., in [58–61].

Account of customers' impatience makes the considered model potentially useful; beyond the evident applications in telecommunications, it is also useful in the following systems. When two competing flows of passengers or vehicles, which arrive from different directions, merge into one flow, a popular mechanism of merging implementation consists of admitting a certain percentage of users from one direction and another. The considered model can help answer the question of how to optimally redistribute the percentage when one of the queues becomes large. In call centers, agents simultaneously handle the voice requests, which are admitted via the buffer of a finite capacity, as well as the requests arriving via the Internet that are placed in the buffer of a potentially infinite capacity. Over time, agents select for service waiting voice requests and Internet requests in some proportion. However, when the buffer for voice requests becomes close to its capacity (exceeds some threshold) and there is a high risk of their loss, the agents can decrease the percentage of Internet users admitted for service and, correspondingly, increase the percentage of voice users.

The structure of this paper is as follows. In Section 2, the model of the considered queueing system is completely defined. In Section 3, a multi-dimensional Markov chain describes how the system operates. The generator of this chain is derived there, and the problems of the existence of the stationary distribution of this Markov chain and the computation of this distribution are briefly discussed. In Section 4, formulas for the computation of some performance measures of the system are presented. Some numerical illustrations are presented in Section 5. The paper is concluded in Section 6.

2. Mathematical Model

We consider an N-server queuing system with two buffers, the structure of which is displayed in Figure 1.

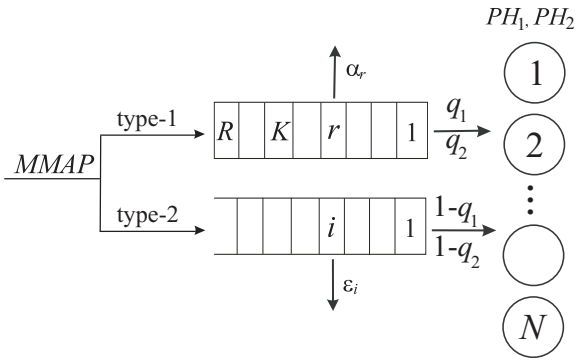

Figure 1. Structure of the system under study.

The system processes customers of two types. The customer's arrival in the system is described by the $MMAP$ which is defined by the underlying continuous-time Markov chain (MC) ν_t, $t \geq 0$, with the state space $\{1, 2, \ldots, W\}$, and the square matrices D_0, D_1, and D_2 of size W. The negative diagonal entries of the matrix D_0 define, up to the sign, the transition rates of the MC ν_t, $t \geq 0$, from the corresponding states. The non-diagonal entries of the matrix D_0 define the transition rates that do not lead to the arrival of any customer. The entries of the matrices D_l, $l = 1, 2$, are the transition rates of the MC ν_t, $t \geq 0$, that cause the arrival of a type-l customer. More information about the $MMAP$ and its characteristics can be found, e.g., in [24].

Let us denote the total customers' arrival rate as λ and type-l customers' arrival rate as λ_l, $l = 1, 2$. These intensities are determined by the formulas:

$$\lambda = \boldsymbol{\theta}(D_1 + D_2)\mathbf{e},\ \lambda_1 = \boldsymbol{\theta}D_1\mathbf{e},\ \lambda_2 = \boldsymbol{\theta}D_2\mathbf{e}$$

where the stationary probability vector $\boldsymbol{\theta}$ of process ν_t is determined to be the one solution to the system

$$\boldsymbol{\theta}(D_0 + D_1 + D_2) = \mathbf{0},\ \boldsymbol{\theta}\mathbf{e} = 1$$

where the row vector $\mathbf{0}$ consists of 0 s and the column vector $\mathbf{e}$ consists of 1 s.

The service time of a type-l, $l = 1, 2$, customer has the phase-type distribution PH_l that is defined by the underlying continuous-time MC $m_t^{(l)}$, $t \geq 0$, with the state space $\{1, 2, \ldots, M_l\}$, the vector $\mathbf{b}_l$ and the matrix S_l. The stochastic row vector $\mathbf{b}_l$ determines the initial state of the process $m_t^{(l)}$ at the service beginning epoch, and the subgenerator S_l defines rates of transition of the process $m_t^{(l)}$ inside the state space $\{1, 2, \ldots, M_l\}$. The service time finishes when the process $m_t^{(l)}$ transits to the absorbing state. The mean service time of type-l customers is calculated as $\mathbf{b}_l(-S_l)^{-1}\mathbf{e}$. For further details on the PH distribution, its moments, and specific cases, see, for example, [24].

If a customer of any type arrives at the system when there is a free server, this customer immediately starts service. A type-2 customer joins the infinite buffer if there are no available servers during its arrival period. Let us call this buffer Buffer-2. If there are no free servers during the arrival epoch of a type-1 customer, this customer tries to join the finite buffer of capacity R (Buffer-1). This trial will be successful in the case of a not full buffer. The arriving customer is lost forever if Buffer-1 is full during the type-1 customer arrival moment.

If during the epoch when some server finishes the service only customers of one type are in the buffers, the customer of such a type is chosen for service. If there are customers of two types in the buffers at the service completion moment, the type of customer that will receive service next is defined according to the following rule. If the number of type-1

customers in the buffer is less than the parameter K, a type-1 customer is chosen for service with the probability q_1, and with the complimentary probability $1 - q_1$, a type-2 customer starts service. If the number of type-1 customers in the buffer is greater or equal to K, a type-1 customer is chosen for service with the probability q_2, $q_2 \geq q_1$, and with the complimentary probability $1 - q_2$ a type-2 customer starts service. We assume that the parameters K, q_1, and q_2 are control parameters, and the task of modeling is to develop a mathematical tool that allows us to find the optimal values of these parameters according to a fixed cost criterion.

Each type of customer is assumed to be impatient and depart from the system without service if the service does not start within an exponentially distributed time since the arrival moment. The total rate of customers' reneging from the corresponding buffer depends on the number of customers staying in the buffer. We assume that if there are r customers in Buffer-1, the total rate of reneging by type-1 customers is equal to α_r, $\alpha_r \geq 0$, $r = \overline{1, R}$. If there are i customers in the infinite Buffer-2, the total rate of type-2 customers reneging is equal to ϵ_i, $\epsilon_i \geq 0$, and $\epsilon_i \to \infty$ if $i \to \infty$.

Let us analyze the stationary behavior of the model under consideration.

3. The Process of the System States and Its Analysis

The operation of the system under study is described by the following regular irreducible $(M_1 + M_2 + 4)$-dimensional continuous-time MC

$$\zeta_t = \{i_t, r_t, n_t, \nu_t, \sigma_t^{(1)}, \ldots, \sigma_t^{(M_1)}, \gamma_t^{(1)}, \ldots, \gamma_t^{(M_2)}\},\ t \geq 0,$$

where at time t, $t \geq 0$,

- i_t is the number of customers in the system, $i_t \geq 0$;
- r_t is the number of type-1 customers in Buffer-1, $r_t = \overline{0, \min\{\max\{0, i - N\}, R\}}$;
- n_t is the number of type-1 customers in service, $n_t = \overline{0, \min\{N, i_t\}}$;
- ν_t is the state of the underlying process of the $MMAP$ of customers, $\nu_t = \overline{1, W}$;
- $\sigma_t^{(m)}$ is the number of type-1 customers at the m-th phase of the PH_1 service process, $\sigma_t^{(m)} = \overline{0, n_t}$, $m = \overline{1, M_1}$, $\sum\limits_{m=1}^{M_1} \sigma_t^{(m)} = n_t$;
- $\gamma_t^{(l)}$ is the number of type-2 customers at the l-th phase of the PH_2 service process, $\gamma_t^{(l)} = \overline{0, \min\{N, i_t - n_t\}}$, $l = \overline{1, M_2}$, $\sum\limits_{l=1}^{M_2} \gamma_t^{(l)} = \min\{N, i_t - n_t\}$.

One can see that the process ζ_t is complicated, and its analysis may be cumbersome. To make the investigation easier, we use the notion of the generalized phase-type distribution, see, [56,57]. Instead of the separate consideration of service processes of each type of customer, we suggest that the customer's service time at a server has the generalized PH (GPH) distribution having the parameters $(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, S)$ where the vectors $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ and subgenerator S are defined by the formulas:

$$\boldsymbol{\beta}_1 = (\mathbf{b}_1, \underbrace{0, 0, \ldots, 0}_{M_2}),\ \ \boldsymbol{\beta}_2 = (\underbrace{0, 0, \ldots, 0}_{M_1}, \mathbf{b}_2),\ \ S = \begin{pmatrix} S_1 & O \\ O & S_2 \end{pmatrix}.$$

The duration of service is governed by the underlying process m_t with the state space $\{1, 2, \ldots, M\}$, where $M = M_1 + M_2$. The state of the process m_t during the service beginning epoch is defined according to the probabilistic vector $\boldsymbol{\beta}_l$, $l = 1, 2$, if a type-l customer starts service. The transition rates into the absorbing state (which implies service completion on one of the busy servers) are given by the entries of the column vector $\mathbf{S}_0 = \begin{pmatrix} -S_1\mathbf{e} \\ -S_2\mathbf{e} \end{pmatrix}$.

The use of GPH allows us to significantly simplify the process of defining the dynamics of the system and, correspondingly, its analysis.

The following regular irreducible continuous-time *MC* can be used to describe the behavior of the system under consideration:

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

The meaning of the first three components $\{i_t, r_t, \nu_t\}$ is the same as for the *MC* ζ_t. The component $\eta_t^{(m)}$ of the vector-valued process $\boldsymbol{\eta}_t = \{\eta_t^{(1)}, \ldots, \eta_t^{(M)}\}$, $t \geq 0$, denotes the number of customers at the m-th phase of the GPH service process at the moment t, $\eta_t^{(m)} = \overline{0, \min\{i_t, N\}}$, $m = \overline{1, M}$, $\sum\limits_{m=1}^{M} \eta_t^{(m)} = \min\{i_t, N\}$.

To describe the transition rates of the *MC* ξ_t, let us first introduce the following auxiliary matrices:

$A_i = A_i(S)$, $i = \overline{1, N}$, define the rates of transitions of the process $\boldsymbol{\eta}_t$, $t \geq 0$, which do not lead to the end of service provided that i servers are busy;

$P_i(\boldsymbol{\beta}_1)$, $P_i(\boldsymbol{\beta}_2)$, $i = \overline{0, N-1}$, define the transition probabilities of the process $\boldsymbol{\eta}_t$, $t \geq 0$, when i servers are busy and a type-1 or type-2 customer is chosen for service, correspondingly;

$L_i = L_i(\mathbf{S}_0)$, $i = \overline{1, N}$, define the transitions rates of the process $\boldsymbol{\eta}_t$, $t \geq 0$, which cause the finish of service if i servers are busy;

Δ_i , $i = \overline{1, N}$, define the rates of exit of the process $\boldsymbol{\eta}_t$, $t \geq 0$, from its states when i customers are in service.

Note that the recursive algorithms for the computation of the matrices A_i, L_i, Δ_i, $i = \overline{1, N}$, and $P_i(\boldsymbol{\beta}_1)$, $P_i(\boldsymbol{\beta}_2)$, $i = \overline{0, N-1}$, are available in paper [55] and represent the advanced modification of the algorithms earlier presented in [53,54] and used, e.g., in [62–64].

We enumerate the components $\{i_t, r_t, \nu_t\}$ of the *MC* ξ_t, $t \geq 0$, in the direct lexicographic order and the components $\{\eta_t^{(1)}, \ldots, \eta_t^{(M)}\}$ in reverse lexicographic order. All states with the value i of the first component i_t are called the level i of the *MC* ξ_t.

We refer to the chain's infinitesimal generator as G. All possible transition rates of the chain under study during an infinitesimally short time interval are defined as the entries of the generator G.

The notation used up until this point, and the new notation are compiled in Table 1 for clarity and to make the description of the shape of the blocks of the generator G simpler.

Table 1. Notation.

N	the number of servers
D_0, D_1, D_2	the $W \times W$ matrices that define the $MMAP$
λ_l, $l = 1, 2$	the average arrival rate of type-l customers
$\mathbf{0}$	row vector which consists of 0s
$\mathbf{e}$	column vector which consists of 1s
$(\boldsymbol{\beta}_l, S_l)$, $l = 1, 2$	an irreducible representation of the type-l customer PH service time distribution
R	Buffer-1 capacity
K, q_1 and q_2	control parameters that define the choice of customers for service
α_r, $r = \overline{1, R}$	the total rate of reneging by type-1 customers if there are r customers in Buffer-1
ϵ_i, $i \geq 1$	the total rate of reneging by type-2 customers if there are i customers in Buffer-2
$(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, S)$	the parameters that define GPH service time distribution with M states

Table 1. *Cont.*

$\mathbf{S}_0$	$\begin{pmatrix} -S_1\mathbf{e} \\ -S_2\mathbf{e} \end{pmatrix}$
$A_i = A_i(S)$, $i = \overline{1,N}$	the matrix that defines the rates of transitions that do not lead to the end of service provided that i servers are busy
$P_i(\boldsymbol{\beta}_1)$, $P_i(\boldsymbol{\beta}_2)$, $i = \overline{0,N-1}$	the matrix that defines the transition probabilities of the process $\boldsymbol{\eta}_t$, $t \geq 0$, when i servers are busy and a type-1 or type-2 is chosen for service, correspondingly
$L_i = L_i(\mathbf{S}_0)$, $i = \overline{1,N}$	the matrix that defines the transition rates that cause the finish of service if i servers are busy
Δ_i, $i = \overline{1,N}$	the matrix that defines the rates of exit of the service process from its states when i customers are in service
T_N	$\binom{N+M-1}{M-1} = \frac{(N+M-1)!}{N!(M-1)!}$
I	the identity matrix
O	a zero matrix
$\otimes$ and $\oplus$	the symbols of Kronecker product and sum of matrices, see, e.g., [65–67]
$\mathrm{diag}\{a_1,\ldots,a_l\}$	a block-diagonal matrix with the diagonal blocks $a_1,\ldots,a_l$
$C_l^{(1)}$	$\mathrm{diag}\{0,\alpha_1,\alpha_2,\ldots,\alpha_l\}$, $l = \overline{1,R-1}$
$C_l^{(2)}$	$\mathrm{diag}\{\epsilon_l,\epsilon_{l-1},\ldots,\epsilon_1,0\}$, $l = \overline{1,R-1}$
C	$\mathrm{diag}\{0,\alpha_1,\alpha_2,\ldots,\alpha_R\}$
$\bar{C}_l$	$\mathrm{diag}\{\epsilon_l,\epsilon_{l-1},\ldots,\epsilon_{l-R-1},\epsilon_{l-R}\}$, $l \geq R$
E_l^+, $l = \overline{0,R-1}$	an $(l+1)\times(l+2)$ matrix having all zero elements except the elements $(E_l^+)_{k,k+1}$, $k = \overline{0,l}$, which are equal to 1
$\bar{E}_l^+$, $l = \overline{0,R-1}$	an $(l+1)\times(l+2)$ matrix having all zero elements except the elements $(\bar{E}_l^+)_{k,k}$, $k = \overline{0,l}$, which are equal to 1
E^+	a square matrix of size $R+1$ having all zero elements except the elements $(E^+)_{k,k+1}$, $k = \overline{0,R-1}$, which are equal to 1
$\hat{I}$	a square matrix of size $R+1$ having all zero elements except the one $(\hat{I})_{R,R} = 1$
E_l^-, $l = \overline{1,R}$	an $(l+1)\times l$ matrix having all zero elements except the element $(E_l^-)_{k,k-1}$, $k = \overline{1,l}$,which are defined as follows: $(E_l^-)_{k,k-1} = \begin{cases} q_1, & if\ k \in \{1,2,\ldots,K-1\}, k \neq l, \\ q_2, & if\ k \in \{K,K+1,\ldots,l-1\} \\ 1, & if\ k = l. \end{cases}$
$\bar{E}_l^-$, $l = \overline{1,R}$	an $(l+1)\times l$ matrix having all zero elements except the elements $(\bar{E}_l^-)_{k,k}$, $k = \overline{0,l-1}$, which are defined as follows: $(\bar{E}_l^-)_{k,k} = \begin{cases} 1, & if\ k = 0, \\ 1-q_1, & if\ k \in \{1,2,\ldots,K-1\}, \\ 1-q_2, & if\ k \in \{K,K+1,\ldots,l-1\}, \end{cases}$
I_l^-, $l = \overline{1,R}$	an $(l+1)\times l$ matrix having all zero elements except the elements $(I_l^-)_{k,k-1}$, $k = \overline{1,l}$, which are equal to 1;
$\bar{I}_l^-$, $l = \overline{1,R}$	an $(l+1)\times l$ matrix having all zero elements except the elements $(\bar{I}_l^-)_{k,k}$, $k = \overline{0,l-1}$, which are equal to 1;

Table 1. *Cont.*

E^-	a square matrix of size $R+1$ having all zero elements except the elements $(E^-)_{k,k-1}$, $k=\overline{1,R}$, which are defined as follows: $(E^-)_{k,k-1}=\begin{cases} q_1, & if\ k\in\{1,2,\dots,K-1\},\\ q_2, & if\ k\in\{K,K+1,\dots,R\}.\end{cases}$
$\bar{E}^-$	a square matrix of size $R+1$ having all zero elements except the elements $(\bar{E}^-)_{k,k}$, $k=\overline{0,R-1}$, which are defined as follows: $(\bar{E}^-)_{k,k}=\begin{cases} 1, & if\ k=0,\\ 1-q_1, & if\ k\in\{1,2,\dots,K-1\},\\ 1-q_2, & if\ k\in\{K,K+1,\dots,R\},\end{cases}$
I^-	a square matrix of size $R+1$ having all zero elements except the elements $(I^-)_{k,k-1}$, $k=\overline{1,R}$, which are equal to 1

Theorem 1. *The generator G has the following block-three-diagonal form:*

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

The non-zero blocks $G_{i,j}$, $i,j\geq 0$, determining the transition rates from level i to level j are defined as follows:

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

$$G_{i,i}=D_0\oplus(A_i+\Delta_i),\ i=\overline{1,N},$$

$$G_{i,i}=I_{i-N+1}\otimes(D_0\oplus(A_N+\Delta_N))-C_{i-N}^{(1)}\otimes I_{WT_N}-C_{i-N}^{(2)}\otimes I_{WT_N},\ i=\overline{N+1,N+R-1},$$

$$G_{i,i}=I_{R+1}\otimes(D_0\oplus(A_N+\Delta_N))-C\otimes I_{WT_N}-\bar{C}_{i-N}\otimes I_{WT_N}+\hat{I}\otimes D_1\otimes I_{T_N},\ i\geq N+R,$$

$$G_{i,i+1}=D_1\otimes P_i(\boldsymbol{\beta}_1)+D_2\otimes P_i(\boldsymbol{\beta}_2),\ i=\overline{0,N-1},$$

$$G_{i,i+1}=E_{i-N}^{+}\otimes D_1\otimes I_{T_N}+\bar{E}_{i-N}^{+}\otimes D_2\otimes I_{T_N},\ i=\overline{N,N+R-1},$$

$$G_{i,i+1}=E^{+}\otimes D_1\otimes I_{T_N}+I_{R+1}\otimes D_2\otimes I_{T_N},\ i\geq N+R,$$

$$G_{i,i-1}=I_W\otimes L_i,\ i=\overline{1,N},$$

$$G_{i,i-1}=E_{i-N}^{-}\otimes I_W\otimes L_NP_{N-1}(\boldsymbol{\beta}_1)+\bar{E}_{i-N}^{-}\otimes I_W\otimes L_NP_{N-1}(\boldsymbol{\beta}_2)+$$

$$+C_{i-N}^{(1)}I_{i-N}^{-}\otimes I_{WT_N}+C_{i-N}^{(2)}\bar{I}_{i-N}^{-}\otimes I_{WT_N},\ i=\overline{N+1,N+R},$$

$$G_{i,i-1}=E^{-}\otimes I_W\otimes L_NP_{N-1}(\boldsymbol{\beta}_1)+\bar{E}^{-}\otimes I_W\otimes L_NP_{N-1}(\boldsymbol{\beta}_2)+$$

$$+CI^{-}\otimes I_{WT_N}+\bar{C}_{i-N}\otimes I_{WT_N},\ i>N+R.$$

Proof. By analyzing every potential transition of the *MC* xi_t during an interval of infinitesimal length and reconstructing the rates of such transitions in block matrix form, the proof of the theorem is implemented.

Customers of both types arrive and are served one at a time. The possibility of more than one consumer entering the system or being served in a moment that lasts an infinite amount of time is negligibly small. Hence, the generator G has a block-tridiagonal structure, and all the matrices $G_{i,j}$ are zero ones for i and j, such as $|i-j|>1$.

The negative diagonal entries of the matrices $G_{i,i}$, $i\geq 0$, define, up to the sign, the total rate of leaving the corresponding state. The *MC* ξ_t can exit from its state in the following cases:

(1) The underlying arrival process changes its state. The rates of the arrival process transitions are defined by the modulus of the corresponding diagonal entries of the matrix D_0 if $i = 0$ (the system is empty), $D_0 \otimes I_{T_i}$ for $i = \overline{1, N}$, and $I_{\min\{\max\{0,i-N\},R\}+1} \otimes D_0 \otimes I_{T_N}$ for $i \geq N+1$. Note that if the underlying arrival process makes a transition from some state to the same state with a type-1 customer arrival when Buffer-1 is full, the customer is lost, and the exit from the corresponding state does not occur. The rates of these transitions are given by the corresponding diagonal elements of the matrix $\hat{I} \otimes D_1 \otimes I_{T_N}$, $i \geq N+R$.

(2) The transition in the service process on one of the servers if there are busy servers in the system. The rates of these transitions are defined by the modulus of the corresponding diagonal elements of the matrix $I_W \otimes \Delta_i$, $i = \overline{1, N}$ (if i customers are served in the system and there is no customer in the buffers) and by the matrix $I_{(\min\{\max\{0,i-N\},R\}+1)W} \otimes \Delta_N$, $i \geq N+1$ (if there are customers in service and in the buffers).

(3) One of the customers leaves the system due to impatience. If a type-1 customer leaves Buffer-1, the rates of these transitions are defined by the diagonal elements of the matrices $C_{i-N}^{(1)} \otimes I_{WT_N}$ for $i = \overline{N+1, N+R-1}$ and of the matrix $C \otimes I_{WT_N}$ for $i \geq N+R$. The diagonal elements of the matrices $C_{i-N}^{(2)} \otimes I_{WT_N}$, $i = \overline{N+1, N+R-1}$, and of the matrix $\bar{C}_{i-N} \otimes I_{WT_N}$, $i \geq N+R$, are the transition rates in the case of a type-2 customer leaving Buffer-2 due to impatience.

The non-diagonal elements of the blocks $G_{i,i}$ determine the transition rates of the *MC* ξ_t that do not lead to the change in the number of customers in the system $i \geq 0$. These transitions are as follows:

(1) The transition in the underlying arrival process without the arrival of a new customer (the rates of these transitions are the non-diagonal elements of the matrix D_0 if $i = 0$, $D_0 \otimes I_{T_i}$ for $i = \overline{1, N}$, and $I_{\min\{\max\{0,i-N\},R\}+1} \otimes D_0 \otimes I_{T_N}$ for $i \geq N+1$).

(2) The arrival of a type-1 customer when N customers are being served in the system and Buffer-1 is full. Note that in such a case, the arriving customer is not accepted into the system. The corresponding rates are given by the elements of the matrix $\hat{I} \otimes D_1 \otimes I_{T_N}$, $i \geq N+R$.

(3) The change of the state of the service underlying process without service completion. The rates of these transitions are defined by the non-diagonal elements of the matrices $I_W \otimes A_i$, $i = \overline{1, N}$ (if i servers are busy but the buffers are empty) and of the matrices $I_{(\min\{\max\{0,i-N\},R\}+1)W} \otimes A_N$, $i \geq N+1$ (if all servers are busy and there are customers in the buffers).

Using all these reasonings, we prove the form of the blocks $G_{i,i}$, $i \geq 0$.

The elements of the matrices $G_{i,i+1}$, $i \geq 0$, define the transition rates of the components of the *MC* ξ_t that lead to the increase in the number of customers i in the system by one. This can happen only if a new customer is admitted to the system. When there is an idle server at an arrival moment, the transition rates are given by the elements of the matrices $D_l \otimes P_i(\boldsymbol{\beta}_l)$, $i = \overline{0, N-1}$, if type-l, $l = 1, 2$, customer arrives. The Kronecker multiplier $P_i(\boldsymbol{\beta}_l)$ reflects the fact that the arriving customer immediately starts service and, therefore, installation of the initial state of its service underlying process is required. If all servers are busy and an arriving type-1 customer joins Buffer-1, the elements of the matrices $E_{i-N}^{+} \otimes D_1 \otimes I_{T_N}$ define the corresponding rates. The Kronecker multiplier $P_i(\boldsymbol{\beta}_l)$ is replaced here by the Kronecker multiplier I_{T_N} because the new service does not start, and no transition of the service underlying process occurs. If a type-2 customer arrives at the system and is appended to Buffer-2, the corresponding rates are defined by the matrices $\bar{E}_{i-N}^{+} \otimes D_2 \otimes I_{T_N}$, $i = \overline{N, N+R-1}$, and $I_{R+1} \otimes D_2 \otimes I_{T_N}$, $i \geq N+R$. The matrices E_{i-N}^{+} and $\bar{E}_{i-N}^{+}$ reflect the fact of increasing the number of customers of the corresponding type in the buffer.

Next, we explain the formation of the blocks $G_{i,i-1}$, $i \geq 1$. Their elements specify the transition rates of the components of the *MC* ξ_t that lead to the decrease in the number of customers i in the system by one. These transitions are the following:

(1) A customer leaves the system due to the successful completion of the service. In this case, the corresponding rates are given as the elements of the matrices: (i) $I_W \otimes$

L_i, $i = \overline{1, N}$, if there are i customers in service and the buffers are empty; (ii) $E^-_{i-N} \otimes I_W \otimes L_N P_{N-1}(\boldsymbol{\beta}_1)$, $i = \overline{N+1, N+R}$, and $E^- \otimes I_W \otimes L_N P_{N-1}(\boldsymbol{\beta}_1)$, $i > N+R$, if all servers are busy and there are customers in the buffers, and a type-1 customer from Buffer-1 is chosen for service. Arguing in a similar way, we obtain matrices $\bar{E}^-_{i-N} \otimes I_W \otimes L_N P_{N-1}(\boldsymbol{\beta}_2)$, $i = \overline{N+1, N+R}$, and $\bar{E}^- \otimes I_W \otimes L_N P_{N-1}(\boldsymbol{\beta}_2)$, $i > N+R$, that are defined the corresponding rates if a type-2 customer from Buffer-2 goes to service.

(2) A customer leaves the buffer due to impatience. If a type-1 customer leaves the system, the corresponding rates are given by the matrices $C^{(1)}_{i-N} I^-_{i-N} \otimes I_{WT_N}$ for $i = \overline{N+1, N+R}$ and by the matrix $CI^- \otimes I_{WT_N}$ for $i > N+R$. If a type-2 customer is lost due to impatience, the related rates are given by the matrices $C^{(2)}_{i-N} \bar{I}^-_{i-N} \otimes I_{WT_N}$, $i = \overline{N+1, N+R}$, and $\bar{C}_{i-N} \otimes I_{WT_N}$, $i > N+R$. □

Having obtained the generator G of the *MC* ξ_t, it is possible to analyze this *MC*.

Theorem 2. *The MC ξ_t is ergodic for any choice of the system parameters.*

Proof. It is easy to prove that, due to the assumption that $\epsilon_i \to \infty$ when $i \to \infty$, the following limits exist:

$$Y_0 = \lim_{i\to\infty} R_i^{-1} G_{i,i-1} = I,$$

$$Y_1 = \lim_{i\to\infty} R_i^{-1} G_{i,i} + I = O,$$

$$Y_2 = \lim_{i\to\infty} R_i^{-1} G_{i,i+1} = O$$

where the matrix $R_i = -I \bigodot G_{i,i}$. Here, $\bigodot$ is the symbol of the Hadamard product of matrices, see [67]. Therefore, the *MC* ξ_t, $t \geq 0$, fits the category of asymptotically quasi-Toeplitz Markov chains (*AQTMC*), see the definition of *AQTMC* in [68].

As follows from [68], a sufficient ergodicity condition of the *MC* ξ_t can be presented in the form

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

where the row-vector $\mathbf{y}$ can be obtained as the only solution

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

Since $Y_0 = I$, $Y_1 = O$, $Y_2 = O$, one can see that the ergodicity condition turns to $1 > 0$ and is fulfilled for any system parameters. □

It follows from Theorem 2, that the following limits (stationary probabilities of the states of the *MC* ξ_t) exist:

$$\pi(i, r, \nu, m) = \lim_{t\to\infty} P\{i_t = i, r_t = r, \nu_t = \nu, \eta_t^{(1)} = \eta^{(1)}, \ldots, \eta_t^{(M)} = \eta^{(M)}\},$$

$$i \geq 0, \ r = \overline{0, \min\{\max\{0, i-N\}, R\}}, \ \nu = \overline{1, W},$$

$$\eta^{(m)} = \overline{0, \min\{i, N\}}, \ m = \overline{1, M}, \ \sum_{m=1}^{M} \eta^{(m)} = \min\{i, N\}.$$

Let us form the row vectors $\boldsymbol{\pi}_i$, $i \geq 0$, of these probabilities enumerated in the same order as the components of the *MC* ξ_t.

To obtain the stationary probability vectors $\boldsymbol{\pi}_i$, $i \geq 0$, we have to find the only solution to the following system:

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

called the equilibrium or Chapman–Kolmogorov equations.

This system has an infinite number of unknowns and equations, and the generator G does not have a level-independent structure. Thus, the problem of solving it is not easy. To solve this system, we recommend using the numerically stable algorithm presented in [69]. In contrast to the general algorithm for $AQTMC$ proposed in [68] and designed for the upper-Hessenberg structure of the generator, the algorithm presented in [69] is oriented, namely, to solving Chapman–Kolmogorov equations with the block tridiagonal structure of the generator, which has the MC ξ_t under study.

4. System Performance Characteristics

The average number of customers in the system is computed by

$$N_{sys} = \sum_{i=1}^{\infty} i\boldsymbol{\pi}_i \mathbf{e}.$$

The mean number of type-1 customers in Buffer-1 can be found as

$$N_{buf}^{(1)} = \sum_{i=N+1}^{\infty} \sum_{r=1}^{\min\{i-N,R\}} r\boldsymbol{\pi}(i,r)\mathbf{e}.$$

The mean number of type-2 customers in Buffer-2 can be found as

$$N_{buf}^{(2)} = \sum_{i=N+1}^{\infty} \sum_{r=0}^{\min\{i-N-1,R\}} (i-N-r)\boldsymbol{\pi}(i,r)\mathbf{e}.$$

The mean number of customers in the buffers is

$$N_{buf} = \sum_{i=N+1}^{\infty} (i-N)\boldsymbol{\pi}_i \mathbf{e} = N_{buf}^{(1)} + N_{buf}^{(2)}.$$

The loss probability of an arbitrary type-1 customer upon arrival due to Buffer-1 overflow can be found as

$$P_1^{ent-loss} = \frac{1}{\lambda_1} \sum_{i=N+R}^{\infty} \boldsymbol{\pi}(i,R)(D_1 \otimes I_{T_N})\mathbf{e}.$$

The loss probability of an arbitrary type-1 customer due to impatience in Buffer-1 can be found as

$$P_1^{imp-loss} = \frac{1}{\lambda_1} \sum_{i=N+1}^{\infty} \sum_{r=1}^{\min\{i-N,R\}} \alpha_r \boldsymbol{\pi}(i,r)\mathbf{e}.$$

The loss probability of an arbitrary type-2 customer due to impatience in Buffer-2 can be found as

$$P_2^{imp-loss} = \frac{1}{\lambda_2} \sum_{i=N+1}^{\infty} \sum_{r=0}^{\min\{i-N-1,R\}} \epsilon_{i-N-r} \boldsymbol{\pi}(i,r)\mathbf{e}.$$

The loss probability of an arbitrary customer due to impatience can be found as

$$P^{imp-loss} = \frac{\lambda_1 P_1^{imp-loss} + \lambda_2 P_2^{imp-loss}}{\lambda}.$$

The rate of the output flow of successfully served type-1 customers can be found as

$$\lambda_1^{out} = \sum_{i=1}^{\infty} \boldsymbol{\pi}_i (I_{(\min\{\max\{0,i-N\},R\}+1)W} \otimes L^{(1)}_{\min\{i,N\}})\mathbf{e}.$$

The rate of the output flow of successfully served type-2 customers can be found as

$$\lambda_2^{out} = \sum_{i=1}^{\infty} \boldsymbol{\pi}_i (I_{(\min\{\max\{0,i-N\},R\}+1)W} \otimes L^{(2)}_{\min\{i,N\}})\mathbf{e}.$$

Remark 1. *The matrices* $L_i^{(1)} = L_i^{(1)}(\mathbf{S}_0^{(1)})$ *and* $L_i^{(2)} = L_i^{(1)}(\mathbf{S}_0^{(2)})$, $i = \overline{1,N}$, *can be calculated using the algorithm from [55]. Their elements define the rates of transitions of the process* $\boldsymbol{\eta}_t$, $t \geq 0$, *which lead to the service completion of a type-1 and type-2 customer correspondingly on the condition that i servers are busy. The matrices* $\mathbf{S}_0^{(l)}$, $l = 1, 2$, *are given as*

$$\mathbf{S}_0^{(1)} = \begin{pmatrix} -S_1\mathbf{e} \\ \mathbf{0}^T \end{pmatrix}, \mathbf{S}_0^{(2)} = \begin{pmatrix} \mathbf{0}^T \\ -S_2\mathbf{e} \end{pmatrix}.$$

The rate of the output flow of successfully served customers can be found as

$$\lambda^{out} = \sum_{i=1}^{\infty} \boldsymbol{\pi}_i (I_{(\min\{\max\{0,i-N\},R\}+1)W} \otimes L_{\min\{i,N\}})\mathbf{e} = \lambda_1^{out} + \lambda_2^{out}.$$

The probability of an arbitrary type-1 customer loss can be found as

$$P_1^{loss} = 1 - \frac{\lambda_1^{out}}{\lambda_1} = P_1^{ent-loss} + P_1^{imp-loss}.$$

The probability of an arbitrary type-2 customer loss can be found as

$$P_2^{loss} = 1 - \frac{\lambda_2^{out}}{\lambda_2} = P_2^{imp-loss}.$$

The probability of an arbitrary customer loss can be found as

$$P^{loss} = 1 - \frac{\lambda^{out}}{\lambda} = \frac{\lambda_1 P_1^{loss} + \lambda_2 P_2^{loss}}{\lambda}.$$

The probability that the system is empty at an arbitrary moment can be found as

$$P_{idle} = \boldsymbol{\pi}_0 \mathbf{e}.$$

The probability that an arbitrary type-1 customer starts service immediately upon arrival can be found as

$$P_1^{imm} = \frac{1}{\lambda_1} \sum_{i=0}^{N-1} \boldsymbol{\pi}_i (D_1 \otimes I_{T_i})\mathbf{e}.$$

The probability that an arbitrary type-2 customer starts service immediately upon arrival can be found as

$$P_2^{imm} = \frac{1}{\lambda_2} \sum_{i=0}^{N-1} \boldsymbol{\pi}_i (D_2 \otimes I_{T_i})\mathbf{e}.$$

The probability that an arbitrary customer starts service immediately upon arrival can be found as

$$P^{imm} = \frac{1}{\lambda} \sum_{i=0}^{N-1} \boldsymbol{\pi}_i [(D_1 + D_2) \otimes I_{T_i}]\mathbf{e} = \frac{\lambda_1 P_1^{imm} + \lambda_2 P_2^{imm}}{\lambda}.$$

5. Numerical Example

In the numerical examples, we investigate the influence of the control parameters q_1, q_2, and K on the system operation. Since we can only present three-dimensional graphs, first, we examine the impact of the probabilities q_1 and q_2 on the main performance characteristics of the system under the fixed value of K. Then, we examine the influence of the probability q_1 and threshold K under the fixed value of probability q_2. Finally, we illustrate the importance of considering the system with the $MMAP$. Namely, we illustrate

that the approximation of the $MMAP$ by the superposition of two stationary Poisson arrival processes may lead to significant errors in the evaluation of the key performance measures of the system.

Experiment 1. Let us assume that the customers arrive at the system according to the $MMAP$ arrival process that is defined by the following matrices:

$$D_0 = \begin{pmatrix} -4.20885 & 0.0884212 \\ 0.0442106 & -0.4863169 \end{pmatrix},$$

$$D_1 = \begin{pmatrix} 1.90105 & 0.0707369 \\ 0.00265263 & 0.0893054 \end{pmatrix}, D_2 = \begin{pmatrix} 2.0160099 & 0.132632 \\ 0.00530527 & 0.344843 \end{pmatrix}.$$

The total rate of type-1 and type-2 customers' arrivals to the system is $\lambda = 1$. The coefficient of correlation of successive inter-arrival times in this arrival process is 0.256958, the squared coefficient of variation is 2.61753. The average rate of type-1 customers' arrivals $\lambda_1 = 0.377074$, the average rate of type-2 customers' arrivals is $\lambda_2 = 0.622926$.

The service process for type-1 customers PH_1 is defined by the vector $\mathbf{b}_1 = (0.2, 0.8)$ and the matrix $S_1 = \begin{pmatrix} -6.16045 & 0.616045 \\ 0.0123209 & -1.23209 \end{pmatrix}$. The mean service time of type-1 customers is 0.7, and the squared coefficient of variation of the service times is 1.24723.

The service process for type-2 customers PH_2 is defined by the vector $\mathbf{b}_2 = (0.4, 0.6)$ and the matrix $S_2 = \begin{pmatrix} -3.50388 & 0.175194 \\ 0.00350388 & -0.700775 \end{pmatrix}$. The mean service time of type-2 customers is 1, and the squared coefficient of variation of the service times is 1.61091.

Let us fix the control parameter $K = 7$ and suggest that the number of servers N is equal to 4 and the capacity R of Buffer-1 is 10. Let the rates of impatience of type-1 customers and type-2 customers be defined as $\alpha_r = 0.01r$, $r = \overline{1, R}$, and $\epsilon_i = 0.03i$, $i \geq 0$, respectively.

Let us vary the probability q_1 over the interval [0, 1] with step 0.05 and the probability q_2 over the interval $[q_1, 1]$ also with step 0.05.

Figure 2 presents the dependence of the probability $P_1^{ent-loss}$ of an arbitrary type-1 customer loss upon arrival due to Buffer-1 overflow on the parameters q_1 and q_2.

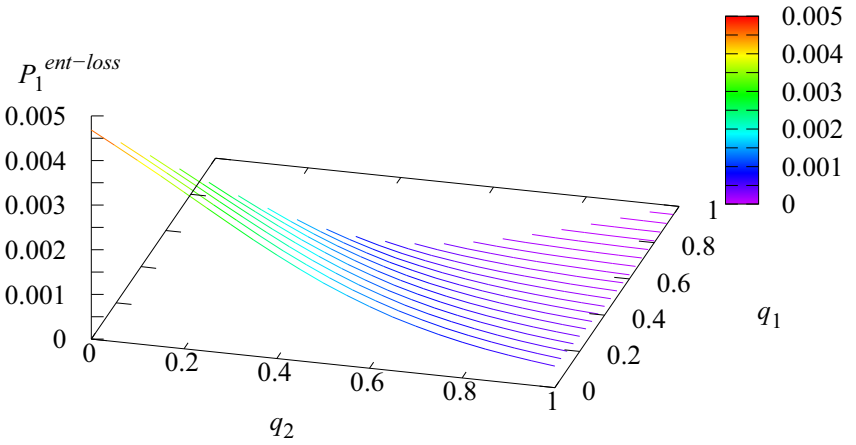

Figure 2. Dependence of the probability $P_1^{ent-loss}$ of an arbitrary type-1 customer loss upon arrival due to Buffer-1 overflow on the parameters q_1 and q_2.

One can see from Figure 2 that in the considered example, the loss probability $P_1^{ent-loss}$ decreases with growth in the probabilities q_1 and q_2. It is evident that when the probabilities q_1 and q_2 grow, type-1 customers are chosen for service more often, and the probability of meeting Buffer-1 full upon type-1 customer arrival decreases. The minimal value of the loss probability $P_1^{ent-loss}$ is achieved for $q_1 = q_2 = 1$ (what corresponds to the standard non-preemptive priority of type-1 customers) and is equal to 0.000018797.

Figure 3 illustrates the dependence of the probability $P_1^{imp-loss}$ of an arbitrary type-1 customer loss due to impatience on the parameters q_1 and q_2.

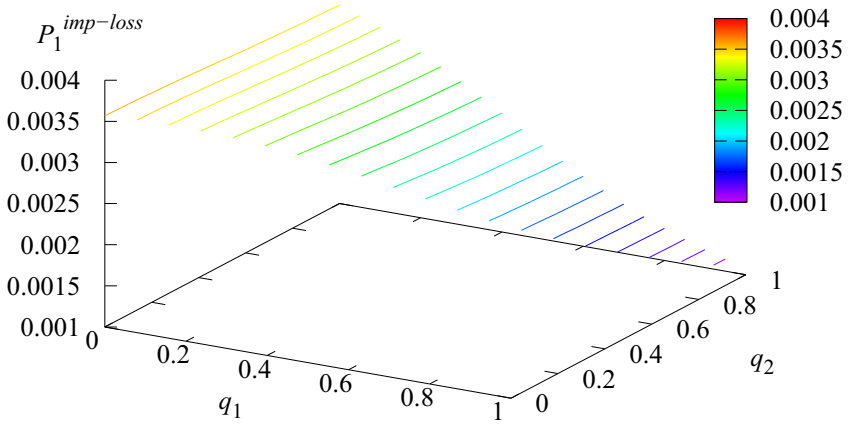

Figure 3. Dependence of the probability $P_1^{imp-loss}$ of an arbitrary type-1 customer loss due to impatience in Buffer-1 on the parameters q_1 and q_2.

One can see from Figure 3 that in the considered example, the loss probability $P_1^{imp-loss}$ sharply decreases with growth in the probability q_1 and slightly decreases with growth in the probability q_2. If the probability q_1 is small and the number of customers in Buffer-1 is less than $K = 7$, type-1 customers are rarely chosen for service. Thus, the number of type-1 customers staying in Buffer-1 is bigger for small values of q_1, and more customers leave the buffer due to impatience.

The dependence of the total probability P_1^{loss} of an arbitrary type-1 customer loss on the parameters q_1 and q_2 is presented in Figure 4.

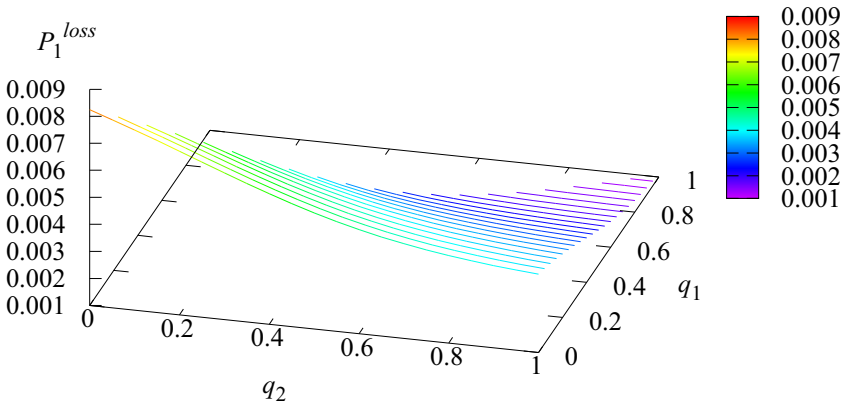

Figure 4. Dependence of the probability P_1^{loss} of an arbitrary type-1 customer loss on the parameters q_1 and q_2.

The minimal value of the loss probability P_1^{loss} is also reached for $q_1 = q_2 = 1$ and is equal to 0.0011.

Figure 5 illustrates the dependence of the probability $P_2^{imp-loss}$ of an arbitrary type-2 customer loss due to impatience on the parameters q_1 and q_2.

This probability evidently increases with growth in the probabilities q_1 and q_2 since the growth in these probabilities leads to worse conditions for type-2 customer service. More customers of this type stay in the infinite buffer, and consequently, more customers leave it due to impatience.

The dependence of the total probability P^{loss} of an arbitrary customer loss on the parameters q_1 and q_2 is presented in Figure 6.

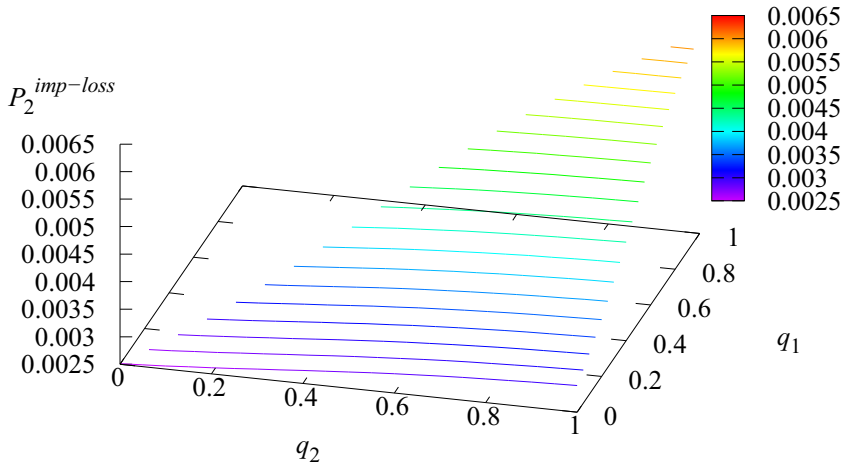

Figure 5. Dependence of the probability $P_2^{imp-loss}$ of an arbitrary type-2 customer loss due to impatience in Buffer-2 on the parameters q_1 and q_2.

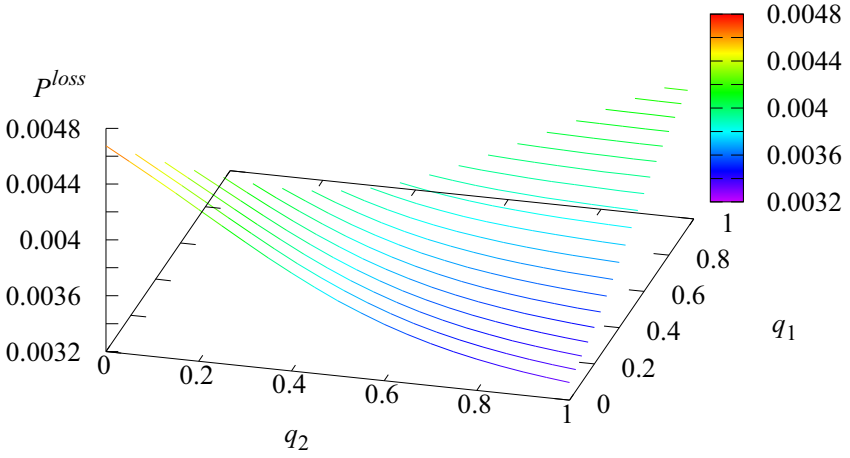

Figure 6. Dependence of the probability P^{loss} of an arbitrary customer loss on the parameters q_1 and q_2.

In the considered case, the probability P^{loss} decreases with growth in the probability q_2 and increases with growth in the probability q_1. Note that for various input parameters, the dependence of the probability P^{loss} can be absolutely different. This is because the increase in probabilities q_l, $l = 1, 2$, implies lower probabilities of type-1 customers lost upon arrival and due to impatience but a higher probability of loss of type-2 customers. In the considered case, the minimal value of the loss probability P^{loss} is reached for $q_1 = 0$ and $q_2 = 1$ and is equal to 0.003326. So, to have minimal customer loss, we should always choose type-2 customers, if any, for service if the number of type-1 customers in the buffer is less than $K = 7$, and always choose type-1 customers otherwise. However, in real-world systems, the customers may have different importance to the system; that is, the charges for losing customers of different types may essentially differ.

Let us assume that the following economic criterion describes the system's operating quality:

$$E = E(q_1, q_2) = c_1 \lambda_1 P_1^{loss} + c_2 \lambda_2 P_2^{imp-loss}$$

where c_1 is a fee paid by the system in the case when a type-1 customer is lost, and c_2 is a fee in the case when a type-2 customer from Buffer 2 is lost due to impatience.

The average losses of the system per unit of time are described by the economic criterion E. In order to optimize the system's performance, we have to determine the probabilities q_1 and q_2 for which the economic criterion E assumes the lowest value.

Let us assume the following cost coefficients: $c_1 = 18$, $c_2 = 10$.

Figure 7 presents the dependence of the cost criterion E on the parameters q_1 and q_2.

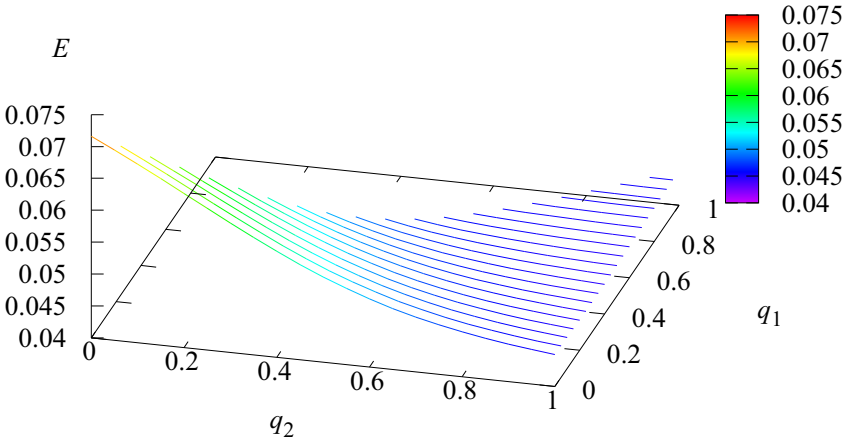

Figure 7. Dependence of the values of the cost criterion E on the parameters q_1 and q_2.

The optimal value of the economic criterion is $E(q_1, q_2) = 0.0446312$ and is achieved when $q_1 = 0.35$ and $q_2 = 1$. Thus, if there are type-1 and type-2 customers in the buffers, it is reasonable to choose the type-1 customer for the next service with a probability of 0.35 if the number of type-1 customers in Buffer-1 is less than $K = 7$, and always choose the type-1 customer otherwise.

Experiment 2. In the second experiment, we analyze the dependence of the system performance measures on the control parameters q_1 and K. To this end, let us fix the probability $q_2 = 1$, i.e., when the number of customers in the Buffer-1 exceeds the threshold K, then type-1 customers are always picked up for service. Additionally, let us increase the system load: we decrease the number of servers N to 3 and increase the capacity of Buffer-1 R from 10 to 20. The rest of the parameters are assumed to be the same as in the first experiment, including the arrival and service processes. We vary the control parameter K over the interval [2, 20] with step 1 and the probability q_1 over the interval [0, 1] with step 0.1.

Figure 8 illustrates the dependence of the probability $P_1^{ent-loss}$ of an arbitrary type-1 customer loss upon arrival due to Buffer-1 overflow on the parameters K and q_1.

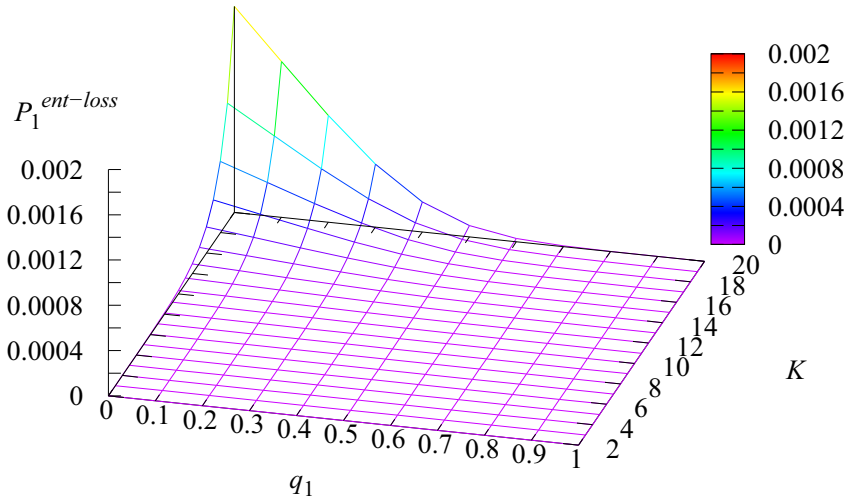

Figure 8. Dependence of the probability $P_1^{ent-loss}$ of an arbitrary type-1 customer loss upon arrival due to Buffer-1 overflow on the parameters K and q_1.

One can see from Figure 8 that the probability $P_1^{ent-loss}$ decreases with an increase in the probability q_1 and increases with a growth in the parameter K. The maximal value of the probability $P_1^{ent-loss}$ is equal to 0.0018176 and is reached for $q_1 = 0$ and $K = 20$, i.e., when there are type-2 customers in the buffer, type-1 customers are chosen for service only if Buffer-1 is full. The minimal value of this probability is equal to 8.43×10^{-6} and is

reached for $q_1 = q_2 = 1$. Note that in the case of $q_1 = q_2$, the parameter K does not impact the system's operation, and the system's performance measures do not depend on K.

The dependence of the probability $P_1^{imp-loss}$ of an arbitrary type-1 customer loss due to impatience in Buffer-1 on the parameters K and q_1 is presented in Figure 9.

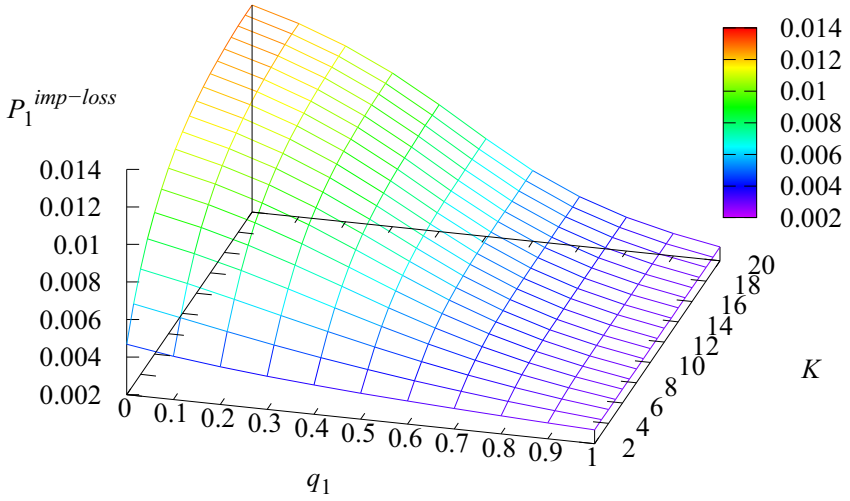

Figure 9. Dependence of the probability $P_1^{imp-loss}$ of an arbitrary type-1 customer loss due to impatience in Buffer-1 on the parameters K and q_1.

The probability $P_1^{imp-loss}$ also increases with the decrease of the probability q_1 and increases with growth in the parameter K. The same behavior we can also observe in Figure 10 that illustrates the dependence of the probability P_1^{loss} of an arbitrary type-1 customer loss on the parameters K and q_1.

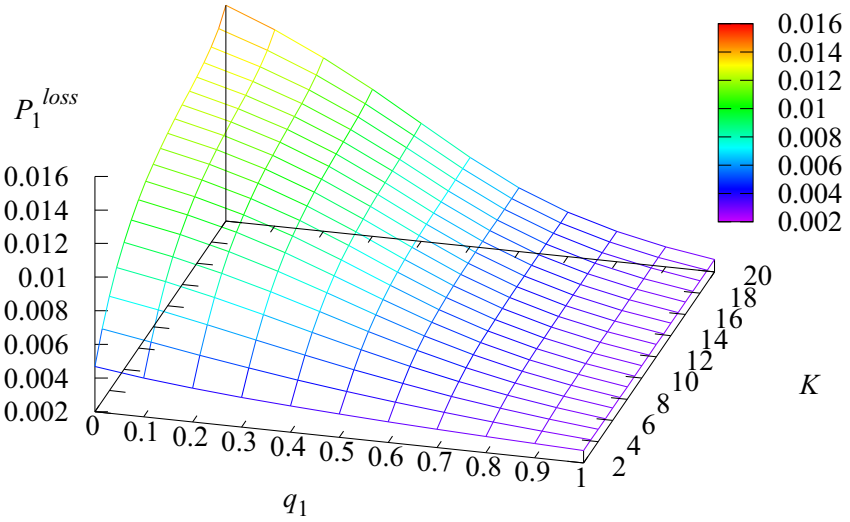

Figure 10. Dependence of the probability P_1^{loss} of an arbitrary type-1 customer loss on the parameters K and q_1.

The loss probability P_1^{loss} takes its minimal value 0.002738 for $q_1 = 1$. This loss mainly occurs due to type-1 customers' impatience.

The probability $P_2^{imp-loss}$ of an arbitrary type-2 customer loss due to impatience in Buffer-2 behaves oppositely. This probability grows with an increase in the probability q_1 and a decrease in the parameter K. The shape of the dependence of the probability $P_2^{imp-loss}$ on the parameters K and q_1 is presented in Figure 11.

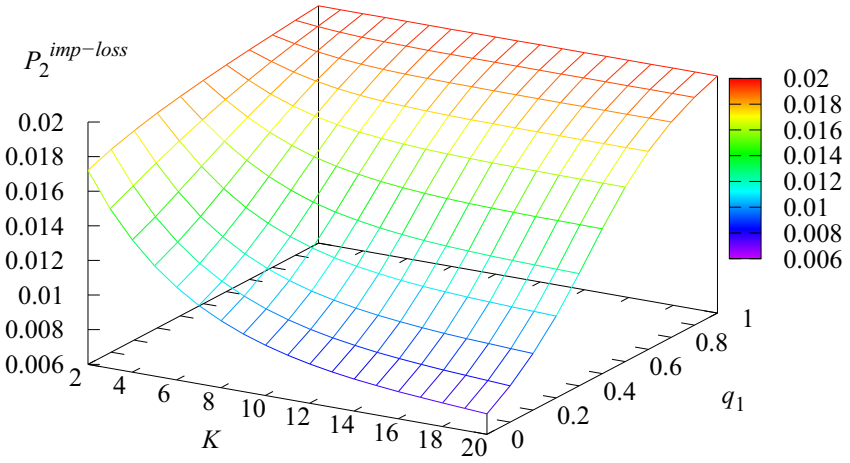

Figure 11. Dependence of the probability $P_2^{imp-loss}$ of an arbitrary type-2 customer loss due to impatience in Buffer-2 on the parameters K and q_1.

The dependence of the probability P_{loss} of an arbitrary customer loss on the values K and q_1 is presented in Figure 12.

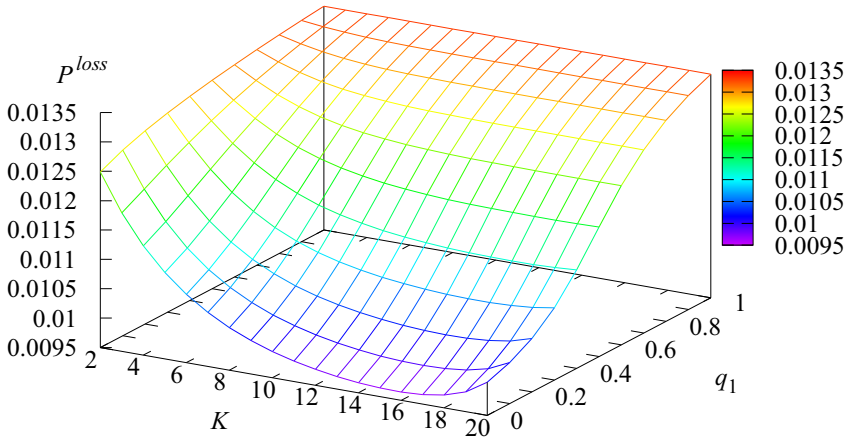

Figure 12. Dependence of the probability P_{loss} of an arbitrary customer loss on the parameters K and q_1.

The minimal value of the probability P_{loss} is equal to 0.009641 for $K = 16$ and $q_1 = 0$.

The quality of system operation can be described by the economic criterion $E = E(q_1, K)$, which is given in the same way as the criterion from the first experiment. The coefficients c_1 and c_2 also coincide with the cost coefficients from the previous experiment. The dependence of the cost criterion $E = E(q_1, K)$ on the parameters q_1 and K is presented in Figure 13.

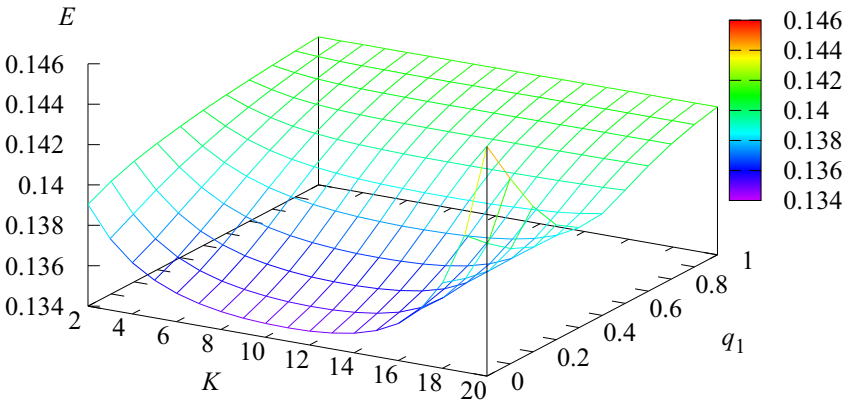

Figure 13. Dependence of the values of the cost criterion E on the parameters K and q_1.

The economic criterion takes the optimal value 0.134653 for $K = 11$ and $q_1 = 0$.

Experiment 3. In the third experiment, instead of a correlated $MMAP$ arrival process, we consider two stationary Poisson arrival processes and will show that accounting for only the mean arrival rates and ignoring the correlation and variation of the inter-arrival times can lead to significant errors in assessing the performance of the system under consideration.

Consider the $MMAP$ arrival process characterized by the value $W = 1$ with the following matrices: $D_0 = (-1)$, $D_1 = (0.377074)$, $D_2 = (0.622926)$. It defines a heterogeneous stationary Poisson arrival process with two types of customers having arrival rates $\lambda_1 = 0.377074$ and $\lambda_2 = 0.622926$. The other parameters coincide with those presented in the second experiment. The control parameters q_1 and K are varied in the same way as in the previous experiment.

Figures 14–16 show the dependence of the probabilities P_1^{loss}, $P_2^{imp-loss}$ and the cost criterion E on the parameters K and q_1 in the case of the heterogeneous stationary Poisson arrival flow.

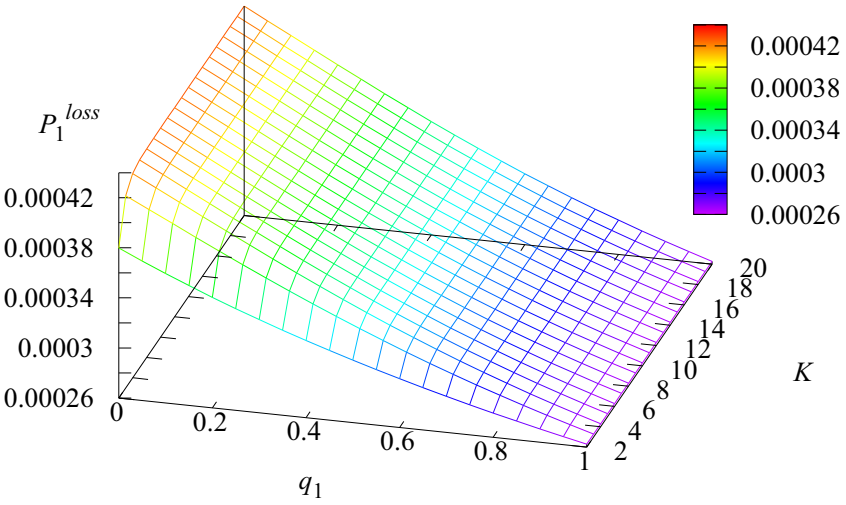

Figure 14. Dependence of the probability P_1^{loss} of an arbitrary type-1 customer loss on the parameters K and q_1 in the case of the heterogeneous stationary Poisson arrival flow.

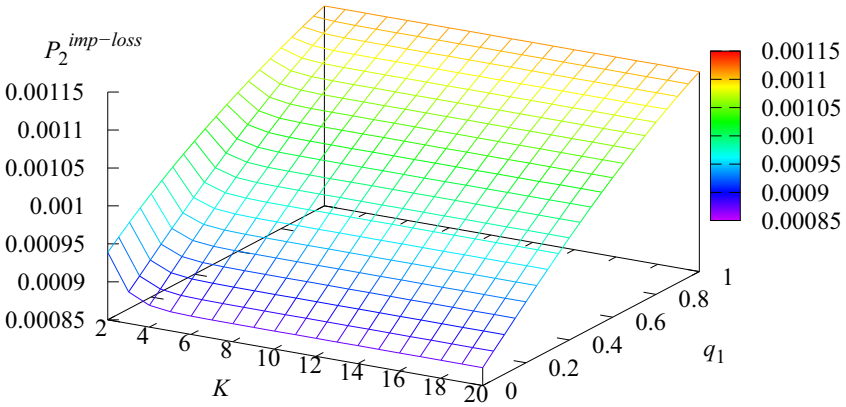

Figure 15. Dependence of the probability $P_2^{imp-loss}$ of an arbitrary type-2 customer loss due to impatience in Buffer-2 on the parameters K and q_1 in the case of the heterogeneous stationary Poisson arrival flow.

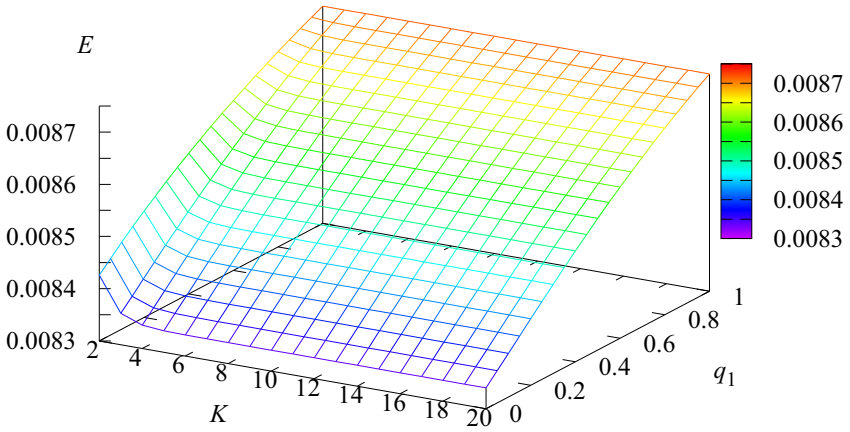

Figure 16. Dependence of the values of the cost criterion E on the parameters K and q_1 in the case of the heterogeneous stationary Poisson arrival flow.

One can draw the conclusion that the correlation and variance of the arrival process have a significant impact on the systems' performance indicators by comparing Figures 10 and 14, 11 and 15, and 13 and 16, respectively. If someone tries to evaluate the system with correlated inter-arrival times using the model with the heterogeneous stationary Poisson arrival flow, he/she can obtain huge errors in the estimation. In the case of the heterogeneous stationary Poisson arrival flow, the minimal value of the probability P_1^{loss} is equal to 0.000262, while in the case of the $MMAP$ the minimal value of this probability is 0.002738. The minimal value of the probability $P_2^{imp-loss}$ is equal to 0.00087 for heterogeneous stationary Poisson arrival flow, while in the case of the $MMAP$ the minimal value of this probability is 0.00717. That means that the real values of the main loss probabilities can be 10 times higher than the values expected based on the model with heterogeneous stationary Poisson arrival flow.

The minimal value of the cost criterion E^* in the case of the heterogeneous stationary Poisson arrival flow is equal to 0.00833988, which is 16 times less than this value for the case of the $MMAP$ arrival process.

Therefore, ignorance of positive correlation leads to an overly optimistic prediction of the performance characteristics of the system and an underestimation of the requirements for providing the desired quality of service, such as service rates or the number of servers. This, along with the observation that flows in many real-world systems exhibit correlation and high variability in inter-arrival times, clearly motivates the necessity of analyzing systems with the $MMAP$.

6. Conclusions

We analyzed the flexible randomized threshold strategy of priority access in a multi-server queueing system with the $MMAP$ arrival process and the phase-type distribution of the service times dependent on the type of customer. The impatience of customers of both types during their stay in the buffers is taken into account. Analysis of the multi-dimensional Markov chain describing the dynamics of the system is implemented via the use of the generalized PH distribution and the results known for asymptotically quasi-Toeplitz Markov chains. Numerical results are presented. They give some insight into the influence of the parameters of the control strategy. The necessity of accounting for correlation in the arrival process is shown.

Analysis can be extended to the cases of several thresholds and to the systems in which service requires the presence of some additional inventory, e.g., energy harvested online.

Author Contributions: Conceptualization, S.A.D. and A.N.D.; methodology, S.A.D., O.S.D. and A.N.D.; software, S.A.D. and O.S.D.; validation, S.A.D. and O.S.D.; formal analysis, S.A.D., O.S.D. and A.N.D.; investigation, S.A.D., O.S.D. and A.N.D.; writing, original draft preparation, O.S.D. and A.N.D.; writing, review and editing, S.A.D., O.S.D. and A.N.D.; supervision, S.A.D. and A.N.D.; project administration, A.N.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Jaiswal, N.K. *Priority Queues*; Academic Press: New York, NY, USA, 1968.
2. Takagi, H. *Queueing Analysis: A Foundation of Performance Evaluation, Volume 1: Vacation and Priority Systems*; Elsevier: Amsterdam, The Netherlands, 1991.
3. Kleinrock, L. *Queueing Systems, Volume 2: Computer Applications*; Wiley: New York, NY, USA, 1976.
4. Gnedenko, B.V.; Danielyan, E.A.; Dimitrov, B.N.; Klimov, G.P.; Matvejev, V.F. *Priority Queueing Systems*; Moscow State University: Moscow, Russia, 1973. (In Russian)
5. Bronshtein, O.I.; Dukhovnyi, I.M. *Priority Queueing Models in Information and Computing Systems*; Nauka: Moscow, Russia, 1976. (In Russian)
6. Chen, Y.; Chen, J. A Multi-Server Priority Agent Service Queueing System with Balking, Reneging and Negative Customers. *J. Syst. Sci. Complex.* **2022**, *42*, 3253–3270.
7. Ghanbari, E.; Soghrati Ghasbe, S.; Aghsami, A.; Jolai, F. A novel mathematical optimization model for a preemptive multi-priority M/M/C queueing system of emergency department's patients, a real case study in Iran. *IISE Trans. Healthc. Syst. Eng.* **2022**, *12*, 305–321. [CrossRef]
8. Nourbakhsh, V.; Turner, J. Dynamized routing policies for minimizing expected waiting time in a multi-class multi-server system. *Comput. Oper. Res.* **2022**, *137*, 105545. [CrossRef]
9. Lee, S.; Dudin, A.; Dudina, O.; Kim, C. Analysis of a priority queueing system with the enhanced fairness of servers scheduling. *J. Ambient. Intell. Humaniz. Comput.* **2022**, 1–13. [CrossRef]
10. Walraevens, J.; Van Giel, T.; De Vuyst, S.; Wittevrongel, S. Asymptotics of waiting time distributions in the accumulating priority queue. *Queueing Syst.* **2022**, *101*, 221–244. [CrossRef]
11. Walraevens, J. Asymptotics in priority queues: From finite to infinite capacities. *Queueing Syst.* **2022**, *100*, 361–363. [CrossRef]
12. Alipour-Vaezi, M.; Aghsami, A.; Jolai, F. Prioritizing and queueing the emergency departments' patients using a novel data-driven decision-making methodology, a real case study. *Expert Syst. Appl.* **2022**, *195*, 116568. [CrossRef]
13. Bai, X.; Jin, S. Performance analysis of an energy-saving strategy in cloud data centres based on a $MMAP[K]/M[K]/N_1 + N_2$ non-preemptive priority queue. *Future Gener. Comput. Syst.* **2022**, *136*, 205–220. [CrossRef]
14. Wang, Z.; Fang, L. The effect of customer awareness on priority queues. *Nav. Res. Logist.* **2022**, *69*, 801–815. [CrossRef]
15. Li, S.; Xu, Q.; Gaber, J.; Yang, N. Modeling and Performance Analysis of Channel Assembling Based on Ps-rc Strategy with Priority Queues in CRNs. *Wirel. Commun. Mob. Comput.* **2022**. [CrossRef]
16. Raj, R.; Jain, V. Optimization of traffic control in $MMAP[2]/PH[2]/S$ priority queueing model with PH retrial times and the preemptive repeat policy. *J. Ind. Manag. Optim.* **2023**, *19*, 2333–2353. [CrossRef]
17. Samouylov, K.; Dudina, O.; Dudin, A. Analysis of Multi-Server Queueing System with Flexible Priorities. *Mathematics* **2023**, *11*, 1040.
18. Rykov, V.V.; Lembert E. Optimal dynamic priorities in single-line queueing systems. *Eng. Cybern.* **1967**, *5*, 21–30.
19. Rykov, V.V. *Controllable Queueing Systems*; Itogi Nauki i Tekhniki, Teoriya Veroyatnostei, Matematicheskaya Statistika, Teoretich-eskaya Kibernetika; VINITI: Moscow, Russia, 1975; Volume 12, pp. 43–153. [CrossRef]
20. Dudin, A.; Dudin, S. Analysis of a priority queue with phase-type service and failures. *Int. J. Stoch. Anal.* **2016**, *2016*, 9152701.
21. Ponomarenko, L.; Kim, C.S.; Melikov, A. *Performance Analysis and Optimization of Multi-Traffic on Communication Networks*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2010. [CrossRef]
22. He, Q.M. Queues with marked customers. *Adv. Appl. Probab.* **1996**, 28, 567–587.
23. He, Q.-M. *Fundamentals of Matrix-Analytic Methods*; Springer: New York, NY, USA, 2014.
24. Dudin, A.N.; Klimenok, V.I.; Vishnevsky, V.M. *The Theory of Queuing Systems with Correlated Flows*; Springer Nature: Cham, Switzerland, 2020.
25. Chakravarthy, S.R. *Introduction to Matrix-Analytic Methods in Queues 1: Analytical and Simulation Approach—Basics*; ISTE Ltd., London and John Wiley and Sons: New York, NY, USA, 2022.
26. Chakravarthy, S.R. *Introduction to Matrix-Analytic Methods in Queues 2: Analytical and Simulation Approach—Queues and Simulation*; ISTE Ltd.: London, UK; John Wiley and Sons: New York, NY, USA, 2022.

27. Chakravarthy, S.R. The Batch Markovian Arrival Process: A Review and Future Work. In *Advances in Probability Theory and Stochastic Processes*; Krishnamoorthy, A., Raju, N., Ramaswami, V., Eds.; Notable Publications, Inc.: NJ, USA, 2001; pp. 21–49. [CrossRef]
28. Latouche, G.; Ramaswami, V. *Introduction to Matrix Analytic Methods in Stochastic Modeling*; SIAM: Philadelphia, PA, USA, 1999. [CrossRef]
29. Lucantoni, D.; Meier-Hellstern, K.S.; Neuts, M.F. A single-server queue with server vacations and a class of nonrenewal arrival processes. *Adv. Appl. Prob.* **1990**, *22*, 676–705.
30. Lucantoni, D. New results on the single server queue with a batch Markovian arrival process. *Stoch. Model.* **1991**, *7*, 1–46. [CrossRef]
31. Lucantoni, D.M. The $BMAP/G/1$ queue: A tutorial. In *Performance Evaluation of Computer and Communication Systems. Performance SIGMETRICS 1993*; Springer: Berlin/Heidelberg, Germany, 2005; Volume 93, pp. 330–358.
32. Neuts, M.F. A versatile Markovian point process. *J. Appl. Prob.* **1979**, *16*, 764–779.
33. Neuts, M.F. Models based on the Markovian arrival processes. *IEICE Trans. Commun.* **1992**, *75*, 1255–1265.
34. Naumov, V.; Gaidamaka, Y.; Yarkina, N.; Samouylov, K. *Matrix and Analytical Methods for Performance Analysis of Telecommunication Systems*; Springer: Berlin/Heidelberg, Germany, 2021. [CrossRef]
35. Li, Q.L. *Constructive Computation in Stochastic Models with Applications: The RG-Factorizations*; Springer Science and Business Media: Berlin/Heidelberg, Germany, 2011. [CrossRef]
36. Heffes, H.; Lucantoni, D. A Markov modulated characterization of packetized voice and data traffic and related statistical multiplexer performance. *IEEE J. Sel. Areas Commun.* **1986**, *4*, 856–868. [CrossRef]
37. Heyman, D.P.; Lucantoni, D. Modelling multiple IP traffic streams with rate limits. *IEEE /ACM Trans. Netw.* **2003**, *11*, 948–958. [CrossRef]
38. Klemm, A.; Lindermann, C.; Lohmann, M. Modelling IP traffic using the batch Markovian arrival process. *Perform. Eval.* **2003**, *54*, 149–173. [CrossRef]
39. Telek, M.; Horváth, G. A minimal representation of Markov arrival processes and a moments matching method. *Perform. Eval.* **2007**, *64*, 1153–1168. [CrossRef]
40. Kang, S.H.; Kim, Y.H.; Sung, D.K.; Choi, B.D. An application of Markovian arrival process (MAP) to modeling superposed ATM cell streams. *IEEE Trans. Commun.* **2002**, *50*, 633–642. [CrossRef]
41. Fralix, B.; Holmes, M.; Löpker, A. A Markovian arrival stream approach to stochastic gene expression in cells. *J. Math. Biol.* **2023**, *86*, 79. [CrossRef]
42. Okamura, H.; Dohi, T.; Trivedi, K.S. Markovian arrival process parameter estimation with group data. *IEEE/ACM Trans. Netw.* **2009**, *17*, 1326–1339. [CrossRef]
43. Buchholz, P.; Kemper, P.; Kriege, J. Multi-class Markovian arrival processes and their parameter fitting. *Perform. Eval.* **2010**, *67*, 1092–1106. [CrossRef]
44. Vishnevskii, V.M.; Dudin, A.N. Queueing systems with correlated arrival flows and their applications to modeling telecommunication networks. *Autom. Remote Control* **2017**, *78*, 1361–1403. [CrossRef]
45. Klimenok, V.; Dudin, A.; Vishnevsky, V. Priority multi-server queueing system with heterogeneous customers. *Mathematics* **2020**, *8*, 1501. [CrossRef]
46. Dudin, S.; Dudina, O.; Samouylov, K.; Dudin, A. Improvement of the fairness of non-preemptive priorities in the transmission of heterogeneous traffic. *Mathematics* **2020**, *8*, 929.
47. Lee, S.; Dudin, S.; Dudina, O.; Kim, C.; Klimenok, V. A priority queue with many customer types, correlated arrivals and changing priorities. *Mathematics* **2020**, *8*, 1292.
48. Neuts, M.F. *Matrix-Geometric Solutions in Stochastic Models*; The Johns Hopkins University Press: Baltimore, MD, USA, 1981. [CrossRef]
49. Asmussen, S. *Applied Probability and Queues*; Springer: New York, NY, USA, 2003; Volume 2. [CrossRef]
50. O'Cinneide, C.A. Phase-type distributions: open problems and a few properties. *Stoch. Model.* **1999**, *15*, 731–757.
51. Altiok, T. On the phase-type approximations of general distributions. *IIE Trans.* **1985**, *17*, 110–116. [CrossRef]
52. Buchholz, P.; Kriege, J.; Felko, I. *Input Modeling with Phase-Type Distributions and Markov Models: Theory and Applications*; Springer: Berlin/Heidelberg, Germany, 2014. [CrossRef]
53. Ramaswami, V. Independent Markov process in parallel. *Commun. Stat. Stoch. Models* **1985**, *1*, 419–432. [CrossRef]
54. Ramaswami, V.; Lucantoni, D. Algorithm for the multi-server queue with phase-type service, Commun. Stat. *Commun. Stat. Stoch. Models* **1985**, *1*, 393–417. [CrossRef]
55. Kim, C.; Dudin, A.; Dudin, S.; Dudina, O. Mathematical model of operation of a cell of a mobile communication network with adaptive modulation schemes and handover of mobile users. *IEEE Access* **2021**, *9*, 106933–106946. [CrossRef]
56. Kim, C.; Dudin, A.; Dudina, O.; Dudin, S. Tandem queueing system with infinite and finite intermediate buffers and generalized phase-type service time distribution. *Eur. J. Oper. Res.* **2014**, *235*, 170–179. [CrossRef]
57. Dudin, A.; Kim, C.; Dudina, O.; Dudin, S. Multi-server queueing system with a generalized phase-type service time distribution as a model of call center with a call-back option. *Ann. Oper. Res.* **2016**, *239*, 401–428. [CrossRef]
58. Swensen, A.R. Remaining loads in a $PH/M/c$ queue with impatient customers. *Methodol. Comput. Appl. Probab.* **2023**, *25*, 25. [CrossRef]

59. Liu, H.L.; Li, Q.L. Matched Queues with Flexible and Impatient Customers. *Methodol. Comput. Appl. Probab.* **2023**, *25*, 4. [CrossRef]
60. Bassamboo, A.; Randhawa, R.; Wu, C. Optimally Scheduling Heterogeneous Impatient Customers. *Manuf. Serv. Oper. Manag.* **2023**, *25*, 811–1208. [CrossRef]
61. Satin, Y.; Razumchik, R.; Kovalev, I.; Zeifman, A. Ergodicity and Related Bounds for One Particular Class of Markovian Time—Varying Queues with Heterogeneous Servers and Customer's Impatience. *Mathematics* **2023**, *11*, 1979. [CrossRef]
62. Kim, C.S.; Mushko, V.V.; Dudin, A.N. Computation of the steady state distribution for multi-server retrial queues with phase type service process. *Ann. Oper. Res.* **2012**, *201*, 307–323. [CrossRef]
63. Kim, C.; Klimenok, V.I.; Dudin, A.N. Analysis of unreliable $BMAP/PH/N$ type queue with Markovian flow of breakdowns. *Appl. Math. Comput.* **2017**, *314*, 154–172.
64. Kim, C.; Dudin, S.; Taramin, O.; Baek, J. Queueing system $MAP/PH/N/N+R$ with impatient heterogeneous customers as a model of call center. *Appl. Math. Model.* **2013**, 37, 958–976.
65. Graham, A. *Kronecker Products and Matrix Calculus: With Applications*; Horwood: Chichester, UK, 1981.
66. Steeb, W.-H.; Hardy, Y. *Matrix Calculus and Kronecker Product*; World Scientific Publishing: Singapore, 2011. [CrossRef]
67. Horn, R.A.; Johnson, C.R. *Topics in Matrix Analysis*; Cambridge University Press: Cambridge, UK, 1991. [CrossRef]
68. Klimenok, V.I.; Dudin, A.N. Multi-dimensional asymptotically quasi-Toeplitz Markov chains and their application in queueing theory. *Queueing Syst.* **2006**, *54*, 245–259.
69. Dudin, S.; Dudina, O. Retrial multi-server queuing system with PHF service time distribution as a model of a channel with unreliable transmission of information. *Appl. Math. Model.* **2019**, *65*, 676–695.