mathematics

MDPI

Article

Queuing System with Two Types of Customers and Dynamic Change of a Priority

Valentina Klimenok [1], Alexander Dudin [1,2,*], Olga Dudina [1] and Irina Kochetkova [2]

[1] Department of Applied Mathematics and Computer Science, Belarusian State University, 4 Nezavisimosti Ave., 220030 Minsk, Belarus; vklimenok@yandex.ru (V.K.); dudina@bsu.by (O.D.)

[2] Applied Mathematics and Communications Technology Institute, Peoples' Friendship University of Russia (RUDN University), 6 Miklukho-Maklaya St, Moscow 117198, Russia; igudkova@sci.pfu.edu.ru

* Correspondence: dudin@bsu.by

Received: 15 April 2020; Accepted: 16 May 2020; Published: 19 May 2020

Abstract: The use of priorities allows us to improve the quality of service of inhomogeneous customers in telecommunication networks, inventory and health-care systems. An important modern direction of research is to analyze systems in which priority of a customer can be changed during his/her stay in the system. We considered a single-server queuing system with a finite buffer, where two types of customers arrive according to a batch marked Markov arrival process. Type 1 customers have non-preemptive priority over type 2 customers. Low priority customers are able to receive high priority after the random amount of time. For each non-priority customer accepted into the buffer, a timer, which counts a random time having a phase type distribution, is switched-on. When the timer expires, the customer with some probability leaves the system unserved and with the complimentary probability gains the high priority. Such a type of queues is typical in many health-care systems, contact centers, perishable inventory, etc. We describe the behavior of the system by a multi-dimensional continuous-time Markov chain and calculate a number of the stationary performance measures of the system including the various loss probabilities as well as the distribution function of the waiting time of priority customers. The illustrative numerical examples giving insights into the system behavior are presented.

Keywords: changing priority queue; batch marked Markov arrival process; phase-type time distribution; waiting time

1. Introduction

Queuing theory is very useful for solving the problems of optimal sharing and scheduling restricted resources in many real world systems. The important branch of this theory is devoted to the consideration of queuing systems that are designed to provide service to heterogeneous flows of customers. These flows may have different value for the system and, therefore, more important flows may deserve a special treatment. This led to the idea to introduce certain priority classes and provide a privilege to the customers from the class with higher priority. Priorities are divided into preemptive and non-preemptive. For example, if the priority defines the choice of the type of the customer that will receive service, preemptive priority suggests immediate interruption of service of a customer if the customer from higher priority class arrives. Non-preemptive priority works only at service completion moments. Next service is provided to the customer from the queue which has the highest priority among waiting for service. Another kind of the priorities classification distinguishes the static and dynamic priorities. Static priorities are strictly fixed and decision making is defined only by these priorities without account of the current lengths of queues of different types of customers. Dynamic priorities take these lengths into account. Generally speaking, the dynamic

priorities are more effective than the static ones. However, the field of their application is narrower because sometimes the queue lengths are not completely observable and management of control is expensive. Therefore, the static priorities are still popular in many real world systems.

The main disadvantage of the classical static priorities is their inflexibility and possible unfairness with respect to low-priority customers. If congestion occurs, the non-priority customers have very small chance to receive service within a reasonable amount of time while the priority customers are serviced quickly. To overcome this disadvantage, various improvements of the static priorities can be offered, e.g., restriction of too fast access of priority customers or mandatory service of a non-priority customer after service in turn of some fixed number of the priority customers. Another popular improvement consists of the possibility of increasing the priority of the customer during his/her staying in the queue. There are works where the customer accumulates the priority from some initial value, which depends on the priority of the customer, according to some linear or non-linear function (again, depending on the priority of the customer) during the stay of the customer in the system. Among these works we can mention papers [1–5]. Another bunch of the works assume that the increase of the priority occurs not deterministically according to some functional dependence, but randomly, after certain random amount of time. Because our paper makes the analogous assumption, to clarify the novelty of our model and results, we overview the related results in Table 1.

Table 1. Results for queues with changing priority.

Paper	Number of Priority Classes	Arrival Flow	Service Time Distribution	Distribution of Time Till Priority Change	Main Result
[6]	N	M	M	M	Ergodicity condition
[7]	N	M MAP for high priority	M	M	Bounds and tails asymptotics
[8,9]	N	$MMAP$	PH	Cox	Ergodicity condition
[10]	2	M	M	M	Optimization via MDP
[11]	2	M	M	M	Asymptotic distribution of queue
[12]	2	M	M	M	Optimization via MDP
Our paper	2	$BMMAP$	PH	PH	Distribution of queue length and waiting time

In this table, the standard Kendall's denotations are used. In particular, the symbol M denotes the exponential distribution of the corresponding variables (inter-arrival and service times as well as the time until the increase of the priority), the symbol PH denotes the phase type distribution.

It is clear from this table that our paper surpasses all cited papers, except [8,9], where N priority classes are considered, whereas we considered only two classes; however, we allow batch arrivals of customers and obtain not just ergodicity condition (as it is done in [8,9]) but exactly compute the stationary distribution of the system states and waiting time for priority customers. No other paper from this table gives the exact formulas for these distributions. It is worth to note also that, except [8,9], all papers assume the exponential distribution of inter-arrival and service times as well as the time until the increase of the priority. We assume a much more general phase type (PH) distributions and the batch marked Markov arrival process ($BMMAP$). This is very important for potential real-world applications. One of popular applications of such kind of models is in the field of health-care. This application is mentioned in practically all cited above papers. For example, we considered queuing system suits for description of operation of the emergency department in a hospital, including an operation room and a team of necessary doctors and nurses. Patients, which suffered in an accident, are delivered to this hospital. There, they are subjected to triage.

This triage is performed by a physician called sorting medic, which decides whether the patient needs very urgent treatment (and, consequently, receives high priority) or can wait for a while (receives low priority). During the waiting time, the state of the low priority patient can become worse and he/she will need urgent treatment (becomes high priority customer). It is worth noting that in the recent paper [13] devoted to analysis of patients managing in emergency departments, the following is stated: "disciplines with changing priorities have been studied in the literature from a queuing theory point of view, which requires assumptions rarely found in real emergency departments, such as homogeneity in the patient arrival pattern and only one service stage." Assumptions about the $BMMAP$ arrival process and phase type service (probably consisting of an arbitrary finite number of sequentially stages) made in our paper a better fit for modeling real emergency departments. The phase type distribution is also much better than the exponential one, as it is suited for the description of the time until the patient will leave the hospital without help (e.g., transfers to another hospital or dies) or become a priority customer. The mode of the exponential distribution is equal to zero, which is hardly true for waiting of service. The model considered in this paper can also be used for description of operation of a contact center. As it is mentioned in the literature, phone calls have high priority, and requests sent by e-mail or a messenger have low priority; however, the customer who used a messenger for receiving information can make a phone call if his/her waiting for a response is too long. It is clarified in [14] that exponential assumptions may be inadequate and more general distributions or flows should be exploited for modeling a contact center, as it is done in our paper.

Our paper has the following structure. In Section 2, the queuing model we studied is described. The operation of the system is described by a multi-dimensional continuous-time Markov chain in Section 3. The infinitesimal generator of this Markov chain is presented there along with the brief proof of the probabilistic meaning of its blocks. The main difficulties in derivation of the expressions for these blocks and computer realization of their computations are caused by the assumption that the times until the change of a priority are not exponential, but a more complicated phase type distribution. To drastically reduce the state space of the multi-dimensional process describing the simultaneous behavior of underlying processes of such a distribution for all non-priority customers, we used and extended the known trick from [15,16]. In Section 4, the problem of computation of the stationary distribution of the states of the constructed Markov chain is briefly touched on and expressions for computation of the most important performance indicators of the system are derived, including expressions for the loss probabilities of an arbitrary customer, high priority customer, low priority customer and customer departed from the buffer without obtaining service. Expressions for computation of distribution functions of the waiting time of an arbitrary priority customer that is accepted into the system and of an arbitrary non-priority customer who becomes the priority customer are presented. These expressions allow us to exactly compute the probability that the waiting time will exceed an arbitrary given value. In turn, having the ability to compute this probability, many managerial problems can be solved, e.g., the choice of the satisfactory service rate (which is predefined by the used equipment and the staff of the team of surgeons, anesthetists, nurses, etc), necessary capacity of the buffer and strategy of rescheduling the patients to other hospitals, etc. Section 6 contains the results of the numerical examples which highlight the effects of variation of arrival rate and buffer capacity as well as correlation in the arrival process.

2. Model Description

We considered a single-server system with a buffer of capacity N and non-homogeneous input flow.

The structure of the system is presented in Figure 1.

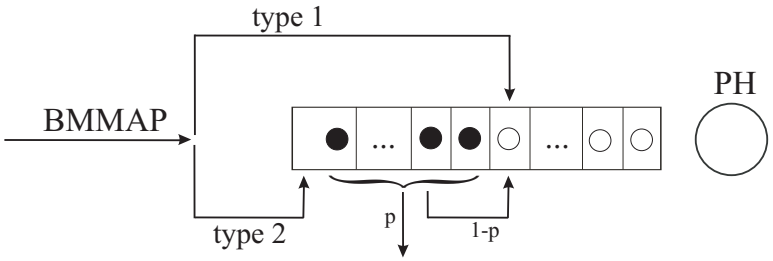

Figure 1. Structure of the system.

Customers of two types arrive to the system according to a $BMMAP$. The brief description of this $BMMAP$ is as follows. Arrivals of the batches of customers occur under the control of an irreducible continuous-time Markov chain ν_t, $t \geq 0$, with finite state space $\{0, \ldots, W\}$. This chain is called as an underlying process of the $BMMAP$. The chain ν_t stays in the state ν during an exponentially distributed time with the parameter $\lambda_\nu, \nu = 0, \ldots, W$. After this time is over, with probability $p_0(\nu, \nu')$ the chain transits into the state ν', $\nu' \neq \nu$, without generating customers or with probability $p_k(\nu, \nu')$ the chain transits into the state ν' and the batch of k customers of lth type is generated, $k \geq 1$, $l = 1, 2$. It is assumed that

$$p_0(\nu, \nu) = 0, \sum_{l=1}^{2} \sum_{k=1}^{\infty} \sum_{\nu'=0}^{W} p_k^{(l)}(\nu, \nu') + \sum_{\nu'=0}^{W} p_0(\nu, \nu') = 1, \nu = 0, \ldots, W.$$

Parameters characterizing the $BMMAP$ are stored in the square matrices D_0, $D_k^{(l)}$, $k \geq 1$, $l = 1, 2$, of size $\bar{W} = W + 1$. These matrices are defined by their entries as follows:

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

$$(D_k^{(l)})_{\nu,\nu'} = \lambda_\nu p_k^{(l)}(\nu, \nu'), \ \nu, \nu' = 0, \ldots, W, \ k \geq 1, \ l = 1, 2.$$

Denote

$$D(1) = D_0 + \sum_{l=1}^{2} \sum_{k=1}^{\infty} D_k^{(l)}, \ \mathcal{D}_1^{(l)} = \sum_{k=1}^{\infty} \sum_{\bar{l}=1, \bar{l} \neq l}^{2} D_k^{(\bar{l})}, \ l = 1, 2.$$

Let $\boldsymbol{\theta}$ be the vector of the stationary distribution of the chain ν_t, $t \geq 0$. The vector $\boldsymbol{\theta}$ is calculated as the unique solution of the system of the linear algebraic equations $\boldsymbol{\theta} D(1) = \mathbf{0}, \boldsymbol{\theta}\mathbf{e} = 1$. Here $\mathbf{e}$ is a column vector consisting of 1's and $\mathbf{0}$ is a row vector consisting of 0's.

The arrival rate of customers of lth type is calculated by the formula

$$\lambda_l = \boldsymbol{\theta} \sum_{k=1}^{\infty} k D_k^{(l)} \mathbf{e}, l = 1, 2.$$

The total rate of customers in the $BMMAP$ is equal to $\lambda = \sum_{l=1}^{2} \lambda_l$. The arrival rate of batches of customers of lth type is calculated by the formula $\lambda_l^{(b)} = \boldsymbol{\theta} \sum_{k=1}^{\infty} D_k^{(l)} \mathbf{e}$, $l = 1, 2$.

The variance of the lengths of the intervals between the moments of arrival of batches of the lth type customers are calculated by

$$v_l = \frac{2\boldsymbol{\theta}(-D_0 - \sum_{k=1}^{\infty} D_k^{(l)})^{-1}\mathbf{e}}{\lambda_l^{(b)}} - \left(\frac{1}{\lambda_l^{(b)}}\right)^2.$$

The coefficient of variation of the lengths of the intervals between the arrival moments of batches of customers of the lth type is calculated by $c_{var}^{(l)} = \lambda_l^{(b)} \sqrt{v_l}$. The coefficient correlation of the lengths of two adjacent intervals between the arrival moments of batches of customers of the lth type is calculated by

$$c_{cor}^{(l)} = \left[\frac{\boldsymbol{\theta}(-D_0 - \mathcal{D}_1^{(l)})^{-1}}{\lambda_l^{(b)}} \sum_{k=1}^{\infty} D_k^{(l)} (-D_0 - \mathcal{D}_1^{(l)})^{-1} \mathbf{e} - \left(\frac{1}{\lambda_l^{(b)}} \right)^2 \right] v_l^{-1}.$$

A more detailed description of a $BMMAP$ and the batch Markov arrival process can be found, for example, in [17,18].

In our model, the type of a customer defines his/her priority. Type 1 customers have non-preemptive priority over type 2 customers. A batch of priority customers, which enters the system and meets the corresponding number of free places in the buffer, is queued in the buffer after the last priority customer staying in the buffer and ahead of all non-priority customers, if any. If the batch size is greater than the number of free places, then the part of the batch is accepted and the rest leaves the system forever (is lost). This means that we consider the so called partial admission discipline. A similar acceptance discipline is valid for an arriving batch of non-priority customers, only these customers are placed at the end of the common queue if there are free places in the buffer.

For every non-priority customer accepted into the buffer, a timer is switched-on. The timer counts the random time having PH distribution defined by an irreducible representation (γ, Γ). The time until the timer expiration (switching-off) is defined by the time until absorption of the continuous-time irreducible Markov chain r_t, $t \geq 0$, with the state space $\{1, \ldots, R, R+1\}$, where the state $R+1$ is an absorbing one. The transition rates of the chain within the set of transient states $\{1, \ldots, R\}$ are defined by the entries of the sub-generator Γ and the rates of transition into the absorbing state $R+1$ are defined by the entries of the vector $\boldsymbol{\Gamma}_0 = -\Gamma \mathbf{e}$. At the moment of switching the timer on, the state of the process r_t is randomly selected among the transient states according to the probability row vector γ. The average time until switching-off the timer is calculated by $t_1 = \gamma \Gamma^{-1} \mathbf{e}$. We denote $\tau = -t_1^{-1}$.

When the timer, which was switched-on for some non-priority customer, expires (switches-off), i.e., the corresponding Markov chain r_t reaches the absorbing state $R+1$, this customer leaves the system unserved with the probability p or gains the higher priority and moves at the end of the priority customers queue with the complimentary probability $1-p$. The priority customers are picked up from the buffer according to first in–first out discipline. If at the service completion epoch the priority customers are absent in the buffer, an arbitrary non-priority customer among those having the maximal value of the corresponding process r_t is picked up for service. Such an assumption is quite realistic when the time counted by the timer has Erlangian distribution or generalized Erlangian distribution having various rates of the phases. In this case, in average, the maximal value of the underlying process r_t have customers having longer stay in the buffer. This is why namely such a customer (or one of such customers if several customers have the maximal value) will be picked up for service.

The service time of any customer has a PH distribution with an irreducible representation $(\boldsymbol{\beta}, S)$ and underlying process m_t, $t \geq 0$. This process has the state space $\{1, \ldots, M, M+1\}$, where the state $M+1$ is an absorbing one. The transition rates of the chain m_t, $t \geq 0$, within the set of transient states $\{1, \ldots, M\}$ are defined by the entries of the sub-generator S. The transition rates into the absorbing state (what implies service completion) are defined by the entries of the vector $\boldsymbol{S}_0 = -S\mathbf{e}$. The average service rate is calculated by $\mu = -(\boldsymbol{\beta} S^{-1} \mathbf{e})^{-1}$, the average service time is given by $b_1 = \mu^{-1}$.

It is worth noting that if we assume that only priority customers arrive to the system then the system under consideration transforms to the $BMAP/PH/1/N$ type queuing system. In this case, the matrices $D_k^{(2)}$ $k \geq 1$, describing the input flow are equal to zero. The system $BMAP/PH/1/N$ is a very special case of the system $BMAP/SM/1/N$ with semi-Markovian service process that was investigated in [19] for partial admission strategy and in [20] for complete admission and complete rejection strategies.

3. Process of the System States

Let at the moment t

- i_t be the number of customers in the buffer, $i_t = 0, \dots, N$;
- j_t be the number of non-priority customers in the buffer, $j_t = 0, \dots, i_t$;
- $\chi_t = 0$, if the server is idle; $\chi_t = 1$, if the server is busy;
- m_t be the state of the underlying process of the PH service on the busy server, $m_t = 1, \dots, M$;
- $n_t^{(r)}$ be the number of non-priority customers staying in the buffer for which the timer is in the state r, $n^{(r)} = 1, \dots, j_t, r = 1, \dots, R$, $\sum\limits_{r=1}^{j_t} n^{(r)} = j_t$;
- ν_t be the state of underlying process of the $BMMAP$, $\nu_t = 0, \dots, W$.

The process of operation of the system is described by a regular irreducible continuous-time Markov chain ξ_t, $t \geq 0$, with state space

$$\Omega = \{(i, \chi, \nu),\ i = 0,\ \chi = 0,\ \nu = 0, \dots, W\} \bigcup$$

$$\{(i, \chi, \nu, m),\ i = 0,\ \chi = 1,\ \nu = 0, \dots, W,\ m = 1, \dots, M\} \bigcup$$

$$\{(i, j, \nu, m, n^{(1)}, \dots, n^{(R)}),\ i = 1, \dots, N, j = 0, \dots, i,\ \nu = 0, \dots, W,\ m = 1, \dots, M,\ n^{(r)} = 1, \dots, j,\ r = 1, \dots, R\}.$$

Note that the number of states in the space Ω with the value $i = 0$ of the first component is

$$K_0 = \bar{W}(M + 1),$$

and the number of states with the value $i = 1, \dots, N$ of the first component is

$$K_i = \bar{W}M \sum_{j=0}^{i} C_{j+R-1}^{R-1},\ i = 1, \dots, N,$$

where

$$C_n^m = \binom{n}{m} = \frac{n!}{m!(n-m)!},\ n \geq 1, m = 0, \dots, n.$$

Then the cardinality of the state space Ω is equal to

$$|\Omega| = \bar{W}(M + 1) + \bar{W}M \sum_{i=1}^{N} \sum_{j=0}^{i} C_{j+R-1}^{R-1}.$$

We suppose that the states of the chain ξ_t, $t \geq 0$, are enumerated in the direct lexicographic order of the components $i_t, j_t, \chi_t, \nu_t, m_t$, and in the reverse lexicographic order of the components $n_t^{(1)}, \dots, n_t^{(R)}$.

In the sequel, we use the following notation:

I (O) is an identity (zero) matrix of an appropriate size. When it is necessary, we identify the size of a matrix with a suffix;

$diag\{a_l,\ l = 1, \dots, L\}$ is a diagonal matrix with the diagonal entries a_l;

$diag^{-}\{A_l, l = 0, \dots, L\}$ is a sub-diagonal matrix with the sub-diagonal blocks A_l;

$\otimes$ and $\oplus$ are the symbols of the Kronecker product and sum of matrices, see, e.g., [21], and $\delta_{m,n}$ is Kronecker's symbol;

$a = \bar{W}M$.

Before proceeding with the construction of the generator of the Markov chain ξ_t, $t \geq 0$, we will consider the important problem of finding the transition rates of the process $\mathbf{n}_t = (n_t^{(1)}, n_t^{(2)}, \ldots, n_t^{(R)})$ where $n_t^{(r)}$ defines the number of timers having the state r of the underlying process r_t, $r = 1, \ldots, R$. To calculate these transition rates, we use the results for R independent Markov processes in parallel, obtained by V. Ramaswami and D. Lucantoni, see [15,16]. To use these results for our model, we introduce the matrix $\hat{\Gamma} = \begin{pmatrix} \mathbf{0} & O \\ \mathbf{\Gamma}_0 & \Gamma \end{pmatrix}$ and the matrices $P_j(\gamma)$, $A_j(l, \Gamma)$, $L_k(l, \hat{\Gamma})$ which describes the transition probabilities or rates of the process $\mathbf{n}_t$.

- The matrix $P_j(\gamma)$ contains the transition probabilities of the process $\mathbf{n}_t$ at the moments of an increase in the number of non-priority customers in the buffer from j to $j+1$.
- The matrix $P_{i,j}(\gamma) = P_i(\gamma)P_{i+1}(\gamma)\ldots P_{j-1}(\gamma)$ contains the probabilities of transition of the process $\mathbf{n}_t$ at the moments of an increase in the number of non-priority customers in the buffer from i to j.
- The matrix $A_j(l, \Gamma)$ contains the transition rates of the process $\mathbf{n}_t$ in the state space of this process without increasing or decreasing the number of non-priority customers in the buffer (here j is the number of non-priority customers in the buffer, l is the total number of free places and non-priority customers in the buffer).
- The matrix $L_k(l, \hat{\Gamma})$ contains the transition rates of the process $\mathbf{n}_t$, which leads to the expiration of the timer on one of the $l-k$ non-priority customers in the buffer (here k is the number of free places in the buffer, l is the total number of free places and non-priority customers in the buffer).

In the following, we assume that $A_0(\cdot, \cdot) = L_*(*, \cdot) = P_{-1}(\cdot) = O$.

A more detailed description of the matrices $P_j(\gamma)$, $A_j(l, \Gamma)$, $L_k(l, \hat{\Gamma})$ and algorithms for their calculation, can be found in [15,22].

Besides the matrices $P_j(\gamma)$, $A_j(l, \Gamma)$, $L_k(l, \hat{\Gamma})$ we need to use the matrix $E_{j,j-1}$, which defines the transition probabilities of the process $\mathbf{n}_t$ when the current service is completed and there are no priority customers in the buffer (in the following we denote this service completion epoch as t^*) and one of j non-priority customers from the buffer moves to the server. The server is occupied by one of the non-priority customers whose timer is in the maximal (among all timers) phase r_{max} and the number of non-priority customers in the buffer becomes equal to $j-1$. Let the rows of the matrix $E_{j,j-1}$ correspond to the possible states of the process $\mathbf{n}_t$ at the moment $t^* - 0$ and the columns correspond to the states of the process $\mathbf{n}_t$ at the moment $t^* + 0$. We assume that in both cases the states are ordered in the reverse lexicographic order. Consider the row of the matrix $E_{j,j-1}$ corresponding the state $(n^{(1)}, n^{(2)}, \ldots, n^{(r_{max})}, 0, \ldots, 0)$ of the process $\mathbf{n}_t$. If at the moment $t^* - 0$ the process $\mathbf{n}_t$ is in such a state, then at the moment $t^* - 0$ it will transit to the state $(n^{(1)}, n^{(2)}, \ldots, n^{(r_{max})} - 1, 0, \ldots, 0)$. This transition occurs with probability 1. To fix this transition, we set the entry $(E_{j,j-1})_{(n^{(1)},n^{(2)},\ldots,n^{(r_{max})},0,\ldots,0),\,(n^{(1)},n^{(2)},\ldots,n^{(r_{max})}-1,0,\ldots,0)}$ of the matrix $E_{j,j-1}$ be equal to 1. The remaining entries of the row are set equal to zero. Using such an algorithm for forming the rows, we get the matrix $E_{j,j-1}$ of size $C_{j+R-1}^{R-1} \times C_{j+R-2}^{R-1}$.

Let $Q_{i,l}$ be the matrix of transition rates of the chain ξ_t, $t \geq 0$, from the set of states corresponding to the value $i_t = i$ to the states corresponding to the value $i_t = l$. Then the infinitesimal generator Q of the chain is formed as $Q = (Q_{i,l})_{i,l=0,\ldots,N}$. The following statement is true.

Lemma 1. *The infinitesimal generator Q of the Markov chain ξ_t, $t \geq 0$, has the following block structure:*

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

where

$$Q_{0,0} = \begin{pmatrix} D_0 & (D_1^{(1)} + D_1^{(2)}) \otimes \boldsymbol{\beta} \\ I_{\bar{W}} \otimes \boldsymbol{S}_0 & D_0 \oplus S \end{pmatrix},$$

$$Q_{0,k} = \begin{pmatrix} D_{k+1}^{(1)} \otimes \boldsymbol{\beta} & O_{\bar{W} \times a \sum_{j=1}^{k-1} C_{j+R-1}^{R-1}} & D_{k+1}^{(2)} \otimes \boldsymbol{\beta} \otimes P_{0,k}(\boldsymbol{\gamma}) \\ D_k^{(1)} \otimes I_M & O_{a \times a \sum_{j=1}^{k-1} C_{j+R-1}^{R-1}} & D_k^{(2)} \otimes I_M \otimes P_{0,k}(\boldsymbol{\gamma}) \end{pmatrix}, k = 1, \ldots, N-1,$$

$$Q_{0,N} = \begin{pmatrix} \sum_{k=N+1}^{\infty} D_k^{(1)} \otimes \boldsymbol{\beta} & O_{\bar{W} \times a \sum_{j=1}^{N-1} C_{j+R-1}^{R-1}} & \sum_{k=N+1}^{\infty} D_k^{(2)} \otimes \boldsymbol{\beta} \otimes P_{0,N}(\boldsymbol{\gamma}) \\ \sum_{k=N}^{\infty} D_k^{(1)} \otimes I_M & O_{a \times a \sum_{j=1}^{N-1} C_{j+R-1}^{R-1}} & \sum_{k=N}^{\infty} D_k^{(2)} \otimes I_M \otimes P_{0,N}(\boldsymbol{\gamma}) \end{pmatrix},$$

$$Q_{1,0} = \begin{pmatrix} O_{a \times \bar{W}} & I_{\bar{W}} \otimes \boldsymbol{S}_0 \boldsymbol{\beta} \\ O_{aR \times \bar{W}} & I_{\bar{W}} \otimes \boldsymbol{S}_0 \boldsymbol{\beta} \otimes \mathbf{e}_R + pI_a \otimes \boldsymbol{\Gamma}_0 \end{pmatrix},$$

$$Q_{i,i-1} =$$

$$\begin{pmatrix} diag\{I_{\bar{W}} \otimes \boldsymbol{S}_0 \boldsymbol{\beta} \otimes I_{C_{j+R-1}^{R-1}}, j = 0, \ldots, i-1\} & +diag^{-}\{pI_a \otimes L_{N-i}(N-i+j, \hat{\Gamma}), j = 1, \ldots, i-1\} \\ O_{aC_{i+R-1}^{R-1} \times a \sum_{j=0}^{i-2} C_{j+R-1}^{R-1}} & I_{\bar{W}} \otimes \boldsymbol{S}_0 \boldsymbol{\beta} \otimes E_{i,i-1} + pI_a \otimes L_{N-i}(N, \hat{\Gamma}) \end{pmatrix},$$

$$i = 2, \ldots, N,$$

$$Q_{i,i} = \Delta_i + diag\{(D_0 \oplus S) \oplus A_j(N-i+j, \Gamma), j = 0, \ldots, i\} +$$

$$(1-p)diag^{-}\{I_a \otimes L_{N-i}(N-i+j, \hat{\Gamma}), j = 1, \ldots, i\}, \ i = 1, \ldots, N-1,$$

$$Q_{N,N} = \Delta_N + diag\{(D(1) \oplus S) \oplus A_j(j, \Gamma), j = 0, \ldots, N\} + (1-p)diag^{-}\{I_a \otimes L_0(j, \hat{\Gamma}), j = 1, \ldots, N\},$$

$$Q_{i,i+k} = \begin{pmatrix} diag\{D_k^{(1)} \otimes I_{MC_{j+R-1}^{R-1}}, j = 0, \ldots, i\} & | & O_{a \sum_{j=0}^{i} C_{j+R-1}^{R-1} \times a \sum_{j=i+1}^{i+k} C_{j+R-1}^{R-1}} \end{pmatrix} +$$

$$+ \begin{pmatrix} O_{a \sum_{j=0}^{i} C_{j+R-1}^{R-1} \times a \sum_{j=0}^{k-1} C_{j+R-1}^{R-1}} & | & diag\{D_k^{(2)} \otimes I_M \otimes P_{j,j+k}(\boldsymbol{\gamma}), j = 0, \ldots, i\} \end{pmatrix},$$

$$i = 1, \ldots, N-1, k = 1, \ldots, N-i-1,$$

$$Q_{i,N} = \begin{pmatrix} diag\{\sum_{k=N-i}^{\infty} D_k^{(1)} \otimes I_{MC_{j+R-1}^{R-1}}, j = 0, \ldots, i\} & | & O_{a \sum_{j=0}^{i} C_{j+R-1}^{R-1} \times a \sum_{j=i+1}^{N} C_{j+R-1}^{R-1}} \end{pmatrix} +$$

$$+ \begin{pmatrix} O_{a \sum_{j=0}^{i} C_{j+R-1}^{R-1} \times a \sum_{j=0}^{N-i-1} C_{j+R-1}^{R-1}} & | & diag\{\sum_{k=N-i}^{\infty} D_k^{(2)} \otimes I_M \otimes P_{j,j+N-i}(\boldsymbol{\gamma}), j = 0, \ldots, i\} \end{pmatrix},$$

$$i = 1, \ldots, N-1.$$

Here, Δ_i, $i = 1, \ldots, N$, *are the diagonal matrices ensuring the equality* $Q\mathbf{e} = \mathbf{0}^T$.

Proof. (1). The entries of the block $Q_{i,i+k}, i = 0, \ldots\ldots, N-1, k = 1, \ldots, N-i$, define the transition rates of the Markov chain ξ_t, $t \geq 0$, that lead to the increase the number of customers in the buffer by k.

If $i = 0$, the transitions can occur as a result of the following events:

- the batch of $(k+1)$ type 1 customers arrives to the idle system and one of the customers occupies the server. The rates of this event are defined by the matrix $D_{k+1}^{(1)} \otimes \boldsymbol{\beta}$, if $k = 1, \ldots, N-1$, and by the matrix $\sum_{k=N+1}^{\infty} D_k^{(1)} \otimes \boldsymbol{\beta}$, if $k = N$.

- the batch of $(k+1)$ type 2 customers arrives to the idle system, one of the customers occupies the server and a timer is set for each of $min\{k, N\}$ customers placed in the buffer. The rates of this event are defined by the matrix $D_{k+1}^{(2)} \otimes \boldsymbol{\beta} \otimes P_{0,k}(\boldsymbol{\gamma})$, if $k = 1, \ldots, N-1$ and by the matrix $\sum_{k=N+1}^{\infty} D_k^{(2)} \otimes \boldsymbol{\beta} \otimes P_{0,N}(\boldsymbol{\gamma})$, if $k = N$.
- the batch of k type 1 customers arrives to the system when the buffer is empty while the server is busy. In this case, $min\{k, N\}$ customers are placed in the buffer. The rates of this event are defined by the matrix $D_k^{(1)} \otimes I_M$, if $k = 1, \ldots, N-1$, and by the matrix $\sum_{k=N}^{\infty} D_k^{(1)} \otimes I_M$, if $k = N$.
- the batch of k type 2 customers arrive to the system when the buffer is empty and the server is busy. In this case $min\{k, N\}$ customers are placed in the buffer and timers are set for these customers. The transition rates of this event are defined by the matrix $D_k^{(2)} \otimes I_M \otimes P_{0,k}(\boldsymbol{\gamma})$, if $k = 1, \ldots, N-1$, and by the matrix $\sum_{k=N}^{\infty} D_k^{(2)} \otimes I_M \otimes P_{0,N}(\boldsymbol{\gamma})$, if $k = N$.

The presence of zero blocks in the matrices $Q_{0,k}$ is explained by the fact that simultaneous arrival of both priority and non-priority customers in the $BMMAP$ over an infinitesimal time interval is possible only with zero probability.

Thus, we have explained the structure of the generator blocks $Q_{i,i+k}$ for $i = 0$. We have considered two cases: the server is idle and the server is busy. If $i = 1, \ldots, N-1$, the server is busy a priori. Because of this, the explanation of the structure of the blocks $Q_{i,i+k}$ for $i \neq 0$ is similar to the explanation given for the case when $i = 0$ and the server is busy.

(2). The entries of the block $Q_{i,i-1}, i = 1, \ldots, N$, define the transition rates of the Markov chain ξ_t, $t \geq 0$, at the moments of the decrease the total number of customers in the buffer by one.

If $i = 1$, the transition occurs as a result of the following events:

- the service of the current customer finishes and the only priority customer staying in the buffer occupies the server. The rates of this event are defined by the matrix $I_{\bar{W}} \otimes \boldsymbol{S}_0 \boldsymbol{\beta}$.
- the service of the current customer finishes and the only non-priority customer staying in the buffer occupies the server. At this moment, the timer set for this customer is switched-off. The rates of this event are defined by the matrix $I_{\bar{W}} \otimes \boldsymbol{S}_0 \boldsymbol{\beta} \otimes \mathbf{e}_R$.
- the timer set for the only non-priority customer staying in the buffer expires while the server is still busy. In this case the non-priority customer with the probability p leaves the system forever. The rates of this event are defined by the matrix $pI_a \otimes \boldsymbol{\Gamma}_0$.

If $i = 2, \ldots, N$, the transitions of the Markov chain ξ_t, $t \geq 0$, related to the decrease the total number of customers in the buffer by one occur as a result of following events:

- the service of the current customer finishes and there is at least one priority customer in the buffer. Then, the first priority customer goes to the server and the number of customers in the buffer decreases by one. The rates of this event are defined by the matrix $diag\{I_{\bar{W}} \otimes \boldsymbol{S}_0 \boldsymbol{\beta} \otimes I_{C_{j+R-1}^{R-1}}, j = 0, \ldots, i-1\}$.
- the timer expires for one of non-priority customers and this customer leaves the buffer forever. The rates of this event are defined by the matrices $pI_a \otimes L_{N-i}(N-i+j, \hat{\Gamma})$, $j = 1, \ldots, i$.
- the service of the current customer finishes and there are no priority customers in the buffer. Then one of the non-priority customers whose timer is in the maximal (among all timers) phase $r, r = 1, \ldots, R$, goes to the server and the number of non-priority customers in the buffer whose timer is in the phase r decreases by one. The rates of this event are defined by the matrix $I_{\bar{W}} \otimes \boldsymbol{S}_0 \boldsymbol{\beta} \otimes E_{i,i-1}$.

(3). The non-diagonal entries of the block $Q_{i,i}$ define the transition rates of the chain ξ_t, that do not lead to the change of the number i of customers in the buffer. The case of $i = 0$ is not commented on here due its simplicity. In the case $i = 1, \ldots, N$, the mentioned transitions occur as a result of the following events:

- the underlying process of $BMMAP$ makes a transition without the generation of customers or the PH service time underlying process makes a transition which does not lead to the service completion. The rates of these events are defined by the non-diagonal entries of matrix $diag\{(D_0 \oplus S) \otimes I_{C_{j+R-1}^{R-1}}, j = 0, \dots, i\}$, if $i = 1, \dots, N-1$ and by the non-diagonal entries of the matrix $diag\{(D(1) \oplus S) \otimes I_{C_{j+R-1}^{R-1}}, j = 0, \dots, N\}$, if $i = N$.
- the underlying process of the PH timer set for one of j non-priority customers in the buffer makes a transition, which does not lead to the timer expiration. The rates of this event are defined by the non-diagonal entries of the matrix $diag\{I_a \otimes A_j(N-i+j, \Gamma), j = 0, \dots, i\}$.
- the timer expires for one of j non-priority customers in the buffer but the server is busy. Then, the mentioned customer with the probability $1-p$ gains the higher priority and joins the tail of the priority customers queue. The rates of this event are defined by the entries of the matrix $(1-p)diag^{-}\{I_a \otimes L_{N-i}(N-i+j, \hat{\Gamma}), j = 1, \dots, i\}$.

The diagonal entries of the block $Q_{i,i}, i = 0, \dots, N$ are negative and the modulus of each entry defines the rate of leaving the corresponding state of the Markov chain $\xi_t, t \geq 0$.

Lemma is proven. □

4. Stationary Distribution: Stationary Performance Measures

Since the Markov chain ξ_t is irreducible with a finite state space, it has the unique stationary distribution.

Let $\mathbf{p}_i$ be the row vector of the steady-state probabilities of the states having the value i of the first component, $i = 0, \dots, N$. The entries of the vector $\mathbf{p}_0$ of size K_0 give the steady-state probabilities that the buffer is empty, the server is idle or busy, the underlying process of the $BMMAP$ is in any of the $(W+1)$ states and, if the server is busy, the service process is in any of the M phases. The entries of the vector $\mathbf{p}_i$ of size $K_i, i = 1, \dots, N$, give the steady-state probabilities that there are i customers in the buffer, the underlying process of the $BMMAP$ is in any of the $(W+1)$ states, the service process is in any of the M phases and the timers process $\mathbf{n}_t$ is in any of the $\sum_{j=0}^{i} C_{j+R-1}^{R-1}$ states.

Denote as $\mathbf{p} = (\mathbf{p}_0, \mathbf{p}_i, \dots, \mathbf{p}_N)$ the stationary probability vector of the chain ξ_t. It is well known that this vector is the unique solution of the following system of linear algebraic equations:

$$\mathbf{p}Q = \mathbf{0}, \mathbf{p}\mathbf{e} = 1. \tag{1}$$

In case of small dimension, the system of Equation (1) can be solved on a computer by standard methods. However, with more or less large values of N, R, the order of this system becomes so large that it is not possible to solve this system directly, for example, using the method of the inverse matrix. In such a case, we use the algorithm that was elaborated in [23]. This algorithm is stable, takes into account the upper-Hessenberg structure of the generator Q and operates with the generator blocks $Q_{i,l}$. The sizes of these blocks are defined by the values $K_0 = \bar{W}(1+M)$, and $K_i = \bar{W}M \sum_{j=0}^{i} C_{j+R-1}^{R-1}$, $i = 1, \dots, N$, whereas the size of whole system of Equation (1) is equal to $\sum_{i=0}^{N} K_i$.

Having calculated the stationary distribution $\mathbf{p}_i, i = 0, \dots, N$, we are able to calculate a number of system performance measures of interest. Below we give the expressions for these performance measures along with brief explanations of nontrivial formulas.

- Probability that the system is idle

$$p_0 = \mathbf{p}_0 \begin{pmatrix} \mathbf{e}_{\bar{W}} \\ \mathbf{0}_a^T \end{pmatrix}.$$

- Probability that the buffer is empty and the server is busy

$$p_1 = \mathbf{p}_0 \begin{pmatrix} \mathbf{0}^T_{\bar{W}} \\ \mathbf{e}_a \end{pmatrix}.$$

- Probability that there are i customers in the buffer, of which j customers are non-priority

$$p_{i,j} = \mathbf{p}_i \begin{pmatrix} \mathbf{0}^T_{\sum_{l=0}^{j-1} C_{l+R-1}^{R-1}} \\ \mathbf{e}_{aC_{j+R-1}^{R-1}} \\ \mathbf{0}^T_{\sum_{l=j+1}^{i} C_{l+R-1}^{R-1}} \end{pmatrix}, \; j = 0, \ldots, i, i = 1, \ldots, N.$$

- Probability that there are $i > 0$ customers in the buffer

$$p_i = \sum_{j=0}^{i} p_{i,j}, i = 1, \ldots, N.$$

- Mean number of customers in the buffer

$$L_{buf} = \sum_{i=1}^{N} i p_i.$$

- Mean number of non-priority customers in the buffer

$$L^{(non-prior)} = \sum_{i=1}^{N} \sum_{j=1}^{i} j p_{i,j}.$$

- Mean number of priority customers in the buffer

$$L^{(prior)} = L_{buf} - L^{(non-prior)}.$$

- Probability that an arbitrary customer will be lost due to lack of buffer space

$$P_{loss} = 1 - \frac{1}{\lambda} \left[\mathbf{p}_0 \begin{pmatrix} O_{\bar{W} \times M} \\ \mathbf{e}_{\bar{W}} \otimes I_M \end{pmatrix} + \sum_{i=1}^{N} \mathbf{p}_i \mathcal{I}_i^{(1)} \right] \boldsymbol{S}_0, \tag{2}$$

where $\mathcal{I}_i^{(1)} = \begin{pmatrix} \mathbf{e}_{\bar{W}} \otimes I_M \otimes \mathbf{e}_{C_{R-1}^{R-1}} \\ \mathbf{e}_{\bar{W}} \otimes I_M \otimes \mathbf{e}_{C_R^{R-1}} \\ \vdots \\ \mathbf{e}_{\bar{W}} \otimes I_M \otimes \mathbf{e}_{C_{i+R-1}^{R-1}} \end{pmatrix}$.

A brief explanation of Equation (2) is as follows. The expression in the braces is the rate of the output flow and λ is the input rate. Then, the ratio of these rates gives the probability that an arbitrary customer will be served, and the complimentary probability gives the desired probability P_{loss}.

Note that the probability P_{loss} can be calculated by an alternative equation obtained by considering the situation at the arrival time. This equation has the form

$$P_{loss} = 1 - \frac{1}{\lambda} \left\{ \mathbf{p}_0 \begin{pmatrix} I_{\bar{W}} \\ O_{\bar{W}} \otimes \mathbf{e}_M \end{pmatrix} \sum_{k=0}^{N+1} (k - N - 1) D_k \mathbf{e} + \right.$$

$$\mathbf{p}_0 \begin{pmatrix} O_{\bar{W}} \\ I_{\bar{W}} \otimes \mathbf{e}_M \end{pmatrix} \sum_{k=0}^{N} (k-N) D_k \mathbf{e} + \sum_{i=1}^{N-1} \mathbf{p}_i \mathcal{I}_i^{(2)} \sum_{k=0}^{N-i} (k-N+i) D_k \mathbf{e} \Big\}, \tag{3}$$

where $D_k = D_k^{(1)} + D_k^{(2)}$,

$$\mathcal{I}_i^{(2)} = \begin{pmatrix} I_{\bar{W}} \otimes \mathbf{e}_M \otimes \mathbf{e}_{C_{R-1}^{R-1}} \\ I_{\bar{W}} \otimes \mathbf{e}_M \otimes \mathbf{e}_{C_R^{R-1}} \\ \vdots \\ I_{\bar{W}} \otimes \mathbf{e}_M \otimes \mathbf{e}_{C_{i+R-1}^{R-1}} \end{pmatrix}.$$

A brief explanation of Equation (3) is as follows. The expression in the braces gives the rate of customers from the input flow that are accepted into the system. Dividing this rate by the input rate λ, we obtain the probability that an arbitrary customer will not be lost due to the lack of buffer space. The complimentary probability gives the probability P_{loss}.

Similarly calculated the probability of loss of customers of each type.

- Probabilities that an arbitrary priority customer ($l = 1$) and an arbitrary non-priority customer ($l = 2$) will be lost due to lack of buffer space

$$P_{loss}^{(l)} = 1 - \frac{1}{\lambda_l} \Big\{ \mathbf{p}_0 \begin{pmatrix} I_{\bar{W}} \\ O_{\bar{W}} \otimes \mathbf{e}_M \end{pmatrix} \Big[\sum_{k=1}^{N+1} k D_k^{(l)} \mathbf{e} +$$

$$(N+1) \sum_{k=N+2}^{\infty} D_k^{(l)} \mathbf{e} \Big] + \mathbf{p}_0 \begin{pmatrix} O_{\bar{W}} \\ I_{\bar{W}} \otimes \mathbf{e}_M \end{pmatrix} \Big[\sum_{k=1}^{N} k D_k^{(l)} \mathbf{e} + N \sum_{k=N+1}^{\infty} D_k^{(l)} \mathbf{e} \Big] +$$

$$\sum_{i=1}^{N-1} \mathbf{p}_i \mathcal{I}_i^{(2)} \Big[\sum_{k=1}^{N-i} k D_k^{(l)} \mathbf{e} + (N-i) \sum_{k=N-i+1}^{\infty} D_k^{(l)} \mathbf{e} \Big] \Big\}, l = 1, 2.$$

- Probability that an arbitrary non-priority customer will be lost due to impatience

$$P_{loss}^{(imp)} = \frac{p}{\lambda_2} \sum_{i=1}^{N} \mathbf{p}_i \mathcal{I}_i^{(\mathcal{L})} \tag{4}$$

where $\mathcal{I}_i^{(\mathcal{L})} = \begin{pmatrix} \mathbf{e}_a \otimes L_{N-i}(N-i, \hat{\Gamma}) \mathbf{e} \\ \mathbf{e}_a \otimes L_{N-i}(N-i+1, \hat{\Gamma}) \mathbf{e} \\ \vdots \\ \mathbf{e}_a \otimes L_{N-i}(N, \hat{\Gamma}) \mathbf{e} \end{pmatrix}$.

A brief explanation of Equation (4) is as follows. The expression $\sum_{i=1}^{N} \mathbf{p}_i \mathcal{I}_i^{(\mathcal{L})}$ gives the mean number of expirations of the timers per unit time. Each non-priority customer from the buffer, for which the timer expires, with the probability p leaves the system (is lost). The ratio of the rate of flow of thus lost non-priority customers to the input rate of non-priority customers, λ_2, is $P_{loss}^{(imp)}$.

- Probability that an arbitrary non-priority customer accepted to the buffer will be lost due to impatience

$$\bar{P}_{loss}^{(imp)} = \frac{P_{loss}^{(imp)}}{1 - P_{loss}^{(2)}}.$$

5. Stationary Distribution of a Priority Customer Waiting Time

Let $w_\tau^{(1)}, \tau \geq 0$, be the process of waiting time of an arbitrary priority customer accepted into the system at the moment τ and $w_\tau^{(2)}, \tau \geq 0$, be the process of waiting time of a customer which changed the low priority to the high priority during the stay in the buffer starting from the moment of changing the priority. We assume that an arbitrary priority customer, which arrives in a batch of size k, is numerated as the ith in the batch with the probability $\frac{1}{k},\ i = 1, \ldots, k$. Let $W_\tau^{(l)}(t) = P\{w_\tau^{(l)} < t\}$ be the distribution function of the process $w_\tau^{(l)}, l = 1, 2$. Denote by $W_l(t)$ the corresponding stationary distribution functions of the waiting time, i.e., $W_l(t) = \lim\limits_{\tau \to \infty} W_\tau^{(l)}(t), l = 1, 2$.

It is seen from the definitions, that $W_l(t), l = 1, 2$, are the conditional distribution functions. To find these functions, we focus on the calculation of the joint probabilities, which are defined as follows:

$\tilde{W}_1(t)$ is the joint probability that an arbitrary priority customer will be accepted into the system and his/her waiting time will be less than t;

$\tilde{W}_2(t)$ is the joint probability that an arbitrary non-priority customer will become the priority customer and after that his/her waiting time will not exceed value t.

The following theorem holds.

Theorem 1. *The functions $\tilde{W}_1(t)$ and $\tilde{W}_2(t)$ have the forms*

$$\tilde{W}_1(t) =$$

$$\lambda^{-1}\Bigg\{\mathbf{p}_0\begin{pmatrix} I_{\bar{W}} \\ O_{a\times\bar{W}}\end{pmatrix}\sum_{k=1}^{\infty} D_k^{(1)}\mathbf{e}\Bigg[1+\sum_{l=2}^{min\{N+1,k\}}(\boldsymbol{\beta}, \mathbf{0}_{(l-2)M})(I - e^{A^{(l-2)}t})\mathbf{e}_{(l-1)M}\Bigg]+$$

$$\mathbf{p}_0\begin{pmatrix} O_{\bar{W}\times a} \\ I_a\end{pmatrix}\sum_{k=1}^{\infty}(D_k^{(1)}\mathbf{e}\otimes I_M)\sum_{l=1}^{min\{N,k\}}(I_M|O_{M\times(l-1)M})(I - e^{A^{(l-1)}t})\mathbf{e}_{lM}+$$

$$\sum_{i=1}^{N-1}\sum_{j=0}^{i}\mathbf{p}_i(j)\sum_{k=1}^{\infty}(D_k^{(1)}\mathbf{e}\otimes I_M\otimes\mathbf{e}_{C_{j+R-1}^{R-1}})\times$$

$$\sum_{l=1}^{min\{N-i,k\}}(I_M|O_{M\times(i-j+l-1)M})(I - e^{A^{(i-j+l-1)}t})\mathbf{e}_{(i-j+l)M}\Bigg\}, \tag{5}$$

$$\tilde{W}_2(t) =$$

$$\frac{1-p}{\hat{\gamma}}\sum_{i=1}^{N}\sum_{j=1}^{i}\mathbf{p}_i(j)\Big[\mathbf{e}_{\bar{W}}\otimes I_M\otimes L_{N-i}(N-i+j,\hat{\Gamma})\mathbf{e}\Big](I_M|O_{M\times(i-j)M})(I - e^{A^{(i-j)}t})\mathbf{e}_{(i-j+1)M}, \tag{6}$$

where

$$\hat{\gamma} = \sum_{i=1}^{N}\sum_{j=1}^{i}\mathbf{p}_i(j)\Big[\mathbf{e}_a\otimes L_{N-i}(N-i+j,\hat{\Gamma})\mathbf{e}\Big].$$

Here the square matrix $A^{(n)}$ of size $M(n+1)$ is defined as

$$A^{(n)} = \begin{pmatrix} S & \mathbf{S}_0\boldsymbol{\beta} & O & \ldots & O & O & O \\ O & S & \mathbf{S}_0\boldsymbol{\beta} & \ldots & O & O & O \\ O & O & S & \ldots & O & O & O \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ O & O & O & \ldots & O & S & \mathbf{S}_0\boldsymbol{\beta} \\ O & O & O & \ldots & O & O & S \end{pmatrix}, n = 0, ..., N-1,$$

and the sub-vector $\mathbf{p}_i(j)$ *is the part of the vector* $\mathbf{p}_i$ *corresponding to presence of j non-priority customers in the buffer and is defined as*

$$\mathbf{p}_i(j) = \mathbf{p}_i \begin{pmatrix} O_{a\sum_{l=0}^{j-1} C_{l+R-1}^{R-1} \times aC_{j+R-1}^{R-1}} \\ I_{aC_{j+R-1}^{R-1}} \\ O_{a\sum_{l=j+1}^{i} C_{l+R-1}^{R-1} \times aC_{j+R-1}^{R-1}} \end{pmatrix}.$$

Proof. We first consider the derivation of Equation (5) for the function $\tilde{W}_1(t)$. Using the equation of total probability, this function can be written in the following form:

$$\tilde{W}_1(t) = \mathbf{p}_0 \begin{pmatrix} I_{\bar{W}} \\ O_{a\times\bar{W}} \end{pmatrix} \sum_{k=1}^{\infty} \frac{kD_k^{(1)}\mathbf{e}}{\lambda} \frac{min\{N+1,k\}}{k} \Bigg[\frac{1}{min\{N+1,k\}} \times 1 +$$

$$\sum_{l=2}^{min\{N+1,k\}} \frac{1}{min\{N+1,k\}} (\boldsymbol{\beta}, \mathbf{0}_{(l-2)M}) \int_0^t e^{A^{(l-2)}x} \mathbf{A}_0^{(l-2)} dx \Bigg] +$$

$$+\mathbf{p}_0 \begin{pmatrix} O_{\bar{W}\times a} \\ I_a \end{pmatrix} \sum_{k=1}^{\infty} (\frac{kD_k^{(1)}\mathbf{e}}{\lambda} \otimes I_M) \frac{min\{N,k\}}{k} \times$$

$$\sum_{l=1}^{min\{N,k\}} \frac{1}{min\{N,k\}} (I_M | O_{M\times(l-1)M}) \int_0^t e^{A^{(l-1)}x} \mathbf{A}_0^{(l-1)} dx +$$

$$+ \sum_{i=1}^{N-1} \sum_{j=0}^{i} \mathbf{p}_i(j) \sum_{k=1}^{\infty} (\frac{kD_k^{(1)}\mathbf{e}}{\lambda} \otimes I_M \otimes \mathbf{e}_{C_{j+R-1}^{R-1}}) \frac{min\{N-i,k\}}{k} \times$$

$$\sum_{l=1}^{min\{N-i,k\}} \frac{1}{min\{N-i,k\}} (I_M | O_{M\times(i-j+l-1)M}) \int_0^t e^{A^{(i-j+l-1)}x} \mathbf{A}_0^{(i-j+l-1)} dx, \tag{7}$$

where

$$\mathbf{A}_0^{(n)} = -A^{(n)}\mathbf{e}.$$

To clarify Equation (7), we first explain the meaning of the matrices $A^{(n)}$ and $\int_0^t e^{A^{(n)}x} \mathbf{A}_0^{(n)} dx$. Let us tag an arbitrary priority customer accepted to the system. Suppose that at the arrival moment of this customer, the server is busy and $max\{0, n-1\}$ priority customers are in the buffer. The waiting time $T^{(n)}$ of the tagged customer is the sum of the remaining service time of customer in the service and the service time of $max\{0, n-1\}$ priority customers from the buffer. Given the known distribution of the service phases at the arrival time of the tagged customer (let this distribution be given by the vector $\boldsymbol{\alpha}$ of size M) the waiting time $T^{(n)}$ of this customer has a PH distribution with $(n+1)M$ phases defined by the irreducible representation $(\tilde{\boldsymbol{\alpha}}, A^{(n)})$, where $\tilde{\boldsymbol{\alpha}} = (\boldsymbol{\alpha}, \mathbf{0}_{nM})$. Then $P\{T^{(n)} < t\} = \tilde{\boldsymbol{\alpha}} \int_0^t e^{A^{(n)}x} \mathbf{A}_0^{(n)} dx$.

Now let us explain the meaning of the terms in Equation (7) for the function $\tilde{W}_1(t)$. The scalar

$$\mathbf{p}_0 \begin{pmatrix} I_{\bar{W}} \\ O_{a\times\bar{W}} \end{pmatrix} \frac{kD_k^{(1)}\mathbf{e}}{\lambda} \frac{min\{N+1,k\}}{k} \frac{1}{min\{N+1,k\}} \times 1$$

is the probability that the system is empty at the arrival moment of the batch of k priority customers, the tagged priority customer arrives in this batch and receives the first position in the batch what implies that his/her waiting time is less than t for any $t > 0$.

As stated above, the waiting time for an arbitrary priority customer has a PH distribution. Taking into account this fact, we can see that the vector

$$\mathbf{p}_0 \begin{pmatrix} I_{\bar{W}} \\ O_{a\times\bar{W}} \end{pmatrix} \frac{kD_k^{(1)}\mathbf{e}}{\lambda} \frac{min\{N+1,k\}}{k} \frac{1}{min\{N+1,k\}} (\boldsymbol{\beta}, \mathbf{0}_{(l-2)M})$$

defines the initial phase distribution of the waiting time of the tagged priority customer conditional he/she enters the empty system in a batch of size k, is accepted into the system, and occupies the $(l-1)$th place in the buffer, $l = 2, \ldots, min\{N+1,k\}$. Multiplying the indicated row vector by the column vector $\int_0^t e^{A^{(l-2)}x}\mathbf{A}_0^{(l-2)}dx$, we obtain the probability that such a tagged customer has his/her waiting time less than t.

The vector

$$\mathbf{p}_0 \begin{pmatrix} O_{\bar{W}\times a} \\ I_a \end{pmatrix} \left(\frac{kD_k^{(1)}\mathbf{e}}{\lambda} \otimes I_M\right) \frac{min\{N,k\}}{k} \frac{1}{min\{N,k\}} (I_M | O_{M\times(l-1)M})$$

defines the initial state of underlying process of phase distribution of the waiting time of the tagged priority customer, which is dependent on whether or not he/she enters the system in the batch of size k when the buffer is empty and the server is busy, is accepted into the system and occupies the lth place in the buffer, $l = 1, \ldots, min\{N,k\}$. Multiplying the indicated vector by the column vector $\int_0^t e^{A^{(l-1)}x}\mathbf{A}_0^{(l-1)}dx$, we obtain the probability that such a tagged customer has his/her waiting time less than t.

The vector

$$\mathbf{p}_i(j)\left(\frac{kD_k^{(1)}\mathbf{e}}{\lambda} \otimes I_M \otimes \mathbf{e}_{C_{j+R-1}^{R-1}}\right) \frac{min\{N-i,k\}}{k} \frac{1}{min\{N-i,k\}} (I_M | O_{M\times(i-j+l-1)M})$$

in double sum on the right-hand side of Equation (7) defines the initial phase distribution of the waiting time for the tagged priority customer, which enters the system in the batch of size k when i customers are in the buffer, $i-j$ of which are priority ones, and occupies the lth place in buffer, $l = i-j+1, \ldots, min\{N-i,k\}$. Multiplying the indicated vector by column vector $\int_0^t e^{A^{(i-j+l-1)}x}\mathbf{A}_0^{(i-j+l-1)}dx$, we obtain the probability that such a tagged priority customer has waiting time less than t.

Now we use the equation for total probability by summing up the described terms on the right-hand side of Equation (7) over k and l. Then, on the right-hand side of Equation (7) we obtain the probability that the arbitrary tagged priority customer was accepted into the system and his/her waiting time is less than t. This is by definition the function $\tilde{W}_1(t)$. Performing the calculation of integrals and some simple transformations in Equation (7), we obtain the desired Equation (5).

Now we prove Equation (6) for $\tilde{W}_2(t)$. In the equation, the mth entry of the vector

$$\mathbf{p}_i(j)\left[\mathbf{e}_{\bar{W}} \otimes I_M \otimes L_{N-i}(N,\hat{\Gamma})\mathbf{e}_{C_{j+R-2}^{R-1}}\right]$$

defines the mean number of expirations of the timers per unit of time conditional that i customers are staying in the buffer, j of which are non-priority ones and the service process is in the phase m.

Note that the value $\hat{\gamma}$ is the total rate of the flow of expiration of timers on non-priority customers. Then, the mth component of the vector

$$\boldsymbol{\phi}^{(i,j)} = \mathbf{p}_i(j)\left[\mathbf{e}_{\bar{W}} \otimes I_M \otimes \frac{L_{N-i}(N,\hat{\Gamma})\mathbf{e}_{C_{j+R-2}^{R-1}}}{\hat{\gamma}}\right] \tag{8}$$

is the probability that at the moment of timer expiration on one of the non-priority customers, the service of the current customer is in the phase m and there are i customers in the buffer of j which are non-priority ones.

If a non-priority customer, whose timer has expired, becomes priority, then his/her waiting time $T^{(i,j)}$ consists of the residual service time of the customer in the service and the service time of $(i-j-1)$ of priority customers standing in the buffer in front of him/her. This time has a PH distribution with the representation $(\tilde{\boldsymbol{\phi}}^{(i,j)}, A^{(i-j)})$, where $\tilde{\boldsymbol{\phi}}^{(i,j)} = (\boldsymbol{\phi}^{(i,j)}, \mathbf{0}_{(i-j)M})$. Then the distribution of the random time $T^{(i,j)}$ is calculated as follows:

$$P\{T^{(i,j)} < t\} = \tilde{\boldsymbol{\phi}}^{(i,j)} \int_0^t e^{A^{(i-j)}x} \mathbf{A}_0^{(i-j)} dx. \tag{9}$$

Multiplying the right-hand side of Equation (8) by the right-hand side of Equation (9), calculating the integrals, summing over the possible values of i, j and multiplying the resulting expression by $1-p$, we obtain Equation (6) for the function $\tilde{W}_2(t)$. □

Corollary 1. *(i) The distribution function of the waiting time of an arbitrary priority customer accepted into the buffer is calculated by the equation*

$$W_1(t) = \frac{\tilde{W}_1(t)}{\tilde{W}_1(\infty)}.$$

(ii) The distribution function of the waiting time of a priority customer that initially was the non-priority customer is calculated by the equation

$$W_2(t) = \frac{\tilde{W}_2(t)}{\tilde{W}_2(\infty)}.$$

Proof. Since the function $W_l(t)$ is a conditional distribution function, it is calculated by the division of the joint probability $\tilde{W}_l(t)$ by the probability of the condition $\tilde{W}_l(\infty), l = 1, 2$. □

Corollary 2. *The average waiting time $\bar{w}_1$ of an arbitrary priority customer accepted into the buffer and the average waiting time $\bar{w}_2$ of a non-priority customer after he/she became the priority customer, are calculated by the equations*

$$\bar{w}_1 = \frac{1}{\lambda \tilde{W}_1(\infty)} \Bigg\{ \mathbf{p}_0 \begin{pmatrix} I_{\bar{W}} \\ O_{a\times\bar{W}} \end{pmatrix} \sum_{k=1}^{\infty} D_k^{(1)} \mathbf{e} \sum_{l=2}^{min\{N+1,k\}} (l-1)b_1 +$$

$$\mathbf{p}_0 \begin{pmatrix} O_{\bar{W}\times a} \\ I_a \end{pmatrix} \sum_{k=1}^{\infty} (D_k^{(1)} \mathbf{e} \otimes I_M) \sum_{l=1}^{min\{N,k\}} [(-S)^{-1} + I_M(l-1)b_1]\mathbf{e}_M +$$

$$\sum_{i=1}^{N-1} \sum_{j=0}^{i} \mathbf{p}_i(j) \sum_{k=1}^{\infty} (D_k^{(1)} \mathbf{e}_{\bar{W}} \otimes I_M \otimes \mathbf{e}_{C_{j+R-1}^{R-1}}) \sum_{l=1}^{min\{N-i,k\}} [(-S)^{-1} + I_M(i-j+l-1)b_1]\mathbf{e}_M \Bigg\}, \tag{10}$$

$$\bar{w}_2 = \frac{1-p}{\hat{\gamma}\tilde{W}_2(\infty)} \sum_{i=1}^{N} \sum_{j=1}^{i} \mathbf{p}_i(j)(\mathbf{e}_{\bar{W}} \otimes I_M \otimes (L_{N-i}(N-i+j, \hat{\Gamma})\mathbf{e}))[(-S)^{-1} + I_M(i-j)b_1]\mathbf{e}_M. \tag{11}$$

Proof. The proof follows from the equations $\bar{w}_l = \int_0^{\infty} t dW_l(t), l = 1, 2$ and Equations (5) and (6) for the functions $W_1(t), W_2(t)$. The derivation of Equations (10) and (11) is based on the following relations that are obtained by using the specific structure of the matrix $A^{(n)}$:

$$(\boldsymbol{\beta}, \mathbf{0}_{nM}) \int_0^{\infty} t d(I - e^{A^{(n)}t}) \mathbf{e}_{(n+1)M} = (\boldsymbol{\beta}, \mathbf{0}_{nM})(-A^{(n)})^{-1} \mathbf{e}_{(n+1)M} =$$

$$= (n+1)\boldsymbol{\beta}(-S)^{-1}\mathbf{e}_M = (n+1)b_1,$$

$$(I_M|O_{M\times nM})\int_0^\infty td(I - e^{A^{(n)}t})\mathbf{e}_{(n+1)M} = (I_M|O_{M\times nM})(-A^{(n)})^{-1}\mathbf{e}_{(n+1)M} =$$

$$= (-S)^{-1}\mathbf{e} + n\boldsymbol{\beta}(-S)^{-1}\mathbf{e}_M = (-S)^{-1}\mathbf{e}_M + \mathbf{e}_M n b_1,$$

where b_1 is the mean service time, $b_1 = \boldsymbol{\beta}(-S)^{-1}\mathbf{e}$. □

6. Numerical Examples

The goal of this section is to bring out the qualitative nature of the model under study through three illustrative numerical examples.

Example 1. *In this example, we plotted graphs of the distribution functions $W_1(t)$ and $W_2(t)$ under different values of the service rate μ. We considered the following input data.*

- $N = 10$, $p = 0.4$
- *To define the BMMAP, we first define the matrices D_0 and D as follows*

$$D_0 = \begin{pmatrix} -8.28142513 & 0 \\ 0 & -0.26874977 \end{pmatrix}, \quad D = \begin{pmatrix} 8.22628993 & 0.0551352 \\ 0.14964989 & 0.11909988 \end{pmatrix}.$$

The matrix D is split into two matrices $D^{(1)}$ and $D^{(2)}$ as

$$D^{(1)} = 0.1D,\ D^{(1)} = 0.9D.$$

It means that the arrival rate of batches of type 2 customers are nine times less than the arrival rate of batches of type 1 customers.

Next, we need to obtain the matrices $D_k^{(1)}$ and $D_k^{(2)}$.

We assume that the maximum batch size of type 1 (priority) customers is 5 and the distribution among batches of different sizes is carried out in accordance with the equation

$$D_k^{(1)} = D^{(1)}q^{k-1}(1-q)/(1-q^5), k = 1,\ldots,5,\ \textit{where } q = 0.8$$

Further, we assume that the maximum batch size of type 2 (non-priority) customers is 2 and the distribution among batches of different sizes is carried out in accordance with the equation

$$D_k^{(2)} = D^{(2)}q^{k-1}(1-q)/(1-q^2), k = 1,2,\ \textit{where } q = 0.2.$$

For this BMMAP $\lambda = 8$, $\lambda_1 = 1.569656$, $\lambda_2 = 6.430344$, $\lambda_1^{(b)} = 0.612413$, $\lambda_2^{(b)} = 5.511723$, $c_{var}^{(1)} = 1.693988$, $c_{var}^{(2)} = 3.417944$, $c_{cor}^{(1)} = 0.02342$, $c_{cor}^{(2)} = 0.187811$.

- *The duration of the random variable that defines the timer has a PH distribution with the representation $(\boldsymbol{\gamma}, \Gamma)$ where*

$$\boldsymbol{\gamma} = (1,0),\ \Gamma = \begin{pmatrix} -10 & 10 \\ 0 & -10 \end{pmatrix}.$$

It means that the time counted by the timer has the Erlangian distribution of order 2. The timer rate $\tau = 5$ and the coefficient of variation $c_{var} = 0.7$.

- *In this example, we considered three service processes, which differ in the service rate. In all these processes, the service time has the Erlangian distribution of order 2. The rates of the phases of this distribution are calculated as $8c$, where $c = 1, 2, 4$. Therefore, the service rates are equal to $4, 8, 16$, respectively.*

When calculating the functions $W_1(t)$ and $W_2(t)$, we consider $t \in [0.01, 4]$ and take 40 points uniformly distributed over the interval. In Table 2 we give the values of the functions $W_1(t)$ and $W_2(t)$ at 21 points.

Table 2. Distribution functions $W_1(t)$ and $W_2(t)$.

	$\mu = 4$		$\mu = 8$		$\mu = 16$	
t	$W_1(t)$	$W_2(t)$	$W_1(t)$	$W_2(t)$	$W_1(t)$	$W_2(t)$
0.01000	0.01036	0.00090	0.04832	0.01564	0.18563	0.09019
0.21462	0.03346	0.02151	0.28419	0.30267	0.80610	0.84344
0.41923	0.06301	0.04435	0.51361	0.52027	0.97322	0.97382
0.62385	0.09762	0.07161	0.70094	0.70086	0.99734	0.99734
0.82846	0.13948	0.10762	0.83818	0.83978	0.99983	0.99984
1.03308	0.19247	0.15798	0.92549	0.92856	0.99999	0.99999
1.23769	0.26089	0.22812	0.97134	0.97371	1.00000	1.00000
1.44231	0.34705	0.32021	0.99080	0.99195	1.00000	1.00000
1.64692	0.44871	0.43047	0.99751	0.99792	1.00000	1.00000
1.85154	0.55856	0.54924	0.99942	0.99954	1.00000	1.00000
2.05615	0.66635	0.66426	0.99988	0.99991	1.00000	1.00000
2.36308	0.80399	0.80770	0.99999	0.99999	1.00000	1.00000
2.77231	0.92138	0.92561	1.00000	1.00000	1.00000	1.00000
3.18154	0.97456	0.97677	1.00000	1.00000	1.00000	1.00000
3.59077	0.99319	0.99399	1.00000	1.00000	1.00000	1.00000
4.00000	0.99845	0.99867	1.00000	1.00000	1.00000	1.00000

As it is seen from Table 2, the functions $W_1(t)$ and $W_2(t)$ differ only slightly. This can be explained by the fact that the waiting time of an arbitrary arrived priority customer and the customer, which initially arrived as the non-priority one, may be different if the distributions of the number of customers at the moments of arrival of a priority customer and transfer of a non-priority customer to the class of priority customers would be different. However, as calculations in this example show, these distributions are close, and the waiting time distributions of two types of priority customers are not much different. Therefore, below in this example we considered only the function $W_1(t)$.

Figure 2 depicts the distribution function $W_1(t)$ under different values of the service rate μ.

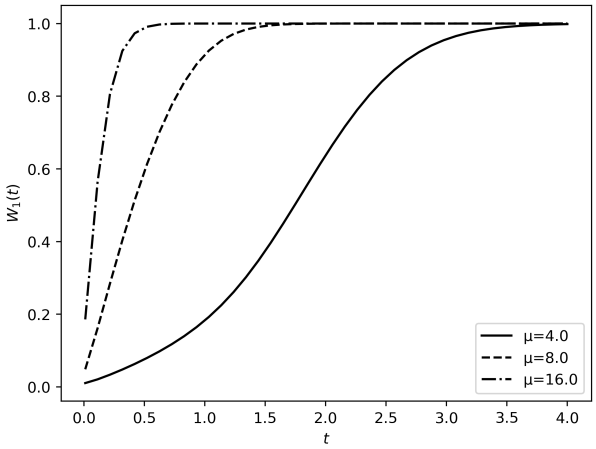

Figure 2. The distribution function $W_1(t)$ under different service rate μ.

As expected, under the fixed value of μ the function $W_1(t)$ increases with increasing t and tends to 1 when $t \to \infty$. It is also clear that the value of the function increases with increasing the service rate μ. Note also that the growth rate is greater for larger values of μ.

Example 2. *In this example, we investigate the dependence of the loss probabilities P_{loss}, $P_{loss}^{(1)}$, $P_{loss}^{(2)}$, $P_{loss}^{(imp)}$ on the buffer capacity N for the system with BMMAPs having the different coefficients of correlation.*

To this end, we considered three BMMAPs. In addition to the BMMAP defined in Example 1 and having the following characteristics: $\lambda = 8$, $\lambda_1 = 1.569656$, $\lambda_2 = 6.430344$, $\lambda_1^{(b)} = 0.612413$, $\lambda_2^{(b)} = 5.511723$, $c_{var}^{(1)} = 1.693988$, $c_{var}^{(2)} = 3.417944$, $c_{cor}^{(1)} = 0.02342$, $c_{cor}^{(2)} = 0.187811$, we considered two more BMMAPs having the same mean arrival rates $\lambda, \lambda_1, \lambda_2$, but different coefficients of correlation.

These additional $BMMAP$s are also defined by certain matrices D_0 and D from which the matrices $D_k^{(1)}, k = 1, \ldots, 5$, $D_k^{(2)}, k = 1, 2$, are defined in the same way as in Example 1.

The first $BMMAP$ is a heterogeneous group Poisson process. For this $BMMAP$, $D_0 = -6.124137$, $D = 6.124137$, the coefficients of correlation $c_{cor}^{(1)} = c_{cor}^{(2)} = 0$.

The second $BMMAP$ is defined by the matrices

$$D_0 = \begin{pmatrix} -29.668038 & 0.003450 \\ 0.006900 & -0.952137 \end{pmatrix}, \; D = \begin{pmatrix} 29.323061 & 0.341527 \\ 0.068995 & 0.876242 \end{pmatrix}.$$

For this $BMMAP$, $c_{cor}^{(1)} = 0.205982$, $c_{cor}^{(2)} = 0.402641$, $c_{var}^{(1)} = 2.394561$, $c_{var}^{(2)} = 3.087863$.

The service time has the Erlangian distribution of order 2 with parameter 20. The timer has the Erlangian distribution of order 2 with parameter 10. The probability that a non-priority customer leaves the system unserved when the timer expires is $p = 0.4$.

Figures 3–6 show the dependence of the loss probabilities $P_{loss}^{(1)}$, $P_{loss}^{(2)}$, P_{loss}, $P_{loss}^{(imp)}$ on the buffer capacity N for the systems with $BMMAP$s having different coefficients of correlation.

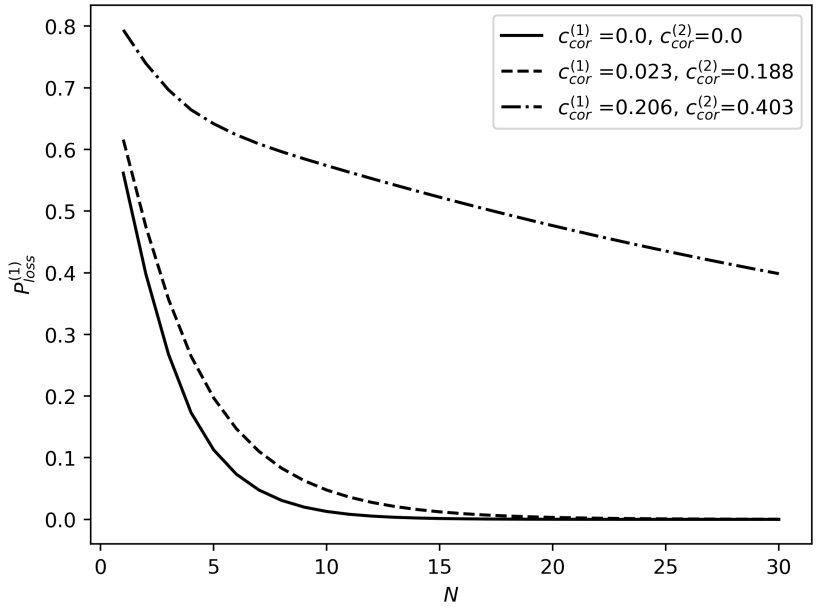

Figure 3. The loss probability $P_{loss}^{(1)}$ as a function of the buffer capacity N for $BMMAP$s with different coefficients of correlation.

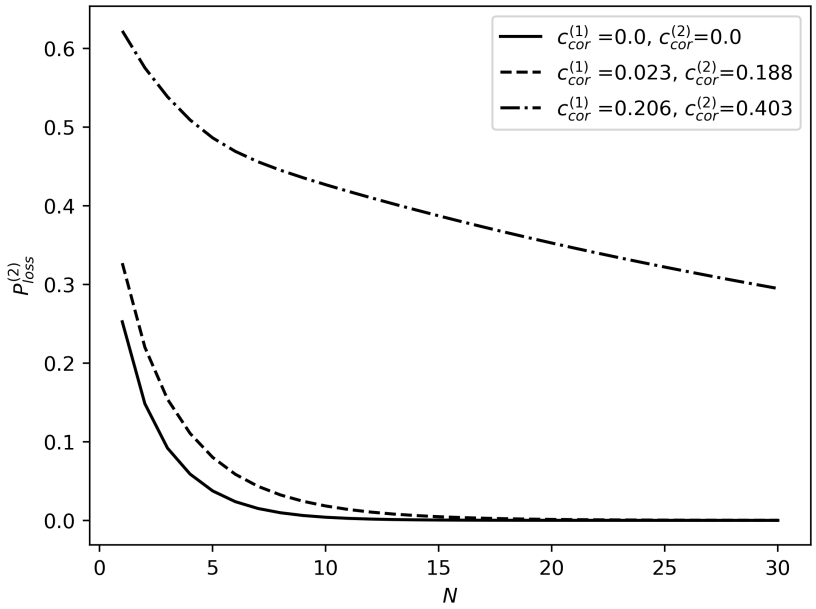

Figure 4. The loss probability $P_{loss}^{(2)}$ as a function of the buffer capacity N for $BMMAP$s with different coefficients of correlation.

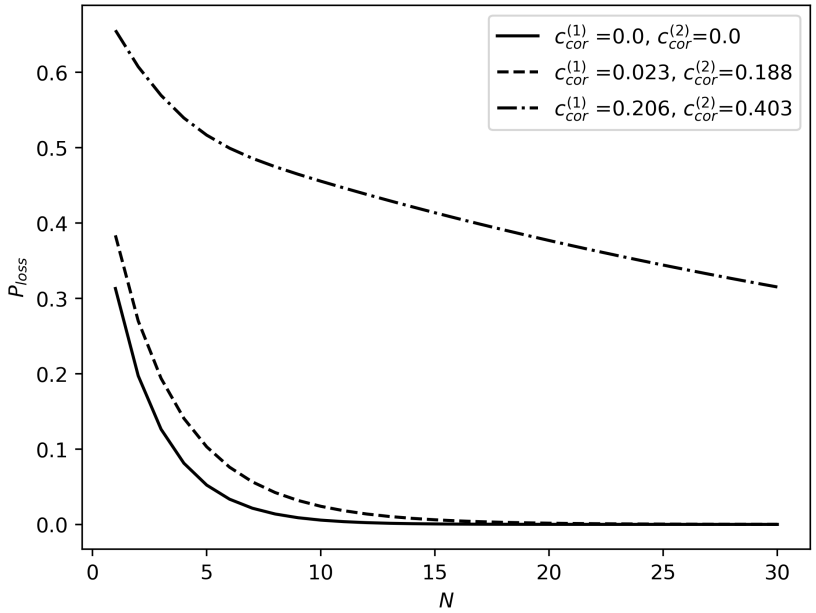

Figure 5. The loss probability P_{loss} as a function of the buffer capacity N for $BMMAP$s with different coefficients of correlation.

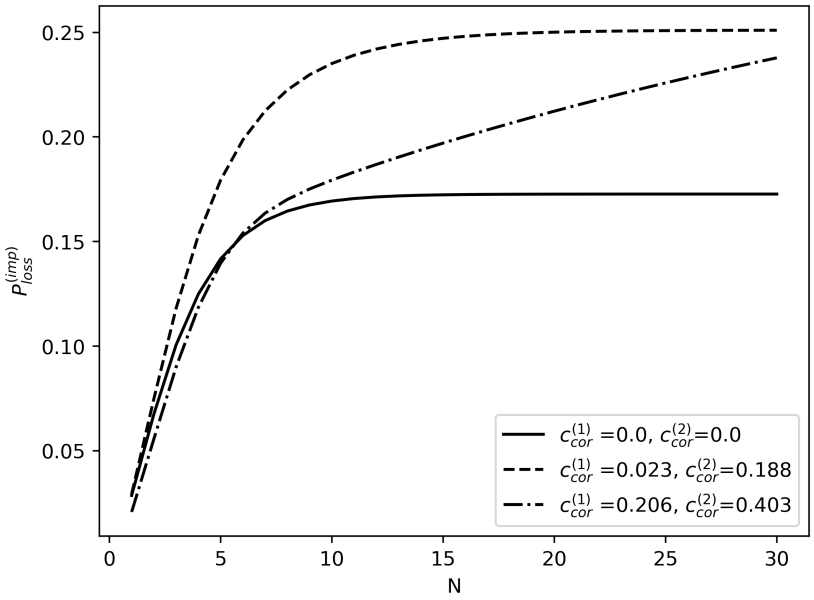

Figure 6. The loss probability $P_{loss}^{(imp)}$ as a function of the buffer capacity N for $BMMAP$s with different coefficients of correlation.

As expected, the loss probabilities of priority and non-priority customers, $P_{loss}^{(1)}$, $P_{loss}^{(2)}$, as well as the general loss probability, P_{loss}, decrease with N increasing. It is also seen that the loss probabilities depend on the correlation in the input flow. Under the same value of N each of these probabilities increases when the correlation increases. The difference in the value of any of these probabilities for $BMMAP$ with different coefficients of correlation may be very significant. This is especially evident if we compare the curves for $BMMAP$s with coefficients of correlation $c_{cor}^{(1)} = c_{cor}^{(2)} = 0$ and $c_{cor}^{(1)} = 0.02342$, $c_{cor}^{(2)} = 0.187811$ with the curves for $BMMAP$ with much higher coefficients of correlation, $c_{cor}^{(1)} = 0.205982$, $c_{cor}^{(2)} = 0.402641$.

Analyzing the graphs for the probability $P_{loss}^{(imp)}$ in Figure 6, one can see that the behavior of the curves is the opposite of the behavior of the curves in Figures 3–5. From Figure 3, the probability $P_{loss}^{(2)}$ decreases with increasing the buffer space N; therefore, larger fraction of non-priority customers is admitted in the system and forms the longer queue. Thus, the rate of the flow of customers leaving the system due to impatience increases. The graphs also confirm that the correlation in the input flow can significantly affect the value $P_{loss}^{(imp)}$, especially for large values of N. The higher correlation implies the higher probability of loss due to impatience.

In general, it can be concluded from Figures 3–6 that ignorance of the correlation in the input flow can negatively affect the accuracy of evaluating the performance of real systems and lead to too optimistic estimates of the performance measures.

Example 3. *In this example, we computed the following performance measures: mean number of priority customers in the buffer,* $L^{(prior)}$*, mean number of all customers in the buffer,* L*, the loss probabilities* $P_{loss}^{(1)}$*,* $P_{loss}^{2)}$*,* P_{loss}*,* $P_{loss}^{(imp)}$*. We investigated the dependence of these performance measures on the input rate* λ *for the systems with BMMAPs having different coefficients of correlation.*

We considered three BMMAPs from Example 2. For convenience, we denote them as $BMMAP_1$*,* $BMMAP_2$ *and* $BMMAP_3$*, where the numbering is done in order of increasing of coefficients of correlation in the BMMAP. Distributions of the service time and timer are the same as in Example 2. The buffer capacity* $N = 10$ *and the probability that a non-priority customer leaves the system unserved when the timer expires is* $p = 0.4$*.*

Figures 7 and 8 depict the mean number, $L^{(prior)}$, of prior customers and the mean number, L, of all customers in the buffer as functions of the input rate λ for the $BMMAP$s with different coefficients of correlation.

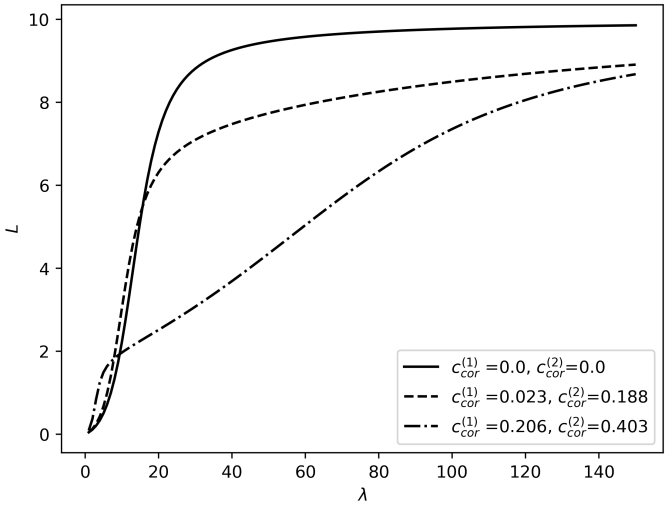

Figure 7. The mean number of customers in the buffer, L, as a function of the input rate λ for the $BMMAP$s with different coefficients of correlation.

As expected, the values of $L^{(prior)}$ and L increase when the input rate λ increases. More interesting is to analyze relative location of the curves for different coefficients of correlation. In the region $\lambda < 10$, a curve for the higher correlation is located above a curve for the lower correlation. It is worth recalling that in the considered example the service rate is $\mu = 10$, therefore, the region $\lambda < 10$ is associated with the region $\rho = \frac{\lambda}{\mu} < 1$. The relative location of the curves is changed when $\lambda > 10$ ($\rho > 1$). In this region, under the same value of λ the mean $L^{(prior)}$ and L are significantly less for the $BMMAP_3$ than for the $BMMAP_1$ and $BMMAP_2$. Also note a relatively low rate of increase in the case of the $BMMAP_3$. Thus, with a large correlation, we observe poor average buffer occupancy (due to the high loss probability).

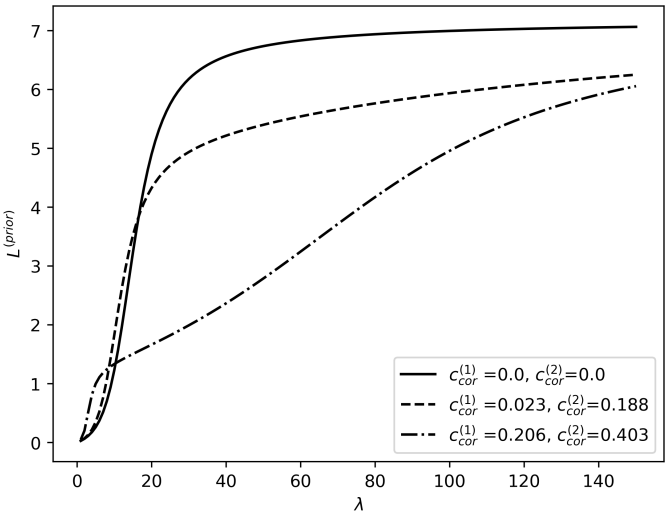

Figure 8. The mean number of priority customers in the buffer, $L^{(prior)}$, as a function of the input rate λ for the $BMMAP$s with different coefficients of correlation.

In order to investigate the behavior of $L^{(prior)}$ and L for $BMMAP_3$ in more detail, we examine the deviations of the number of priority customers and the total number of customers in the buffer from their average values. Figures 9 and 10 show the behavior of standard deviations σ and $\sigma^{(prior)}$.

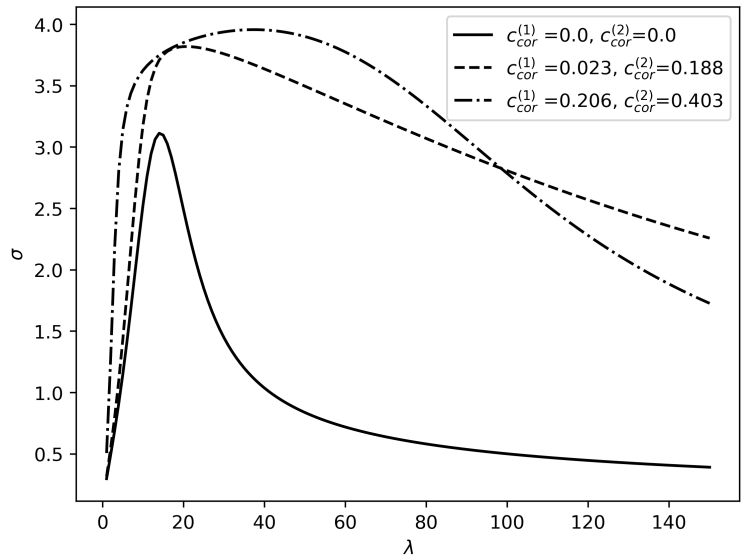

Figure 9. The standard deviations σ of the number of customers in the buffer for the $BMMAP$s with different coefficients of correlation.

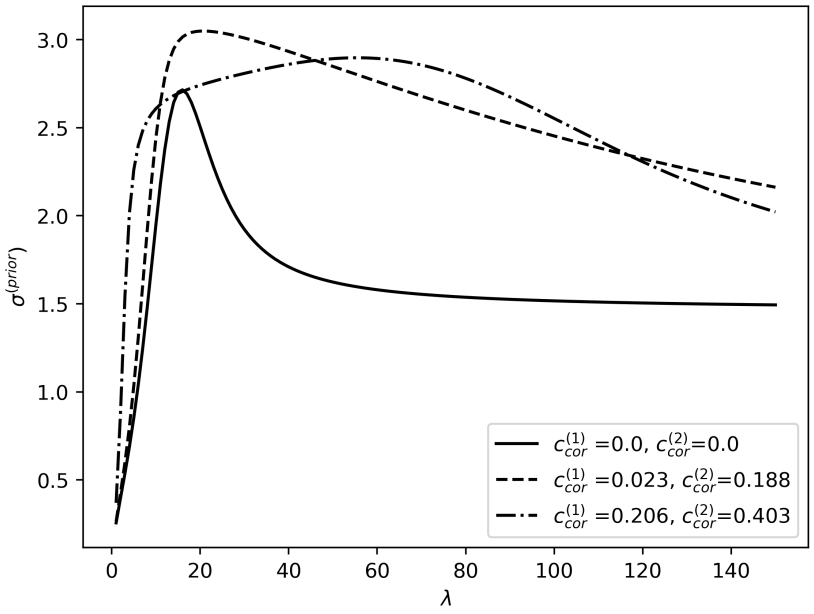

Figure 10. The standard deviations $\sigma^{(prior)}$ of the number of priority customers in the buffer for the $BMMAP$s with different coefficients of correlation.

As seen in the figures, the values of σ and $\sigma^{(prior)}$ for the $BMMAP_3$ retain large values in a wide region $\lambda > 10$. Taking this into account, we can explain the behavior of the means $L^{(prior)}$ and L by the fact that, under a high correlation in the $BMMAP$, periods of time when customers arrive rarely (and the server starves) alternate with the periods when customers arrive frequently (and many customers are lost due to the full occupancy of the buffer). This implies the non-uniform filling of the buffer. On average, the buffer may be filled weakly, but there is the large deviation of the number of customers in the buffer from the average value.

7. Conclusions

We considered a priority single-server queue with a finite buffer and two types of customers. Customers of low priority may become high priority customers after a random time having PH distribution. This queuing system may be helpful for modeling the work of emergency department in a hospital, contact center, inventory with perishable foods, etc. An analysis of the system was implemented under more realistic situations than found in the majority of the existing literature, and making assumptions about the arrival process and distributions of the service time and time until the change of the priority. Exact algorithmic results are obtained for stationary distributions of the number of customers of two types in the system and waiting time of priority customers. Numerical illustrations are presented, which show the dependencies of important performance indicators of the system in arrival rate and capacity of the buffer. Significant effect of correlation in arrival process is shown what evidently motivates the choice of the $BMMMAP$ as the model of arrival process, whereas the overwhelming majority of the existing research is devoted to the system with the stationary Poisson arrival process having zero correlation. The presented results are planned to be extended to the systems in which the increase of the priority can be dependent on the availability of some additional items (equipment, expendable materials, etc.), see, e.g., [24].

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

Funding: The publication has been prepared with the support of the RUDN University Program 5-100.

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

References

1. Bilodeau, B.; Stanford, D.A. Average Waiting Times in the Two-Class $M/G/1$ Delayed Accumulating Priority Queue. *arXiv* **2020**, arXiv:2001.06054.
2. Fajardo, V.A.; Drekic, S. Waiting Time Distributions in the Preemptive Accumulating Priority Queue. *Methodol. Comput. Appl. Probab.* **2017**, *19*, 255–284. [CrossRef]
3. Mojalal, M.; Stanford, D.A.; Caron, R.J. The lower-class waiting time distribution in the delayed accumulating priority queue. *INFOR Inf. Syst. Oper. Res.* **2020**, *58*, 60–86. [CrossRef]
4. Sharma, K.C.; Sharma, G.C. A delay dependent queue without preemption with general linearly increasing priority function. *J. Oper. Res. Soc.* **1994**, *45*, 948–953. [CrossRef]
5. Stanford, D.A.; Taylor, P.; Ziedins, I. Waiting time distributions in the accumulating priority queue. *Queueing Syst.* **2014**, *77*, 297–330. [CrossRef]
6. Xie, O.; He, Q.-M.; Zhao, X. Stability of a priority queueing system with customer transfers. *Oper. Res. Lett.* **2008**, *36*, 705–709. [CrossRef]
7. Xie, O.; He, Q.-M.; Zhao, X. On the stationary distribution of queue lengths in a multi-class priority queueing system with customer transfers. *Queueing Syst.* **2009**, *62*, 255–277. [CrossRef]
8. He, Q.M.; Xie, J.G.; Zhao, X.B. Stability conditions of a preemptive repeat priority $MMAP[N]/PH[N]/S$ queue with customer transfers (short version). In Proceedings of the 2009 Conference Proceedings on ASMDA (Advanced Stochastic Models and Data Analysis), Vilnius, Lithuania, 30 June–3 July 2009; pp. 463–467.
9. He, Q.-M.; Xie, J.; Zhao, X. Priority Queue with Customer Upgrades. *Nav. Res. Logist.* **2012**, *59*, 362–375. [CrossRef]
10. Xie, J.; Cao, P.; Huang, B.; Ong, M.E.H. Determining the conditions for reverse triage in emergency medical services using queuing theory. *Int. J. Prod. Res.* **2012**, *54*, 3347–3364. [CrossRef]
11. Xie, J.; Zhu, T.; Chao, A.K.; Wang, S. Performance analysis of service systems with priority upgrades. *Ann. Oper. Res.* **2017**, *253*, 683–705. [CrossRef]
12. Cao, P.; Xie, J. Optimal control of a multiclass queueing system when customers can change types. *Queueing Syst.* **2016**, *82*, 285–313.

13. Cildoz, M.; Ibarra, A.; Mallor, F. Accumulating priority queues versus pure priority queues for managing patients in emergency departments. *Oper. Res. Health Care* **2019**, *23*, 100224. [CrossRef]
14. Brown, L.; Gans, N.; Mandelbaum, A.; Sakov, A.; Shen, H.; Zeltyn, S.; Zhao, L. Statistical analysis of a telephone call center: A queueing-science perspective. *J. Am. Stat. Assoc.* **2005**, *100*, 36–50. [CrossRef]
15. Ramaswami, V. Independent Markov processes in parallel. *Comm. Statist.-Stoch. Models* **1985**, *1*, 419–432. [CrossRef]
16. Ramaswami, V.; Lucantoni, D. Algorithms for the multi-server queue with phase-type service. *Comm. Statist.-Stoch. Models* **1985**, *1*, 393–417.
17. He, Q.M. Queues with marked calls. *Adv. Appl. Probab.* **1996**, *28*, 567–587. [CrossRef]
18. Dudin, A.N.; Klimenok, V.I.; Vishnevsky, V.M. *The Theory of Queuing Systems with Correlated Flows*; Springer: Berlin/Heidelberg, Germany, 2020; 430p.
19. Dudin, A.N.; Klimenok, V.I.; Tsarenkov, G.V. A single-server queueing system with batch Markov arrivals, semi-Markov service, and finite buffer: Its characteristics. *Autom. Remote Control* **1985**, *63*, 1285–1297. [CrossRef]
20. Dudin, A.N.; Shaban, A.A.; Klimenok, V.I. Analysis of a $BMAP/G/1/N$ queue. *Int. J. Simul. Syst. Sci. Technol.* **2005**, *6*, 13–23.
21. Graham, A. *Kronecker Products and Matrix Calculus with Applications*; Ellis Horwood: Cichester, UK, 1981.
22. Kim, C.S.; Dudin, S.; Taramin, O.; Baek, J. Queueing system $MMAP/PH/N/N+R$ with impatient heterogeneous customers as a model of call center. *Appl. Math. Model.* **2013**, *37*, 958–976. [CrossRef]
23. Klimenok, V.I.; Kim, C.S.; Orlovsky, D.S.; Dudin, A.N. Lack of invariant property of Erlang $BMAP/PH/N/0$ model. *Queueing Syst.* **2005**, *49*, 187–213. [CrossRef]
24. Sun, B.; Dudin, A.; Dudin, S. Queueing System with Impatient Customers, Visible Queue and Replenishable Inventory. *Appl. Comput. Math.* **2018**, *17*, 161–174.