mathematics

MDPI

Article

A Priority Queue with Many Customer Types, Correlated Arrivals and Changing Priorities

Seokjun Lee [1], Sergei Dudin [2], Olga Dudina [2], Chesoong Kim [3,*] and Valentina Klimenok [2]

[1] Department of Management Information Systems, Sangji University, Wonju 26339, Korea; digitaldesign@sangji.ac.kr

[2] Laboratory of Applied Probabilistic Analysis, Belarusian State University, 4, Nezavisimosti Ave., 220030 Minsk, Belarus; dudin85@mail.ru (S.D.); dudina@bsu.by (O.D.); vklimenok@yandex.ru (V.K.)

[3] Department of Business Administration, Sangji University, Wonju 26339, Korea

* Correspondence: dowoo@sangji.ac.kr

Received: 15 July 2020; Accepted: 3 August 2020; Published: 5 August 2020

Abstract: A single-server queueing system with a finite buffer, several types of impatient customers, and non-preemptive priorities is analyzed. The initial priority of a customer can increase during its waiting time in the queue. The behavior of the system is described by a multi-dimensional Markov chain. The generator of this chain, having essential dependencies between the components, is derived and formulas for computation of the most important performance indicators of the system are presented. The dependence of some of these indicators on the capacity of the buffer space is illustrated. The profound effect of the phenomenon of correlation of successive inter-arrival times and variance of the service time is numerically demonstrated. Results can be used for the optimization of dispatching various types of customers in information transmission systems, emergency departments and first aid stations, perishable foods supply chains, etc.

Keywords: priority system; marked Markov arrival process; phase-type distribution; change of the priority; dispatching

1. Introduction

Queueing theory is successfully applied in various fields of human activity for optimization of the consumption and scheduling certain restricted resources and provisioning the high quality of service. The overwhelming majority of the existing literature in this theory is devoted to the systems with homogeneous customers; see, e.g., [1]. Because real-world customers are very often heterogeneous in many respects, new developments in the analysis of queues with heterogeneous customers are of great importance. The heterogeneity of the customers with respect to the required resources, level of service, and their economical or social value causes the necessity of the optimal management of their service. Such management can be implemented, e.g., in various generalizations of polling disciplines, processor sharing, applying versatile priority schemes. For some references, see, e.g., [2]. Priority schemes assume the assignment of a certain priority to each class of customers and providing the advantage of access to the restricted resource (we will call this resource as a server) to available customers having the highest priority. Static priorities suggest that once the priorities are assigned, a low priority customer does not have any chance to start service until the server finishes service of all high priority customers presenting in the system. This may cause a low priority customer to wait in the queue much longer than the just arrived high priority customer. To avoid this evident unfairness to the low priority customers, dynamic priorities were taken into consideration. The dynamic priority assumes, e.g., that the low priority customers obtain the chance to start service in presence of high priority customers when: (i) the queue of the low priority customers exceeds some threshold values, see, e.g., [3–6]; or (ii) some relation between the queue lengths of priority and non-priority customers is fulfilled, see, e.g., [7]; or (iii) a certain limit of the number of high priority customers that can overtake the low priority customers is exceeded, see, e.g., [8]. The use of dynamic priorities allows to essentially improve the quality of the system operation. The shortcomings of such priorities are: (i) the necessity to permanently monitor the values of the queue length of different classes of customers what is not always possible (or costly) in some real-world systems and (ii) dependence of the waiting time of a concrete low priority customer on the rate of future arrival of other low priority customers. Another opportunity of providing more fair access to low priority customers is assumed in the models where a low priority customer can become higher priority customer after a certain period of waiting in the buffer. A currently popular model assumes that the low priority customers accumulate a priority during the stay in the queue. The accumulation of the priority may be described as some function, e.g., linear or piece-wise linear function, of the time spent by the customer in a queue. The rate of the increase of the priority may depend on the class to which the customer belongs. Such a type of model was considered, e.g., in the papers [9–14]. The main interest to the queues with accumulating priorities stems from their applicability to modeling operation of emergency departments of hospitals. Arriving customers (patients) are preliminarily sorted (triaged) into several groups according to the severity of the patient's condition. However, during the waiting for treatment by the doctors, a state of health of some patient, which was initially classified as not requiring very urgent treatment, can become essentially worse and this patient has to be transferred to the group of very urgent patients. Because in the described situation the increase of the priority of a customer is not defined by some deterministic function of the elapsed waiting time, another type of model, with the randomized change of a priority, exists in the literature. This type of model was considered, e.g., in [15,16] and the recent paper [17]. The table presenting the state of art in the analysis of queues with priority change after some random amount of time is presented in [17]. It follows from that table that only a few papers consider the models where the arrival processes of customers of different types are not defined by the stationary Poisson arrival process, while it is already well recognized that the flows in many real systems and networks are poorly described by the stationary Poisson arrival process. The rare exceptions, when a more complicated arrival process is considered, are the papers [18–20]. In all these papers, an arbitrary number of priority classes is suggested. In [18], it is assumed that all the flows, except the flow having the highest priority, are described by the stationary Poisson arrival process. The arrival flow of customers having the highest priority is described by a much more general Markov arrival process (MAP); see, e.g., [21–23] for more details. In [19,20], the arrival flow is described by even more general marked Markov arrival process ($MMAP$). The $MMAP$, as the essential generalization of the MAP to the case of heterogeneous customers, was introduced in [24]. The models with the MAP or $MMAP$ are much more difficult for analysis than the models with the stationary Poisson arrival process. This explains why only some bounds and tail distributions were obtained in [18] and only the problem of establishing the ergodicity condition (but not the problem of computation of the stationary distribution of the system states and performance measures) is solved in [19,20]. The problem of computation of the stationary distribution of the system states is successfully solved in [17] but only for two classes of customers. The advantage of our paper over [17] is that we suggest any finite number R of priority classes. The arrival process is described by the $MMAP$. The system has a finite buffer and any arriving customer is admitted to the buffer if it is not full. If the buffer is full while some waiting customers have lower priority than the arriving customer, the arriving customer pushes out from the buffer a customer having the lowest priority among the presenting ones. During the stay in the buffer, after an exponentially distributed time, any customer can increase its priority. The service time has a phase-type distribution. After the service completion, the next service is provided for a customer with the highest priority among the presented in the buffer.

It is worth mentioning that the problem of assigning the priorities to different classes of customers is often closely related to the problem of the account of possible impatience of customers from different classes, e.g., if customers of two types are almost equally valuable for the system, the more impatient customers should be given higher priority (and the possibility to increase the priority during the waiting time in a buffer) to avoid the loss of the customer and possible starvation (and poor utilization) of the server in the future. In our model, we pay significant attention to the account of impatience.

Besides the above-mentioned popular model of treatment of patients in a hospital emergency department, we mention the following examples of potential applications of the considered model to the analysis and optimization of real-world systems.

(1) Let us consider the operation of an information transmission channel. Several kinds of information having approximately the same transmission times, but having different importance for the system and different tolerance to the delay are transmitted through this channel. Initially, the priorities can be assigned to the different types of information depending on their importance. However, to avoid the loss of low priority and delay-sensitive information units (and possible under-utilization of the channel in the future), it makes sense to allow a low priority information unit whose obsolescence time is almost expired to become a high priority information unit and receive the service soon.

(2) Let us consider the operation of a first aid station. The station has to accept the calls for help, categorize the urgency of the required help, and to manage the assignment of the necessary ambulance car for providing help, e.g., in the Republic of Belarus (as of 1 January 2020), there are three possible categories of the urgency of the required help.

    (a) An emergency call—when a patient suddenly has diseases, conditions and (or) exacerbation of chronic diseases that pose a threat to the patient's life and (or) others requiring emergency medical intervention;

    (b) An urgent call is associated with a sharp deterioration in the patient's health status when it is not possible to clarify the reasons for treatment;

    (c) A less urgent call—when the patient suddenly has diseases, conditions, and/or exacerbation of chronic diseases without obvious signs of a threat to the patient's life, requiring urgent medical intervention.

Accordingly, the emergency call has the highest priority, the urgent call has the middle priority, and the less urgent calls have the lowest priority. However, along with this categorization and establishing the priority in service, there exist strict standards for starting the provisioning of help. A dispatcher has to assign an ambulance car for providing help to patients before the fixed deadlines. In Minsk, the capital of the Republic of Belarus, these standards are fixed as four minutes for the emergency call, fifteen minutes for the urgent call, and sixty minutes for the less urgent call. Violation of this standard is punished. In this example, the service time can be interpreted as a time between the sequential release of ambulance cars. The service time essentially depends on the number of available cars and medical teams. The results of the analysis of the model given in our paper can be useful for the optimization of the work of the described first aid station via a proper choice of the number of ambulance teams to guarantee the required quality of service.

Methodological value of the paper consists of presenting a way for analysis of various transitions of a set of interacting Markov processes, which define the dynamics of the number of customers of several types in the system, caused by new customers of various types arrival, service completion, departure due to impatience, changing the priority, and pushing out the low priority customers in the case of the buffer overflow.

The organization of the text is as follows. In Section 2, the mathematical model is described and graphically illustrated. The multi-dimensional Markov chain including as components the total number of customers in the system, the states of the underlying processes of customers arrival and service, and the number of customers of all types presenting in the system is defined in Section 3. The set of matrices defining the probabilities or intensities of transitions of the number of customers of all types are given and the generator of the Markov chain is written down. Formulas for computation of the main performance measures of the system are presented in Section 4. The numerical example illustrating the dependence of performance measures of the system on the capacity of the buffer is presented in Section 5. The importance of account of a complicated pattern of arrival process and variance of the service time is demonstrated there. Section 6 concludes the paper.

2. Mathematical Model

We consider a single-server queuing system where service is provided to R types of customers. The structure of the system is presented in Figure 1.

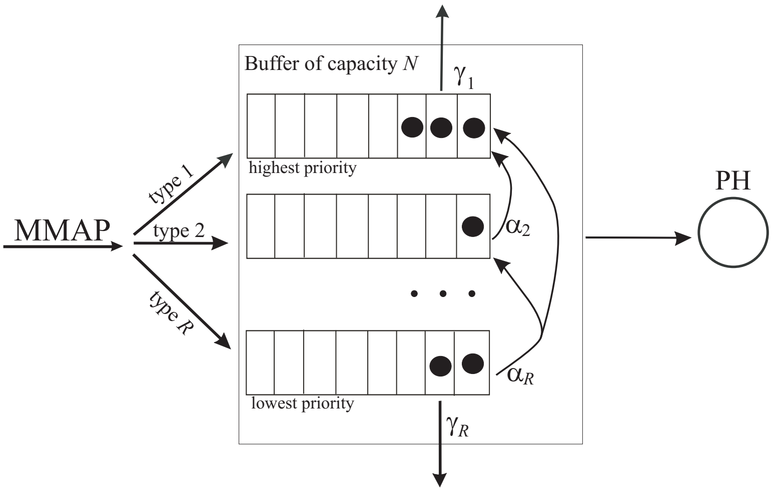

Figure 1. Structure of the system.

The customer arrival process is assumed to be defined by the *MMAP* (see, e.g., [24]). As the recent papers where the queuing models with the *MMAP* are analyzed, we can mention, e.g., [25–27].

Customer arrivals in the *MMAP* can occur at the moments of the transitions of the irreducible continuous-time Markov chain ν_t, $t \geq 0$, having a state space $\{1, 2, ..., W\}$. The *MMAP* is completely described by the square matrices D_0, D_r, $r = \overline{1, R}$. Hereinafter, the denotation like $r = \overline{1, R}$ means that the parameter r takes values $\{1, \ldots, R\}$.

The matrix D_r defines the transition intensities of the underlying process ν_t that lead to arrival of a type-r customer, $r = \overline{1, R}$. The non-diagonal entries of the matrix D_0 define the transition intensities of the underlying process that do not lead to any arrival. The moduli of the diagonal entries of the matrix D_0 define the intensity of the the process ν_t departure of from its states. The matrix $D(1) = D_0 + D$ where $D = \sum\limits_{r=1}^{R} D_r$ is the generator of the underlying process.

The mean arrival rate λ is defined by $\lambda = \boldsymbol{\theta} D \mathbf{e}$ where $\boldsymbol{\theta}$ is the invariant probability row vector of the underlying process. This vector is computed as the unique solution for the finite system $\boldsymbol{\theta} D(1) = \mathbf{0}$, $\boldsymbol{\theta}\mathbf{e} = 1$. Hereinafter, $\mathbf{e}$ denotes a column vector of appropriate size consisting of 1s and $\mathbf{0}$ denotes a row vector consisting of zeroes.

The mean rate λ_r of type-r customers arrival is computed as $\lambda_r = \boldsymbol{\theta} D_r \mathbf{e}$, $r = \overline{1, R}$. The squared coefficient of variation c_{var}^2 of the intervals between successive arrivals is given by $c_{var}^2 = 2\lambda\boldsymbol{\theta}(-D_0)^{-1}\mathbf{e} - 1$. The coefficient of correlation c_{cor} of two successive intervals between arrivals is given by

$$c_{cor} = (\lambda\boldsymbol{\theta}(-D_0)^{-1} D (-D_0)^{-1}\mathbf{e} - 1)/c_{var}^2.$$

The system has the finite common buffer space for storing the customers that arrive when the server is busy. The capacity of the buffer is N, $N \geq 1$. Therefore, the total number of customers of all types, which can stay in the system simultaneously, is restricted by the number $N + 1$. If a customer of any type arrives when the server is idle, the customer immediately starts processing by the server (service). If the server is busy but the buffer is not full, the customer of any type is placed into the buffer dedicated to this type of customers. There is no specific restriction on the capacity of the dedicated buffers, except that the total number of the customers staying in all these buffers always does not exceed the capacity N.

Customers of different types have different priorities. The priority defines the fate of the customer if it arrives when the buffer is full and the order of picking up the customers from the buffer when the server finishes service. We assume that type-r, $r = \overline{1, R}$, customers have the non-preemptive priority over type-l customers, $l = \overline{r+1, R}$. This means the following.

(1) If during the arrival of a type-r customer the server is busy and the number of customers in the buffer is N and there are no type-l, $l = \overline{r+1, R}$, customers, the arriving customer is lost. If there are type-l, $l = \overline{r+1, R}$, customers in the buffer then, with the probability q, the arriving customer is accepted to the buffer and one of the customers with the lowest priority among the presenting in the buffer is lost. With the complimentary probability $1 - q$, the arriving customer is lost despite the presence in the system of customers with lower priority.

(2) Type-1 customers have the highest priority among all types of customers and if type-1 customers present in the buffer at a service completion epoch, one of these customers starts service, ..., type R customers have the lowest priority. A customer of such a type has a chance to start service only if customers of types $1, 2, \ldots, R-1$ are absent in the buffer. Service of any customer cannot be preempted (interrupted) in the case of an arrival of a customer having a higher priority.

We assume that during the stay in the system, each customer of type-r, $r = \overline{2, R}$, can increase its priority. It means that after exponentially distributed time with the parameter α_r a type-r customer becomes a type-l customer with the probability $p_{r,l}$, $l = \overline{1, r-1}$, independently of other customers. Here, $\sum\limits_{l=1}^{r-1} p_{r,l} = 1$, $r = \overline{2, R}$.

It is worth noting that more popular in the existing literature assumption is that only the head-of-the-line customer of each type can make a jump to the end of the queue of higher priority customers. We assume that each customer of any type can jump to higher priority class, independently of other customers. This means that not only the head-of-the-line customer has a clock counting the time till the jump, but each customer (not of the highest priority) has its own clock. Our assumption seems more realistic in some potential applications, e.g., health of any patient, not only the head-of-the-line patient in emergency department modeling example, can suddenly become worse. The same is true in applications where various information units become obsolete independently of the other units or different perishable foods have independent spoiling times. Note also, that, using the slight modification of some matrix blocks defined and constructed in the next section, the presented results can be extended to the models with the head-of-the-line customer priority jumps as well.

Customers staying in the buffer are impatient and can leave the system without service, independently of other customers, if the waiting time is too long. A type-r customer leaves the system without service after an exponentially distributed patience time with the parameter γ_r , $\gamma_r \geq 0$. Let us denote $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \ldots, \gamma_R)$. If the customer changes the priority, its patience time starts from the early beginning with the parameter corresponding to the new priority.

We assume that the service time of any type customer has a PH distribution with the underlying Markov process $m_t, t \geq 0$, having a finite state space $\{1, \ldots, M, M+1\}$ and the irreducible representation $(\boldsymbol{\beta}, S)$, see, [28]. We denote $\mathbf{S_0} = -S\mathbf{e}$. The mean service time is given by $b_1 = \boldsymbol{\beta}(-S)^{-1}\mathbf{e}$. The mean service rate can be compute as $\mu = b_1^{-1}$.

If during the service completion epoch there are customers in the buffer, the first customer among having the highest priority starts service. Otherwise, the server remains idle until the next arrival moment.

3. Process of the System States

The behavior of the system under study can be described by the regular irreducible continuous-time Markov chain

$$\xi_t = \{n_t, \nu_t, m_t, \eta_t^{(1)}, \dots, \eta_t^{(R)}\},\ t \geq 0,$$

where, during the epoch t,

- n_t is the number of customers in the system, $n_t = \overline{0, N+1}$;
- ν_t is the state of the underlying process of the $MMAP$, $\nu_t = \overline{1, W}$;
- m_t is the state of the underlying process of PH service process, $m_t = \overline{1, M}$;
- $\eta_t^{(r)}$ is the number of type-r customers in the buffer, $\eta_t^{(r)} = \overline{0, n_t - 1}$, $r = \overline{1, R}$, $\sum\limits_{r=1}^{R} \eta_t^{(r)} = n_t - 1$, $n_t > 1$.

To investigate the Markov chain $\xi_t,\ t \geq 0$, let us enumerate its states in the direct lexicographic order of the components ν_t and m_t, and in the reverse lexicographic order of the components $\eta_t^{(1)}, \dots, \eta_t^{(R)}$.

The most technically difficult and important part of the research is the analysis of the transitions of the process of the number of different type customers in the buffer. Let us firstly consider the process $\zeta_t^{(n)} = \{\eta_t^{(1)}, \dots, \eta_t^{(R)}\}$, $t \geq 0$, $\eta_t^{(r)} = \overline{0, n}$, $r = \overline{1, R}$, $\sum\limits_{r=1}^{R} \eta_t^{(r)} = n$. The process $\zeta_t^{(n)}$ describes the transitions of the number of different types customers in the buffer when the total number of customers in the buffer is n. First, we present the algorithms for computing the set of the matrices that define the transition probabilities or transition intensities of the process $\zeta_t^{(n)}$ at the moments of the changes, due to various reasons, of the components of this process when n, $n = \overline{1, N}$, customers stay in the buffer.

Lemma 1.

(a) *Let $L_n(\gamma)$ be the matrix the entries of which define the intensities of transitions when some customer leaves the buffer due to impatience.*

The matrices $L_n(\gamma)$, $n = \overline{1, N}$, can be computed by the following way:

1. *Calculate the matrices $L_n^{(l)}(\gamma)$ using the recursive formulas:*

$$L_n^{(0)}(\gamma) = n\gamma_R,$$

$$L_n^{(l)}(\gamma) = \begin{pmatrix} n\gamma_{R-l}I & O & \cdots & O \\ L_1^{(l-1)}(\gamma) & (n-1)\gamma_{R-l}I & \cdots & O \\ O & L_2^{(l-1)}(\gamma) & \cdots & O \\ \vdots & \vdots & \ddots & \vdots \\ O & O & \cdots & \gamma_{R-l}I \\ O & O & \cdots & L_n^{(l-1)}(\gamma) \end{pmatrix},\ l = \overline{1, R-1}.$$

Here and after, I is the identity matrix and O is a zero matrix of an appropriate dimension;

2. *Calculate the matrices $L_n(\gamma)$ as $L_n(\gamma) = L_n^{(R-1)}(\gamma)$, $n = \overline{1, N}$.*

(b) *Let $Y_n = Y_n(H)$ be the matrix the entries of which define the intensities of transitions that occur when some customer increases its priority. Here, the matrix H defines the intensities of priorities increasing and has the following form:*

$$H = \begin{pmatrix} 0 & 0 & 0 & \cdots & 0 & 0 \\ \alpha_2 & 0 & 0 & \cdots & 0 & 0 \\ p_{3,1}\alpha_3 & p_{3,2}\alpha_3 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ p_{R-1,1}\alpha_{R-1} & p_{R-1,2}\alpha_{R-1} & p_{R-1,3}\alpha_{R-1} & \cdots & 0 & 0 \\ p_{R,1}\alpha_R & p_{R,2}\alpha_R & p_{R,3}\alpha_R & \cdots & p_{R,R-1}\alpha_R & 0 \end{pmatrix}.$$

Calculation of the matrices $Y_n(H)$, $n = \overline{1,N}$, can be performed as follows:

1. *Calculate the matrices H_j, $j = \overline{1,R-2}$, which are obtained by deletion of $R-2-j$ first rows and columns from the matrix H.*
2. *Calculate the matrices $Z_n^{(l)}(H_j)$ using the recursive formulas:*

$$Z_n^{(0)}(H_j) = nh^j_{r_j,1},\ n = \overline{1,N},\ j = \overline{1,R-2},$$

$$Z_n^{(l)}(H_j) = \begin{pmatrix} nh^j_{r_j-l,1}I & O & \cdots & O \\ Z_1^{(l-1)}(H_j) & (n-1)h^j_{r_j-l,1}I & \cdots & O \\ O & Z_2^{(l-1)}(H_j) & \cdots & O \\ \vdots & \vdots & \ddots & \vdots \\ O & O & \cdots & h^j_{r_j-l,1}I \\ O & O & \cdots & Z_n^{(l-1)}(H_j) \end{pmatrix},$$

$$l = \overline{1,\ldots,r_j-2},\ n = \overline{1,N},\ j = \overline{1,R-2},$$

where $h^j_{a,b}$ is the (a,b)th entry of the matrix H_j and r_j is the number of rows of the matrix H_j.

3. *Calculate the matrices $X_n^{(l)}(H_j)$ using the recursive formulas:*

$$X_n^{(0)}(H_j) = h^j_{1,r_j},\ n = \overline{0,N-1},\ j = \overline{1,R-2},$$

$$X_n^{(l)}(H_j) = \begin{pmatrix} h^j_{1,r_j-l}I & X_0^{(l-1)}(H_j) & O & \cdots & O & O \\ O & h^j_{1,r_j-l}I & X_1^{(l-1)}(H_j) & \cdots & O & O \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ O & O & O & \cdots & h^j_{1,r_j-l}I & X_n^{(l-1)}(H_j) \end{pmatrix},$$

$$l = \overline{1,r_j-2},\ n = \overline{0,N-1},\ j = \overline{1,R-2}.$$

4. *Calculate the matrices $Z_n(H_j) = Z_n^{(r_j-2)}(H_j)$, $n = \overline{1,N}$, and $X_n(H_j) = X_n^{(r_j-2)}(H_j)$, $n = \overline{0,N-1}$, $j = \overline{1,R-2}$.*

5. *Calculate the matrices $Y_n^{(j)}$, $n = \overline{1,N}$, using the recursive formulas:*

$$Y_n^{(0)} = \begin{pmatrix} 0 & nH_{M-1,M} & 0 & \cdots & 0 & 0 \\ H_{M,M-1} & 0 & (n-1)H_{M-1,M} & \cdots & 0 & 0 \\ 0 & 2H_{M,M-1} & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & H_{M-1,M} \\ 0 & 0 & 0 & \cdots & nH_{M,M-1} & 0 \end{pmatrix},$$

$$Y_n^{(j)} = \begin{pmatrix} O & nX_0(H_j) & O & \cdots & O & O \\ Z_1(H_j) & Y_1^{(j-1)} & (n-1)X_1(H_j) & \cdots & O & O \\ O & Z_2(H_j) & Y_2^{(j-1)} & \cdots & O & O \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ O & O & O & \cdots & Y_{n-1}^{(j-1)} & 1X_{n-1}(H_j) \\ O & O & O & \cdots & Z_n(H_j) & Y_n^{(j-1)} \end{pmatrix},$$

$$j = \overline{1, R-2}.$$

6. *Calculate the matrices $Y_n(H)$ as $Y_n(H) = Y_n^{(R-2)}$, $n = \overline{1,N}$.*

(c) *Let $A_n(\mathbf{h})$, $n = \overline{0, N-1}$, be the matrix the entries of which define the transition probabilities at the moment when a new customer arrives to the system and the system capacity is not exhausted (there are n, $0 \leq n < N$, customers in the buffer). Here, the row vector $\mathbf{h}$ has the following form $\mathbf{h} = (h_1, h_2, \ldots, h_R)$ where h_r is the probability that the arrived to the system customer has type-r, $r = \overline{1,R}$.*

Computation of the matrices $A_n(\mathbf{h})$ can be performed as follows:

$A_0(\mathbf{h}) = \mathbf{h}$ and $A_n(\mathbf{h}) = A_n^{(R-2)}(\mathbf{h})$ where the matrices $A_n^{(l)}(\mathbf{h})$ of block size $(n+1) \times (n+2)$, $n = \overline{1, N-1}$, are recursively computed as

$$A_n^{(0)}(\mathbf{h}) = \begin{pmatrix} h_{R-1} & h_R & 0 & \cdots & 0 & 0 \\ 0 & h_{R-1} & h_R & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & h_{R-1} & h_R \end{pmatrix},$$

$$A_n^{(l)}(\mathbf{h}) = \begin{pmatrix} h_{R-l-1} & \bar{\mathbf{h}}^{(l)} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ \mathbf{0}^T & h_{R-l-1}I & A_1^{(l-1)} & O & \cdots & O & O \\ \mathbf{0}^T & O & h_{R-l-1}I & A_2^{(l-1)} & \cdots & O & O \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{0}^T & O & O & O & \cdots & h_{R-l-1}I & A_n^{(l-1)} \end{pmatrix},$$

$$l = \overline{1, R-2},$$

where the vectors $\bar{\mathbf{h}}^{(l)}$ are defined as $\bar{\mathbf{h}}^{(l)} = (h_{R-l}, h_{R-l+1}, \ldots, h_R)$, $l = \overline{1, R-2}$.

(d) *Let E_n^-, $n = \overline{1,N}$, be the matrix the entries of which define the transition probabilities at the moment when a customer with the maximal (among currently presenting in the system) priority is chosen for service.*

The matrices E_n^- can be computed as

$$E_1^- = (\underbrace{1, 1, \ldots, 1}_{R})^T,$$

$$E_n^- = \begin{pmatrix} \multicolumn{2}{c}{I_{K_R^{(n)}}} \\ O_{K_{R-1}^{(n)} \times (K_R^{(n)} - K_{R-1}^{(n)})} & I_{K_{R-1}^{(n)}} \\ \multicolumn{2}{c}{\dots} \\ O_{K_2^{(n)} \times (K_R^{(n)} - K_2^{(n)})} & I_{K_2^{(n)}} \\ O_{K_1^{(n)} \times (K_R^{(n)} - K_1^{(n)})} & I_{K_1^{(n)}} \end{pmatrix}, n = \overline{2, N},$$

where

$$K_r^{(n)} = \binom{n+r-2}{r-1}, r = \overline{1, R}.$$

Here, $\binom{n+r-2}{r-1} = C_{n+r-2}^{r-1}$ *is the binomial coefficient.*

(e) *Let the entries of the square matrix* $\hat{E}_r$, $r = \overline{1, R}$, *of size* $\binom{N+R-1}{R-1}$ *define the transition probabilities at the moment when a type-r customer arrives at the system when there are N customers in the buffer and the arriving customer tries to force out a customer with a lower priority from the buffer. All entries in each row of this matrix are equal to zero except one entry which is equal to 1. We assume that each row and column of the matrix* $\hat{E}_r$ *correspond to some state* $\{\eta_1, \eta_2, \dots, \eta_R\}$ *of the process* ζ_t, $t \geq 0$. *Note, that all states of the process* ζ_t, $t \geq 0$, *are enumerated in the reverse lexicographical order of components* $\eta_t^{(1)}, \dots, \eta_t^{(R)}$. *For example, the first row and column of the matrix* $\hat{E}_r$ *correspond to the state* $\{N, 0, 0, \dots, 0\}$, *the second row and column correspond to the state* $\{N-1, 1, 0, \dots, 0\}$,..., *the last row and column correspond to the state* $\{0, 0, 0, \dots, N\}$. *In the row of the matrix* $\hat{E}_r$ *that corresponds to the state* $\{\eta_1, \eta_2, \dots, \eta_R\}$, *the entry 1 is located in the column that corresponds to the same state* $\{\eta_1, \eta_2, \dots, \eta_R\}$ *only in the case if* $\eta_l = 0$ *for all* l, $R \geq l > r$. *In this case, the arriving type-r customer is lost, because the customers with lower priority are absent in the buffer. If* $\eta_l > 0$ *for some* l, $R \geq l > r$ *and* r^* *is a maximum of such values* l, *then the entry 1 is located in the column that corresponds to the state* $\{\eta_1, \dots, \eta_{r-1}, \eta_r + 1, \eta_{r+1}, \dots, \eta_{r*-1}, \eta_{r*} - 1, 0, \dots, 0\}$. *In this case, the customer of type-*r^* *has the lowest priority among the customers presenting in the system and an arriving type-r customer forces out one type-*r^* *customer which departs from the system (is lost).*

Proof. The derivation of the form of the matrices that describe the transitions of the process $\zeta_t^{(n)}$, $t \geq 0$, is quite complicated and cumbersome. In derivations, we used some ideas of the paper [29]. To explain the scheme of the derivation of the form of the presented matrices, we show here how to compute the matrices $L_n(\gamma)$, $n = \overline{1, R}$, the entries of which define the intensities of transitions of the components of the process $\zeta_t^{(n)}$, $t \geq 0$, when some customer leaves the buffer due to impatience. The rest of the matrices that define the intensities of transition of the components of the process $\zeta_t^{(n)}$, $t \geq 0$, can be obtained by the same way based on the careful account of possible transitions.

Computation of the matrices $L_n(\gamma)$ can be performed as follows. Let us introduce the matrices $L_n^{(l)}(\gamma)$ of the transition intensities of the components $n_t^{(R)}, \dots, n_t^{(R-l)}$ at the moment when there are n customers in the buffer and one of the customers leaves it due to impatience conditional on the fact that all customers have types $R, R-1, \dots, R-l$, where $l = \overline{0, R-1}$.

It is clear, that for $l = 0$, the matrices $L_n^{(0)}(\gamma)$ have the scalar form $L_n^{(0)}(\gamma) = n\gamma_R$, because all n customers are of type-R in this situation.

Let us consider the matrix $L_n^{(1)}$. This matrix defines the transition intensities of the components $n_t^{(R)}, n_t^{(R-1)}$ at the moment when there are n customers in the buffer and one of the customers leaves it due to impatience conditional on the fact that all customers have types R or $R-1$. Taking into account the reverse lexicographic order of components, by definition the first row of the matrix $L_n^{(1)}(\gamma)$ corresponds to the state where all n customers are of type-$(R-1)$, the second row corresponds to the state where $n-1$ customers are of type-$(R-1)$ and one customer is of type-R, etc., the last row corresponds to the state where all n customers are of type-R. After the customer leaves the system, the number of customers in the buffer decreases by 1. Thus, the first column of the matrix $L_n^{(1)}(\gamma)$

corresponds to the state where all $n-1$ customers are of type-$(R-1)$, the second column corresponds to the state where $n-2$ customers are of type-$(R-1)$ and one customer is of type-R, etc., the last column corresponds to the state where all $n-1$ customers are of type-R. Taking into account these considerations, it is easy to verify that the matrix $L_n^{(1)}(\gamma)$ of size $(n+1)\times n$ has the form

$$L_n^{(1)}(\gamma) = \begin{pmatrix} n\gamma_{R-1} & 0 & \cdots & 0 \\ \gamma_R & (n-1)\gamma_{R-1} & \cdots & 0 \\ 0 & 2\gamma_R & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \gamma_{R-1} \\ 0 & 0 & \cdots & n\gamma_R \end{pmatrix},$$

or

$$L_n^{(1)}(\gamma) = \begin{pmatrix} n\gamma_{R-1} & 0 & \cdots & 0 \\ L_1^{(0)}(\gamma) & (n-1)\gamma_{R-1} & \cdots & 0 \\ 0 & L_2^{(0)}(\gamma) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \gamma_{R-1} \\ 0 & 0 & \cdots & L_n^{(0)}(\gamma) \end{pmatrix}.$$

Using the same reasonings, it can be shown that the matrix $L_n^{(l)}(\gamma)$ of block size $(n+1)\times n$ has the following form

$$L_n^{(l)}(\gamma) = \begin{pmatrix} n\gamma_{R-l}I & O & \cdots & O \\ L_1^{(l-1)}(\gamma) & (n-1)\gamma_{R-l}I & \cdots & O \\ O & L_2^{(l-1)}(\gamma) & \cdots & O \\ \vdots & \vdots & \ddots & \vdots \\ O & O & \cdots & \gamma_{R-l}I \\ O & O & \cdots & L_n^{(l-1)}(\gamma) \end{pmatrix}, \ l=\overline{2,R-1}.$$

It is clear that the required matrices $L_n(\gamma)$ can be computed as $L_n(\gamma) = L_n^{(R-1)}(\gamma)$, $n=\overline{1,N}$. This proves the proposed formulas for computation of the matrices $L_n(\gamma)$. □

Remark 1. *Derivation of the form of the matrices defined in Lemma 1 creates an opportunity to analyze not only the system under study in this paper but also many other queueing systems with a finite buffer and many types of customers having different priorities.*

Let us introduce the following notation:

- $\otimes$ and $\oplus$ indicate the symbols of the Kronecker product and sum of matrices, respectively, see [30];
- $\mathbf{h}_r = (\underbrace{0,\ldots,0}_{r-1},1,\underbrace{0,\ldots,0}_{R-r})$, $r=\overline{1,R}$;
- $\hat{I}_n = -\text{diag}\{Y_n\mathbf{e} + L_n\mathbf{e}\}$, $n=\overline{1,N}$, where diag$\{\ldots\}$ denotes the diagonal matrix with the diagonal entries defined by the vector in the brackets;
- $K_n = \binom{n+R-1}{R-1}$, $n=\overline{1,N}$.

By analyzing all possible transitions of the Markov chain $\xi_t, t \geq 0$, during an interval of infinitesimal length and rewriting the intensities of these transitions in the block matrix form, we obtain the following result.

Theorem 1. *The infinitesimal generator Q of the Markov chain $\xi_t,\ \ t \geq 0$, has the following block-tridiagonal structure*

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

The non-zero blocks are defined as follows:

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

$$Q_{1,1} = D_0 \oplus S,$$

$$Q_{n,n} = D_0 \oplus S \otimes I_{K_{n-1}} + I_{WM} \otimes (Y_{n-1} + \hat{I}_{n-1}),\ n = \overline{2,N},$$

$$Q_{N+1,N+1} = (D_0 \oplus S) \otimes I_{K_N} + I_{WM} \otimes (Y_N + \hat{I}_N) + (1-q)\sum_{r=1}^{R} D_r \otimes I_{MK_N} +$$

$$q\sum_{r=1}^{R} D_r \otimes I_M \otimes \hat{E}_r,$$

$$Q_{0,1} = \sum_{r=1}^{R} D_r \otimes \boldsymbol{\beta},$$

$$Q_{n,n+1} = \sum_{r=1}^{R} D_r \otimes I_M \otimes A_{n-1}(\mathbf{h}_r),\ n = \overline{1,N},$$

$$Q_{1,0} = I_W \otimes \mathbf{S_0},$$

$$Q_{n,n-1} = I_W \otimes \mathbf{S_0}\boldsymbol{\beta} \otimes E_{n-1}^{-} + I_{WM} \otimes L_{n-1}(\gamma),\ n = \overline{1,N+1}.$$

The Markov chain $\xi_t,\ t \geq 0$, is an irreducible and has a finite state space. Therefore, the stationary probabilities of the system states

$$\pi(n,\nu,m,\eta^{(1)},\dots,\eta^{(R)}) =$$

$$= \lim_{t\to\infty} P\{n_t = n, \nu_t = \nu, m_t = m, \eta_t^{(1)} = \eta^{(1)}, \dots, \eta_t^{(R)} = \eta^{(R)}\}$$

always exist.

Let us form the row vectors $\boldsymbol{\pi}_n,\ n = \overline{0,N+1}$, of these probabilities which are enumerated in the reverse lexicographic order of the components $\eta_t^{(1)},\dots,\eta_t^{(R)}$ and the direct lexicographic order of the components ν_t and m_t.

It is well known that the probability vectors $\boldsymbol{\pi}_n,\ n = \overline{0,N+1}$, satisfy the following system of linear algebraic equations:

$$(\boldsymbol{\pi}_0, \boldsymbol{\pi}_1, \dots, \boldsymbol{\pi}_{N+1})Q = \mathbf{0}, \tag{1}$$

$$(\boldsymbol{\pi}_0, \boldsymbol{\pi}_1, \dots, \boldsymbol{\pi}_{N+1})\mathbf{e} = 1$$

where Q is the infinitesimal generator of the Markov chain $\xi_t,\ t \geq 0$.

To compute the steady-state distribution of this Markov chain, it is necessary to solve system (1). The matrix of this system has the block-tridiagonal structure. Markov chains having the structure of the generator similar to the one defined in Theorem 1 are sometimes called in the existing literature as the Level-Dependent Quasi-Birth-and-Death processes; see, e.g., [31]. System (1) is finite and can be directly solved via the use of the variety of the standard computer programs. However, the number of equations of the finite system (1) for queueing model under study can be large especially when the buffer capacity N or the number of priority classes is large. Therefore, to effectively solve this system, it is desirable to apply an algorithm that exploits the sparse block-tridiagonal structure of the generator Q. In particular, the algorithm given in [32] can be recommended.

4. Performance Measures

The average number of customers in the buffer is

$$N_{buffer} = \sum_{n=2}^{N+1} (n-1)\boldsymbol{\pi}_n \mathbf{e}.$$

The average number $N^{(r)}_{buffer}$ of type-r, $r = \overline{1,R}$, customers in the buffer can be computed as

$$N^{(r)}_{buffer} = \sum_{n=2}^{N+1} \boldsymbol{\pi}_n (I_{WM} \otimes L_{n-1}(\mathbf{h}_r))\mathbf{e}.$$

The intensity of the output flow of successfully serviced customers is

$$\lambda_{out} = \sum_{n=1}^{N+1} \boldsymbol{\pi}_n (I_W \otimes \mathbf{S_0} \otimes I_{K_{n-1}})\mathbf{e}.$$

The intensity of the output flow of customers who leave the buffer due to impatience is

$$\lambda_{imp} = \sum_{n=2}^{N+1} \boldsymbol{\pi}_n (I_{WM} \otimes L_{n-1}(\boldsymbol{\gamma}))\mathbf{e}.$$

The probability P_{loss} of loss of an arbitrary customer is computed

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

The probability $P_{imp-loss}$ of loss of an arbitrary customer due to impatience is computed

$$P_{imp-loss} = \frac{\lambda_{imp}}{\lambda}.$$

The intensity $\lambda^{(r)}_{imp}$ of the output flow of the type-r, $r = \overline{1,R}$, customers who leave the buffer due to impatience is

$$\lambda^{(r)}_{imp} = \sum_{n=2}^{N+1} \boldsymbol{\pi}_n (I_{WM} \otimes L_{n-1}(\boldsymbol{\gamma}_r))\mathbf{e}$$

where $\boldsymbol{\gamma}_r$ is the row vector of size R with all zero entries except the r-th entry which is equal to γ_r.

The average intensity $\tilde{\lambda}^{(r)}$ of the type-l, $l = \overline{r+1,R}$, customers transformation to the type-r, $r = \overline{1,R-1}$, customers is computed as

$$\tilde{\lambda}^{(r)} = \sum_{l=r+1}^{R} \alpha_l N^{(l)}_{buffer} p_{l,r}.$$

The probability $P^{(r)}_{imp-loss}$, $r=\overline{1,R}$, of loss of an arbitrary type-r customer due to impatience can be computed

$$P^{(r)}_{imp-loss}=\frac{\lambda^{(r)}_{imp}}{\lambda_r+\tilde{\lambda}^{(r)}}.$$

Here, we assume that $\tilde{\lambda}^{(R)}=0$.

The probability of an arbitrary type-r customer loss upon arrival without trying to force out a customer with lower priority is

$$P^{(r)}_{ent-loss-without-force-out}=(1-q)\lambda_r^{-1}\boldsymbol{\pi}_{N+1}(D_r\otimes I_{MK_N})\mathbf{e},\ r=\overline{1,R}.$$

The probability of an arbitrary type-r customer loss upon arrival despite an attempt to force out a customer with lower priority is

$$P^{(r)}_{ent-loss-with-force-out}=q\lambda_r^{-1}\boldsymbol{\pi}_{N+1}(D_r\otimes I_M\otimes\tilde{E}_r)\mathbf{e},\ r=\overline{1,R},$$

where the matrix $\tilde{E}_r$ has all zero entries except the diagonal entries which are equal to the diagonal entries of the matrix $\hat{E}_r$.

The probability of an arbitrary customer loss upon arrival is

$$P_{ent-loss}=\frac{\sum\limits_{r=1}^{R}((1-q)\boldsymbol{\pi}_{N+1}(D_r\otimes I_{MK_N})\mathbf{e}+q\boldsymbol{\pi}_{N+1}(D_r\otimes I_M\otimes\tilde{E}_r)\mathbf{e})}{\lambda}.$$

The probability of an arbitrary type-r customer loss upon arrival is

$$P^{(r)}_{ent-loss}=P^{(r)}_{ent-loss-with-force-out}+P^{(r)}_{ent-loss-without-force-out},\ r=\overline{1,R}.$$

The probability that an arbitrary type-r customer meets the full buffer upon arrival and forces out a customer with lower priority is

$$P^{(r)}_{force-out}=q\lambda_r^{-1}\boldsymbol{\pi}_{N+1}(D_r\otimes I_M\otimes\bar{E}_r)\mathbf{e},\ r=\overline{1,R},$$

where the matrix $\bar{E}_r=\hat{E}_r-\tilde{E}_r$.

Let the square matrix $\hat{E}_{r,l}$, $r=\overline{1,R-1}$, $l=\overline{r+1,R}$, of size $\binom{N+R-1}{R-1}$ define the transition probabilities of the process $\zeta^{(N)}_t$, $t\geq 0$, at the moment when a type-r customer arrives to the system when there are N customers in the buffer and the arriving customer forces out a type-l customer from the buffer. This matrix is defined by analogy with the matrix $\hat{E}_r$ defined above. All entries in each row of this matrix are equal to zero except one entry which can be equal to 1. We assume that each row and column of the matrix $\hat{E}_{r,l}$ correspond to some state $\{\eta_1,\eta_2,\ldots,\eta_R\}$ of the process $\zeta^{(N)}_t$, $t\geq 0$. In the row of the matrix $\hat{E}_{r,l}$ that corresponds to the state $\{\eta_1,\eta_2,\ldots,\eta_R\}$, the entry 1 is located in the column that corresponds to the state $\{\eta_1,\ldots,\eta_{r-1},\eta_r+1,\eta_{r+1},\ldots,\eta_{l-1},\eta_l-1,0,\ldots,0\}$ only in the case if $\eta_m=0$ for all m, $R\geq m>l$, and $\eta_l>0$. If this condition is false, all entries of this row are zero entries.

The intensity $\lambda^{(r)}_{force-out}$ of forcing out from the buffer type-r, $r=\overline{2,R}$, customers is

$$\lambda^{(r)}_{force-out}=q\sum_{l=1}^{r-1}\boldsymbol{\pi}_{N+1}(D_l\otimes I_M\otimes\hat{E}_{l,r})\mathbf{e}.$$

The probability $P_{force-loss}$ of the loss of an arbitrary customer due to forcing out is

$$P_{force-loss} = \frac{\sum\limits_{r=2}^{R} \lambda^{(r)}_{force-out}}{\lambda}.$$

The probability $P^{(r)}_{force-loss}$ of the loss of an arbitrary type-r, $r = \overline{2,R}$, customer due to forcing out is

$$P^{(r)}_{force-loss} = \frac{\lambda^{(r)}_{force-out}}{\lambda_r + \tilde{\lambda}^{(r)}}.$$

5. Numerical Example

In this section, we illustrate the dependencies of some performance measures of the system on the buffer capacity N and show the poor quality of evaluation of the value of the loss probability via the following three simplifications of the model: (i) the arrival flow is assumed to be described not by the $MMAP$, but by the superposition of the stationary Poisson processes; (ii) the service time distribution is assumed to be not of a general phase-type, but exponential; (iii) the arrival flow is assumed to be the superposition of the stationary Poisson processes and the service time distribution is assumed to be exponential.

In this illustrative example, we consider a small information transmission device that is designed for transmission of four types of information. We assume that the distribution of the size of various types information units is the same. The information units of various types have different importance for the system and, correspondingly, have different priority. Let us assume that the arrivals of the units (customers) of different types are modeled by the $MMAP$ arrival process defined by the matrices:

$$D_0 = \begin{pmatrix} -1.8 & 0.0 \\ 0.0 & -0.4458 \end{pmatrix},\ D_1 = \begin{pmatrix} 0.51 & 0.04 \\ 0.006 & 0.1047 \end{pmatrix},$$

$$D_2 = \begin{pmatrix} 0.31 & 0.01 \\ 0.0 & 0.2641 \end{pmatrix},\ D_3 = \begin{pmatrix} 0.41 & 0.01 \\ 0.002 & 0.058 \end{pmatrix},\ D_4 = \begin{pmatrix} 0.5 & 0.01 \\ 0.001 & 0.01 \end{pmatrix}.$$

It has the average arrival intensity $\lambda = 0.600076$, the coefficient of correlation $c_{cor} = 0.148534$, and the coefficient of variation $c^2_{var} = 1.46139$. The intensities of type-r customer arrivals are $\lambda_1 = 0.160747$, $\lambda_2 = 0.270468$, $\lambda_3 = 0.101013$, $\lambda_4 = 0.0678481$, respectively.

The PH service process is defined by the vector $\boldsymbol{\beta} = (0.01, 0.99)$ and the sub-generator

$$S = \begin{pmatrix} -0.1 & 0.1 \\ 0.02 & -2 \end{pmatrix}.$$

The average service time is $b_1 = 0.706060$ and the coefficient of variation is $c^2_{var} = 8.781$.

The rest parameters are as follows: $\gamma_1 = 0.012$, $\gamma_2 = 0.011$, $\gamma_3 = 0.01$, $\gamma_4 = 0.009$, $\alpha_r = 0.1$, $r = \overline{2,4}$, $p_{2,1} = 1$, $p_{3,1} = p_{3,2} = 0.5$, $p_{4,1} = p_{4,2} = p_{4,3} = \frac{1}{3}$, $q = 0.5$.

Let us vary the buffer capacity N over the interval $[1, 25]$ and calculate the main performance measures of the system. It is worth to note that capacity of the buffer not exceeding 25 is realistic in many real-world applications, e.g., in application for modeling emergency departments in a hospital, the number of waiting patients cannot be large because if this number grows, the ambulance cars will deliver new patients to other neighboring hospitals. In modeling the operation of an information transmission device, the capacity of the buffer can also be not very large due to fast obsolescence of the transmitted information.

For computations, we use a PC with an Intel Core i7-8700 CPU and 16 GB RAM, Mathematica 11.0. The computation time for all 25 different buffer capacities is about 15 min.

Figure 2 illustrates the dependence of the average number of customers in the buffer N_{buffer} and the average numbers $N_{buffer}^{(r)}$, $r = \overline{1, R}$, of type-r customers in the buffer on the buffer capacity N. As it is expected, the values N_{buffer} and $N_{buffer}^{(r)}$, $r = \overline{1, R}$, increase with the growth of the buffer capacity N.

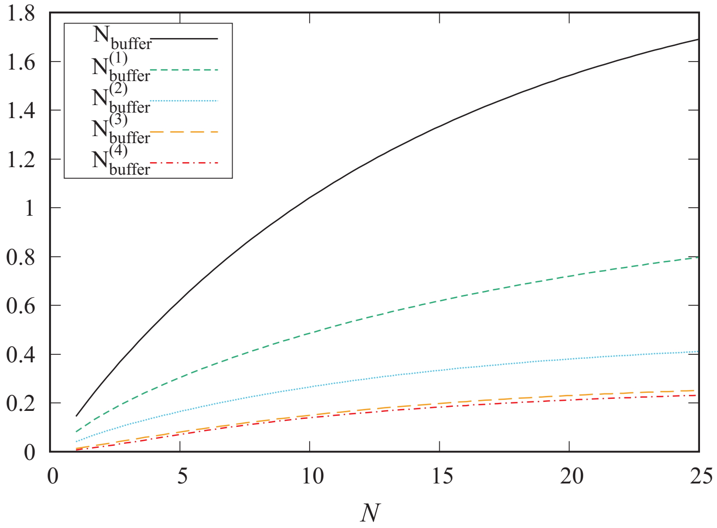

Figure 2. The dependence of N_{buffer} and $N_{buffer}^{(r)}$, $r = \overline{1, R}$, on the buffer capacity N.

Figure 3 illustrates the dependence of the average intensities $\tilde{\lambda}^{(r)}$ of type-l, $l = \overline{r+1, R}$, customers transformation to the type-r, $r = \overline{1, R-1}$, customers on the buffer capacity N. All these intensities increase with the growth of the buffer capacity N because the larger capacity of the buffer implies the longer stay of a customer in the buffer and, therefore, higher chances to increase the priority. The highest value of the intensity $\tilde{\lambda}^{(1)}$ among the values $\tilde{\lambda}^{(r)}$, $r = \overline{1, R-1}$, is easily explained by the fact that about 45 percent of arriving customers are type-2 customers that can increase their priority only to type-1, a half of type-3 customers may increase the priority directly to type-1 and one third of type-4 customers may also increase the priority directly to type-1.

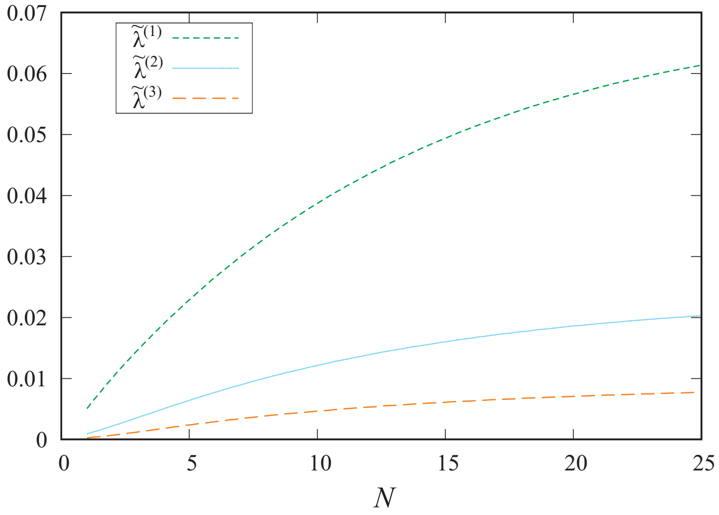

Figure 3. The dependence of the average intensities $\tilde{\lambda}^{(r)}$, $r = \overline{1, R-1}$, on the buffer capacity N.

Figure 4 illustrates the dependence of the probability of an arbitrary customer loss upon arrival $P_{ent-loss}$ and the probabilities of an arbitrary type-r, $r = \overline{1, R}$, customer loss upon arrival $P_{ent-loss}^{(r)}$ on the buffer capacity N. This figure confirms the intuitively clear fact that all these loss probabilities decrease with the growth of the buffer capacity.

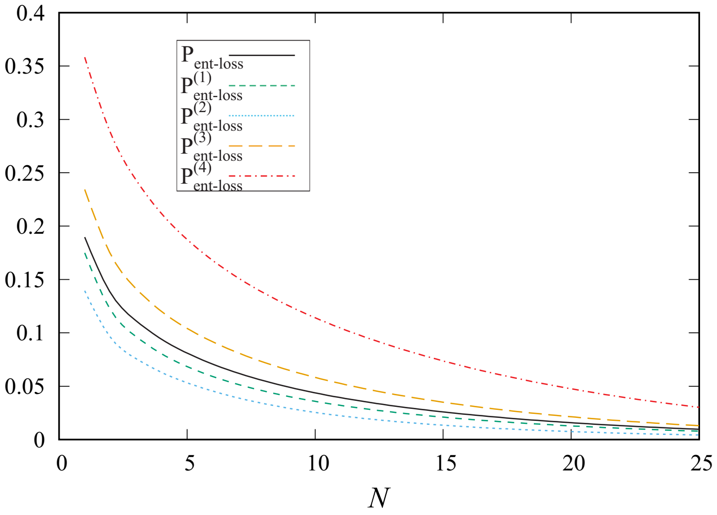

Figure 4. The dependence of the probabilities $P_{ent-loss}$ and $P^{(r)}_{ent-loss}$, $r = \overline{1, R}$, on the buffer capacity N.

Figure 5 illustrates the dependence of the probability $P_{force-loss}$ of the loss of an arbitrary customer due to forcing out and the probability $P^{(r)}_{force-loss}$ of the loss of an arbitrary type-r, $r = \overline{2, R}$, customer on the buffer capacity N. The behavior of these probabilities for type-3 and type-4 customers is explained as follows. For small values of N, these probabilities are small because there is a high probability that such customers are not admitted to the system at all (are lost at the entrance to the system). Then, when the buffer capacity N increases, fewer customers of these types are lost at the entrance and, therefore, more customers are accepted to the buffer and are forced out by the high priority customers. After the buffer capacity N reaches the values about 2 or 3, the probability that the high priority customers will meet full buffer essentially decreases and these customers have no need to force out type-3 and type-4 customers. Consequently, the probabilities $P^{(r)}_{force-loss}$, $r = 3, 4$, decrease when N further increases.

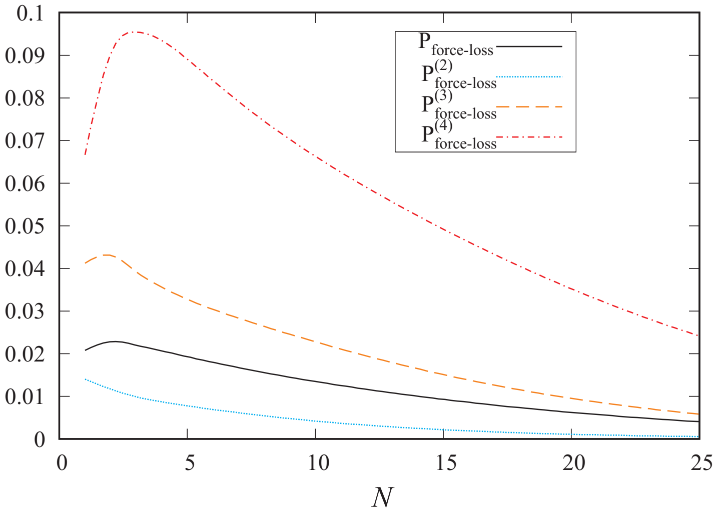

Figure 5. The dependence of the probabilities $P_{force-loss}$ and $P^{(r)}_{force-loss}$, $r = \overline{2, R}$, on the buffer capacity N.

Figure 6 illustrates the dependence of the probability $P_{imp-loss}$ of the loss of an arbitrary customer due to impatience and the probability $P^{(r)}_{imp-loss}$, $r = \overline{1, R}$, of loss of an arbitrary type-r customer due to impatience on the buffer capacity N. When the buffer capacity increases, customers of all types spend more time in the buffer and are lost due to the impatience more frequently.

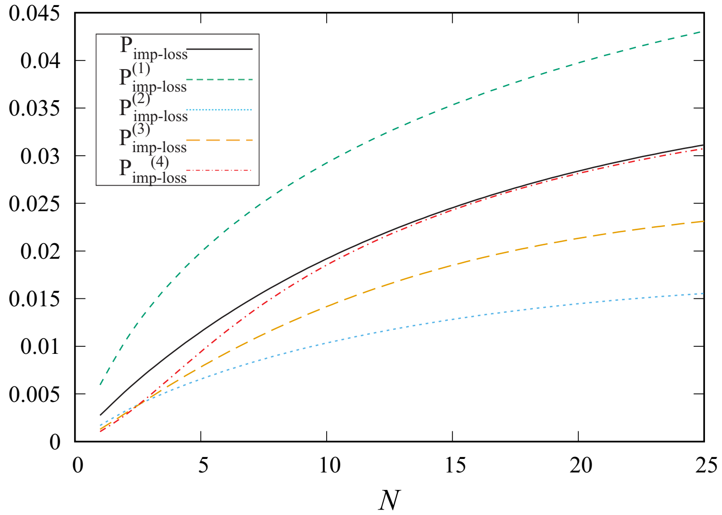

Figure 6. The dependence of the probabilities $P_{imp-loss}$ and $P^{(r)}_{imp-loss}$, $r = \overline{1, R}$, on the buffer capacity N.

As it was announced above, one of the important goals of our numerical example is to demonstrate the poor quality of approximation of the value of the loss probability in the considered $MMAP/PH/1/N$ model with dynamically variable non-preemptive priorities by the value of the loss probability in more simple models coded below as $MMAP/M/1/N$, $M/PH/1/N$ and $M/M/1/N$ type priority models with the same rates of the arrival of different types of customers and the service rate. Using the $MMAP/M/1/N$ model, one ignores that we assumed that the service time has the coefficient of variation $c^2_{var} = 8.781$, not $c^2_{var} = 1$, as the exponential distribution of the service time suggests. Using the $M/PH/1/N$ model, one ignores that the inter-arrival times have the coefficient of correlation $c_{cor} = 0.148534$, and the coefficient of variation $c^2_{var} = 1.46139$, not $c^2_{var} = 1$, as the exponential distribution of inter-arrival times of different types of customers suggests. Using the $M/M/1/N$ model, one assumes a zero coefficient of inter-arrival times and the coefficient of variation of inter-arrival of all types of customers and the service times equal to 1.

Figure 7 illustrates the dependence of the probability P_{loss} of the loss of an arbitrary customer on the buffer capacity N for the considered $MMAP/PH/1/N$ priority system and its particular cases coded as the $MMAP/M/1/N$, $M/PH/1/N$ and $M/M/1/N$ type systems.

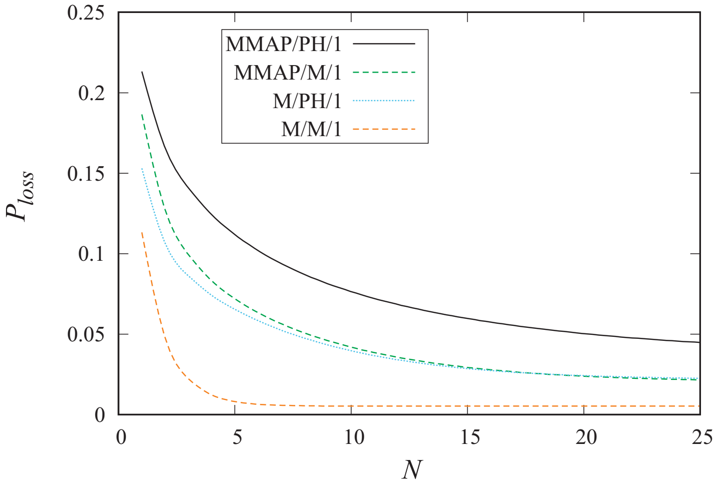

Figure 7. The dependence of the probability P_{loss} on the buffer capacity N for the considered set of the system parameters.

One can see that the values of the loss probabilities computed for the approximating models are essentially smaller than the actual value. It is well known that queueing models with a finite buffer can help to solve the important problem of computing the required capacity N of the buffer, e.g., the problem of finding the minimum value of N such as the loss probability P_{loss} is less than 0.05 can be considered. Using the approximate value of this loss probability computed via the $M/M/1/N$ type system, one can compute that the buffer capacity $N = 2$ is enough to guarantee the fulfillment of the inequality $P_{loss} < 0.05$. Using the approximate value of this loss probability computed via the $M/PH/1/N$ type system, one can compute that the required buffer capacity is $N = 8$. Using the approximate value of this loss probability computed via the $MMAP/M/1/N$ type system, one can compute that the required buffer capacity is $N = 9$. Furthermore, finally, if one properly accounts the values of the coefficients of correlation and variation via the use of the $MMAP/PH/1/N$ model, he/she obtains that the required buffer capacity is $N = 21$. For $N = 2, 8$ and 9 the loss probability has values 0.1659179, 0.087093, and 0.081367, correspondingly, and is essentially larger than 0.05. Therefore, the simplified models give a quite poor estimation of the required capacity of the buffer.

6. Conclusions

We analyzed a quite general single-server queue with heterogeneous customers and a finite buffer. The arrival flow is defined by the $MMAP$ what allows us to take into account the possible correlation of inter-arrival intervals of customers of different types. The service time distribution is of phase-type which allows to approximate more general distributions. Customers of various types have different impatience. It is assumed that the problem of assigning the non-preemptive priorities to different types of customers is solved in the assumption that during staying in the buffer customers can improve their priority. Presented above results allow computing the steady-state distribution of the system and the key performance measures of the system under any fixed set of the system parameters. This creates an opportunity for further use of the obtained results for the optimal scheduling of the flows (assigning the priorities and permissions to increase the priority) under any fixed cost criterion. The criterion may include, e.g., the profit gained via the service of different types of customers or the coefficient of utilization of the server and loss probabilities (rejection at the entrance of the system, pushing out by a high priority customer, leaving the system due to impatience) of different types of customers.

Results can be applied for optimization of the scheduling of: (i) information flows in communication networks where users are categorized into several groups according to their importance, in particular, possible damage caused by the loss or obsolescence of the corresponding information; (ii) patients with different degree of life threat in emergency departments; (iii) perishable goods and foods in warehouses, etc. As future directions of generalization of the considered model we can mention the account of possibility of different distribution of service time for different types of customers and possibility of unreliable service of customers similar to [33].

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

Funding: This work has been supported by Sangji University Grant 2018. This work was also partially supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2018K2A9A1A06072058) and grant No. F19KOR-001 of the Belarusian Republican Foundation for Fundamental Research.

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

References

1. Kalashnikov, V.V. *Mathematical Methods in Queuing Theory*; Springer: Berlin/Heidelberg, Germany, 2013.
2. 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. [CrossRef]
3. Fratini, S. Analysis of a dynamic priority queue. *Commun. Stat. Stoch. Model.* **1990**, *6*, 415–444. [CrossRef]

4. Kim, C.S.; Klimenok, V.; Dudin, A. Priority tandem queueing system with retrials and reservation of channels as a model of call center. *Comput. Ind. Eng.* **2016**, *96*, 61–71. [CrossRef]
5. Knessl, C.; Tier, C.; Cho, D. A dynamic priority queue model for simultaneous service of two traffic types. *SIAM J. Appl. Math.* **2003**, *63*, 398–422. [CrossRef]
6. Ramaswami, V.; Lucantoni, D.M. Algorithmic analysis of a dynamic priority queue. In *Applied Probability—Computer Science: The Interface*; Birkhäuser: Boston, MA, USA, 1982; pp. 157–206,
7. Xin, J.; Zhu, Q.; Liang, G.; Zhang, T. Performance Analysis of D2D Underlying Cellular Networks Based on Dynamic Priority Queuing Model. *IEEE Access* **2019**, *7*, 27479–27489. [CrossRef]
8. De Clercq, S.; Steyaert, B.; Wittevrongel, S.; Bruneel, H. Analysis of a discrete-time queue with time-limited overtake priority. *Ann. Oper. Res.* **2015**, *238*, 69–97. [CrossRef]
9. De Boeck, K.; Carmen, R.; Vandaele, N. Needy boarding patients in emergency departments: An exploratory case study using discrete-event simulation. *Oper. Res. Health Care* **2019**, *21*, 19–31. [CrossRef]
10. Bilodeau, B.; Stanford, D.A. Average Waiting Times in the Two-Class $M/G/1$ Delayed Accumulating Priority Queue. *arXiv* **2020**, arXiv:2001.06054.
11. Fajardo, V.A.; Drekic, S. Waiting Time Distributions in the Preemptive Accumulating Priority Queue. *Methodol. Comput. Appl. Probab.* **2017**, *19*, 255–284. [CrossRef]
12. 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]
13. 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]
14. Stanford, D.A.; Taylor, P.; Ziedins, I. Waiting time distributions in the accumulating priority queue. *Queueing Syst.* **2014**, *77*, 297–330. [CrossRef]
15. Lim, Y.; Kobza, J.E. Analysis of a delay-dependent priority discipline in an integrated multiclass traffic fast packet switch. *IEEE Trans. Commun.* **1990**, *38*, 659–665. [CrossRef]
16. Maertens, T.; Bruneel, H.; Walraevens, J. On priority queues with priority jumps. *Perform. Eval.* **2006**, *63*, 1235–1252. [CrossRef]
17. Klimenok, V.; Dudin, A.; Dudina, O.; Kochetkova, I. Queuing System with Two Types of Customers and Dynamic Change of a Priority. *Mathematics* **2020**, *8*, 824. [CrossRef]
18. 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]
19. 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.
20. He, Q.-M.; Xie, J.; Zhao, X. Priority Queue with Customer Upgrades. *Nav. Res. Logist.* **2012**, *59*, 362–375. [CrossRef]
21. 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.: Branchburg, NJ, USA, 2001; pp. 21–29.
22. Lucantoni, D. New results on the single server queue with a batch Markovian arrival process. *Commun. Stat. Stoch. Model.* **1991**, *7*, 1–46. [CrossRef]
23. Dudin, A.N.; Klimenok, V.I.; Vishnevsky, V.M. *The Theory of Queuing Systems with Correlated Flows*; Springer: Berlin/Heidelberg, Germany, 2019.
24. He, Q.M. Queues with marked customers. *Adv. Appl. Probab.* **1996**, *28*, 567–587. [CrossRef]
25. Kim, C.S.; Dudin, S.; Dudina, O.; Dudin, A.N. Mathematical Model of a Cell With Bandwidth Sharing and Moving Users. *IEEE Trans. Wirel. Commun.* **2020**, *19*, 744–755. [CrossRef]
26. Sun, B.; Dudin, S.; Dudina, O.; Samouylov, K. Optimization of admission control in tandem queue with heterogeneous customers and pre-service. *Optimization* **2020**, *69*, 165–185. [CrossRef]
27. Dudin, S.; Dudin, A.; Dudina, O.; Samouylov, K. Competitive queueing systems with comparative rating dependent arrivals. *Oper. Res. Perspect.* **2020**, *7*, 100139. [CrossRef]
28. Neuts, M. *Matrix-Geometric Solutions in Stochastic Models*; The Johns Hopkins University Press: Baltimore, MD, USA, 1981.

29. Ramaswami, V.; Lucantoni,D. Algorithms for the multi-server queue with phase-type service. *Commun. Stat. Stoch. Model.* **1985**, *1*, 393–417. [CrossRef]
30. Graham, A. *Kronecker Products and Matrix Calculus with Applications*; Horwood, E., Ed.; Courier Dover Publications: Cichester, UK, 1981.
31. Latouche, G.; Ramaswami, V. *Introduction to Matrix Analytic Methods in Stochastic Modeling*; Society for Industrial and Applied Mathematics: Philadelphia, PA, USA, 1999.
32. Baumann, H.; Sandmann, W. Numerical solution of level dependent quasi-birth-and-death processes. *Procedia Comput. Sci.* **2010**, *1*, 1561–1569. [CrossRef]
33. 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. [CrossRef]