mathematics

MDPI

Article

Analysis of Single-Server Multi-Class Queue with Unreliable Service, Batch Correlated Arrivals, Customers Impatience, and Dynamical Change of Priorities

Alexander Dudin [1,2,*], Olga Dudina [1], Sergei Dudin [1] and Konstantin Samouylov [2]

[1] Department of Applied Mathematics and Computer Science, Belarusian State University, 4, Nezavisimosti Ave., 220030 Minsk, Belarus; dudina@bsu.by (O.D.); dudin85@mail.ru (S.D.)
[2] Applied Mathematics and Communications Technology Institute, Peoples' Friendship University of Russia (RUDN University), 6 Miklukho-Maklaya St., 117198 Moscow, Russia; ksam@sci.pfu.edu.ru
* Correspondence: dudin@bsu.by or dudin_alexander@mail.ru

Abstract: A single-server non-pre-emptive priority queueing system of a finite capacity with many types of customers is analyzed. Inter-arrival times can be correlated and batch arrivals are allowed. Possible unreliability of the server, implying the loss of a customer or the necessity of its service from the early beginning or some phase of the service, is taken into account. Initial priorities provided to various types of customers at the arrival moment can be varied (increased or decreased) after the random amount of time during the customer stay in the buffer. Such a type of queues arises in the modeling operation of various emergency care systems, information, and perishable goods delivering systems, etc. The stationary behavior of the system is described by the finite state multi-dimensional continuous-time Markov chain with the upper-Hessenberg block structure of the generator. The stationary distribution of the system states and some important characteristics of the system are calculated. The presented numerical examples illustrate opportunities to quantitatively evaluate the impact of the buffer capacity and customers' mean arrival rate on the most important characteristics of the system. The possibility of solving optimization problems is briefly shown.

Keywords: dynamic priority queue; batch marked Markov arrival process; phase-type with failures time distribution; performance evaluation

Citation: Dudin, A.; Dudina, O.; Dudin, S.; Samouylov, K. Analysis of Single-Server Multi-Class Queue with Unreliable Service, Batch Correlated Arrivals, Customers Impatience and Dynamical Change of Priorities. *Mathematics* **2021**, *9*, 1257. https://doi.org/10.3390/math9111257

Academic Editor: János Sztrik

Received: 28 April 2021
Accepted: 25 May 2021
Published: 31 May 2021

1. Introduction

Queueing theory provides a powerful tool for design, capacity planning, performance evaluation, and optimization of many real-world systems. In the simplest settings, all customers of a queueing system are assumed to be homogeneous, having equal requirements and rights to receive service. However, in many real-world systems, this is not true. Some customers or classes of customers, for certain reasons, are more important for the system and are provided higher priority in access to the servers comparing to other classes. These reasons can be versatile. For example, in information transmission systems, signaling information is much more important than the information sent by the users, time-sensitive information should be transferred earlier than the elastic insensitive information. In intellectual vehicular networks, the transmission of driving safety information is much more urgent than the transmission of the info-entertainment information. Processing of information generated by the certified user is more urgent than the processing of information by the cognitive user in cognitive radio systems. The handover user who arrived at the cell of a mobile network has to be treated differently than the new user trying to establish a connection within a given cell, etc. In emergency service systems, in particular, an emergency healthcare system, the patients can be sorted and treated by injury severity. In food delivery services, the most rapidly deteriorating (perishable) items have to be delivered first. In communication systems, the clients can sign agreements with

different service levels (and different fees). The ultra-reliable low-latency communication (URLLC) applications in 5G networks have higher priority than the enhanced mobile broadband (eMBB) applications, etc. A proper choice of the priorities can significantly increase the economic profit gained from the operation of a corresponding system and revenue generating businesses. Therefore, the priority queueing models have attracted a lot of attention from researchers. In the overwhelming majority of the existing research, the priorities (non-pre-emptive or pre-emptive) are assigned from the early beginning and remain permanently valid.

However, there exist a lot of various situations where it is necessary to be more flexible and change the current priorities of customers. For example, in an emergency healthcare system, after assigning the priorities to various customers as the result of the initial triage, the state of the health of a patient can become essentially worse during his or her waiting for the treatment and his or her priority has to be increased. In dispatching the ambulance cars, the priority of an initially non-urgent patient can increase due to approaching some existing deadline for the provision of an ambulance. The same situation occurs in signal processing systems with information obsolescence in which the further processing of the signal becomes meaningless after some amount of time.

The short surveys of the literature devoted to the queueing systems with changing the priority of customers are given, e.g., in [1–11]. In [5–8], the change of the priority occurs deterministically as a function of the elapsed sojourn time. In the rest of the cited papers, the change of the priority occurs stochastically after some random time. In [1], a table giving the brief classification of the existing results for the case of the stochastic change is presented. The existing papers are classified according to the number of priority classes, arrival process, distribution of the service time, and time until the change of the priority and the obtained results. The model considered in paper [1] is the most general with respect to the pattern of the arrival process and the distribution of the service time and time until the change of the priority. The main restriction of this paper is that only the case of two priority classes is dealt with. The closest queueing system to the model considered in this paper was analyzed in paper [12]. That queueing system has a single-server and a buffer of a finite capacity. Arriving customers are heterogeneous and impatient. Impatience depends on the type of a customer. Different types of customers are assigned the different non-pre-emptive priorities at an arrival moment. However, after the exponentially distributed interval of time during the stay in the buffer, the established priority of a customer can increase. This can occur several times during the customer waiting time. The dynamics of the considered system is described in [12] by the multi-dimensional continuous-time Markov chain. The generator of this chain is obtained, the problem of computation of the stationary state probabilities is touched. Explicit formulas for computing the key performance measures of the system via the obtained stationary probabilities are derived. The dependencies of certain performance measures on the capacity of the buffer are graphically illustrated. The importance of account of correlation in the arrival process and variance of the service time is numerically confirmed.

The distinguishing features of the model considered in this paper comparing to [12] are the following ones:

- The batch arrival of customers of different classes in the batch marked Markov arrival process (BMMAP), which is typical in many real-world systems, is supposed while the customers arrive only one-by-one in the marked Markov arrival process (MMAP) in [12];
- The server is supposed to be unreliable with the options of the loss of a customer, during service of which the breakdown occurs, service repetition from the beginning or from the phase of service at which the breakdown occurs, while the server is assumed reliable in [12];
- Both possibilities of the priority increase and decrease are explored here while only the case of the possible increase of the priority is supposed in [12].

The brief outline of the paper is as follows. In Section 2, the mathematical model is described. In Section 3, the operation of the system is described by the multi-dimensional continuous-time Markov chain with the upper-Hessenberg block structure of the generator. Some results from [12] and their extensions are used to compute the blocks of the generator. A numerically stable algorithm for computation of the stationary distribution of this Markov chain, which essentially exploits the revealed structure of the generator, is presented. Formulas for computation of the key performance measures of the system based on the knowledge of the stationary distribution of the Markov chain are presented in Section 4. Section 5 contains the numerical results that show the dependencies of some performance measures (including the average number of customers of different types in the buffer and customer loss probabilities due to the buffer overflow, breakdown of the server, and impatience of the customers) on the buffer capacity and the average arrival rate. The possibility of the optimal choice of the capacity of the buffer, which guarantees that the loss probability does not exceed the pre-assigned small value, is illustrated. It is numerically shown that sometimes, due to the existence of customer losses caused not by the buffer overflow but by the server unreliability or customers impatience, it is not possible to decrease the loss probability to some admissible level by means of the corresponding increase of the buffer capacity. Section 6 concludes the paper.

2. Mathematical Model

We consider a single-server queueing system with a finite buffer of capacity N and R types of customers. The structure of the system is presented in Figure 1.

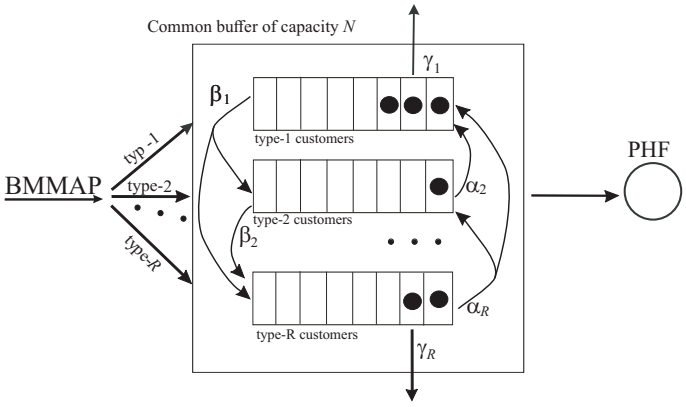

Figure 1. Structure of the system.

Customers of different types have identical requirements to the service process but different priorities. We assume that the types of customers are enumerated in the decreasing order of the priority. This means that type-1 customers have the highest priority, etc., type-R customers have the lowest priority. We assume that a customer having the highest priority among the presented in the buffer is selected for service at any service completion epoch.

The input flow of customers is described by the $BMMAP$. Customers arrival in the $BMMAP$ is defined by the irreducible continuous time Markov chain ν_t, $t \geq 0$, having the finite state space $\{1, \ldots, W\}$. The sojourn time of the chain ν_t, $t \geq 0$, in the state ν is exponentially distributed with a positive parameter λ_ν. After this time expires, with probability $p_0(\nu, \nu')$ this chain jumps into the state ν', $\nu' \in \{1, \ldots, W\}$, $\nu' \neq \nu$, without generation of customers and with probability $p_r^{(k)}(\nu, \nu')$ the chain jumps into the state ν', $\nu' \in \{1, \ldots, W\}$, and a batch consisting of k customers of type-r is generated. Here, we assume that the maximal batch size of type r customers is limited by the parameter $K_r, K_r \geq 1$. Let us denote as K the maximal batch size among all types of customers, i.e., $K = \max\{K_r, r = \overline{1, R}\}$.

The parameters defining the $BMMAP$ can be stored in the square matrices D_0, $D_r^{(k)}$, $r = \overline{1,R}$, $k = \overline{1,K_r}$, of size W defined by their entries:

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

$$(D_0)_{\nu,\nu'} = \lambda_\nu p_0(\nu,\nu'),\ (D_r^{(k)})_{\nu,\nu'} = \lambda_\nu p_r^{(k)}(\nu,\nu'),\ \nu,\nu' = \overline{1,W},\ k = \overline{1,K_r}, r = \overline{1,R}.$$

The matrix

$$D(1) = D_0 + \sum_{r=1}^{R}\sum_{k=1}^{K_r} D_r^{(k)}$$

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

Let us denote as $\boldsymbol{\theta}$ the stationary probability vector of the states of the Markov chain ν_t, $t \geq 0$.

This vector can be found as the unique solution to the system

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

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

The average intensity λ_r of type-r customers arrival is defined as

$$\lambda_r = \boldsymbol{\theta}\sum_{k=1}^{K_r} kD_r^{(k)}\mathbf{e}, r = \overline{1,R}.$$

The average intensity λ_r^{batch} of batches of type-r customers arrival is calculated by

$$\lambda_r^{batch} = \boldsymbol{\theta}\sum_{k=1}^{K_r} D_r^{(k)}\mathbf{e},\ \ r = \overline{1,R}.$$

The average intensity λ of customers arrival is defined as

$$\lambda = \sum_{r=1}^{R}\lambda_r.$$

For more information about the $BMMAP$, see, e.g., [13].

If a batch of customers of any type arrives when the server is idle, the first customer of the batch immediately starts processing by the server (service), and the rest of the customers are placed into the buffer. If the capacity of the buffer is not enough for storing all the customers from the batch, the customers, for which there is not enough buffer space, leave the system permanently (are lost). This means that we assume the partial admission discipline.

During the stay in the buffer, each customer of type-r, $r = \overline{1,R}$, can change (increase or decrease) its priority. Obviously, for most applications, it is sufficient to consider the model with increasing priority of customers. However, for the mathematical generality of the model and having in mind the potential application in information processing systems where long waiting implies the obsolescence of information and the decrease of its value, we consider the possibility of both increasing and decreasing the priority. We assume that after an exponentially distributed time with the parameter α_r any type-r, $r = \overline{2,R}$, customer becomes a type-l customer with the probability $p_{r,l}$, $l = \overline{1,r-1}$, independently of other customers. Here, $\sum_{l=1}^{r-1} p_{r,l} = 1,\ r = \overline{2,R}$. After the exponentially distributed time with the parameter β_r, $r = \overline{1,R-1}$, any type-r customer becomes a type-l customer with the probability $q_{r,l}$, $l = \overline{r+1,R}$, independently of other customers. Here, $\sum_{l=r+1}^{R} q_{r,l} = 1,\ r = \overline{1,R-1}$.

Customers can be impatient and leave the buffer without service, independently of other customers. If, during an exponentially distributed with parameter γ_r time, a type-r customer does not succeed to enter the service, this customer departs from the system permanently (is lost), $r = \overline{1,R}$. If, during the stay in the buffer, the customer changes his or her priority and becomes type-r' customer, the patience time restarts and has an exponential distribution with the parameter $\gamma_{r'}$, $r' = \overline{1,R}$. We denote $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_R)$.

Customers' service times have the so called phase-type distribution with failures (PHF), see [14]. This distribution is an essential generalization of the classical phase-type distribution. The duration of a service time and the result of service are defined by the continuous-time underlying Markov chain m_t, $t \geq 0$, having a finite state space $\{1, \ldots, M, M+1, M+2\}$. Here, the states $\{1, \ldots, M\}$ are transient while the states $M+1$ and $M+2$ are the absorbing states. The initial state of the underlying Markov chain at the service beginning moment is randomly chosen from the set of the transient states. The probabilities defining the choice are given by the components of the stochastic row vector $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_M)$. The intensities of the transition between transient states (phases) of the process m_t are defined by the entries of the sub-generator S.

The transition of the chain to one of the two absorbing states corresponds to the end of the current stage of the service. The transition intensities of the chain to the absorbing state $M+1$ are given by the components of the column vector $\mathbf{S}_1$. The transitions to the absorbing state $M+1$ correspond to the successful service of a customer. The intensities of the transition to this state $M+2$ are given by the components of the column vector $\mathbf{S}_2$. The transition of the chain to the absorbing state $M+2$ means the end of the stage of the service due to a failure occurrence. When a failure occurs, then the serviced customer is lost with probability q_1, restarts service from the early beginning with probability q_2, and continues service from the phase at which the failure occurred with probability $1 - q_1 - q_2$.

For more information about PHF distribution and its properties see [14,15].

3. Process of the System States

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

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

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 $BMMAP$, $\nu_t = \overline{1,W}$;
- m_t is the state of the underlying process of PHF 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 ξ_t, $t \geq 0$, let us enumerate its states in the direct lexicographic order of the components ν_t and m_t, and the reverse lexicographic order of the components $\eta_t^{(1)}, \ldots, \eta_t^{(R)}$.

Let us firstly consider the process $\zeta_t^{(n)} = \{\eta_t^{(1)}, \ldots, \eta_t^{(R)}\}$, $t \geq 0$, $\eta_t^{(r)} = \overline{0,n}$, $r = \overline{1,R}$, $\sum\limits_{r=1}^{R} \eta_t^{(r)} = n$. Under the fixed total number n of customers in the buffer, the components of the R-dimensional process $\zeta_t^{(n)}$ describe the dynamics of the number of customers of different types currently presenting in the buffer. To describe transition intensities of this process, we use several auxiliary matrices.

(a) The intensities of transition of this process when some customer departs from the buffer due to impatience are the elements of the matrix $L_n(\gamma)$;

(b) The intensities of transition of this process when some customer changes the priority are the elements of the matrix $Y_n = Y_n(H)$. Here, the matrix H defines the intensities of increasing and decreasing of priorities. It is described by formula:

$$H = \begin{pmatrix} 0 & q_{1,2}\beta_1 & q_{1,2}\beta_1 & \cdots & q_{1,R-1}\beta_1 & q_{1,R}\beta_1 \\ \alpha_2 & 0 & q_{2,3}\beta_2 & \cdots & q_{2,R-1}\beta_2 & q_{2,R}\beta_2 \\ p_{3,1}\alpha_3 & p_{3,2}\alpha_3 & 0 & \cdots & q_{3,R-1}\beta_3 & q_{3,R}\beta_3 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ p_{R-1,1}\alpha_{R-1} & p_{R-1,2}\alpha_{R-1} & p_{R-1,3}\alpha_{R-1} & \cdots & 0 & q_{R-1,R}\beta_{R-1} \\ p_{R,1}\alpha_R & p_{R,2}\alpha_R & p_{R,3}\alpha_R & \cdots & p_{R,R-1}\alpha_R & 0 \end{pmatrix}.$$

(c) The transition probabilities of the process $\zeta_t^{(n)}$ when a new customer arrives and is admitted to the system are the elements of the matrix $A_n(\mathbf{h})$, $n = \overline{0, N-1}$. Here, the entries h_r of the row vector $\mathbf{h} = (h_1, h_2, \ldots, h_R)$ are the probabilities that the arrived to the system customer has type-r, $r = \overline{1, R}$;

(d) The transition probabilities of the process $\zeta_t^{(n)}$ when new customer is picked up for service are the elements of the matrix E_n^-, $n = \overline{1, N}$. This customer has the highest priority among all customers currently waiting in the system.

The algorithms for computation of the matrices $L_n(\gamma)$, $Y_n(H)$, $A_n(\mathbf{h})$ and E_n^- and their proofs are presented in [12].

Let us introduce the following notation:

- $\otimes$ and $\oplus$ indicate the symbols of the Kronecker product and sum of matrices, respectively, see [16];
- $\mathbf{h}_r = (\underbrace{0, \ldots, 0}_{r-1}, 1, \underbrace{0, \ldots, 0}_{R-r})$, $r = \overline{1, R}$;
- $\hat{I}_n = -\mathrm{diag}\{Y_n\mathbf{e} + L_n\mathbf{e}\}$, $n = \overline{1, N}$, where $\mathrm{diag}\{\ldots\}$ is the diagonal matrix with the diagonal elements listed in the brackets;
- $T_n = \binom{n+R-1}{R-1} = \frac{(n+R-1)!}{n!(R-1)!}$, $n = \overline{1, N}$, $T_0 = 1$.

By analyzing all possible transitions of the Markov chain ξ_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 ξ_t, $t \geq 0$, has the following block-tridiagonal structure*

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

The non-zero blocks are defined as follows:

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

$$Q_{1,1} = D_0 \oplus S + I_W \otimes [(1 - q_1 - q_2)\mathrm{diag}\{(\mathbf{S}_2)_l,\ l = \overline{1, M}\} + q_2\mathbf{S}_2\boldsymbol{\sigma}], \tag{1}$$

$$\begin{aligned} Q_{n,n} =& D_0 \oplus S \otimes I_{T_{n-1}} + I_W \otimes [(1 - q_1 - q_2)\mathrm{diag}\{(\mathbf{S}_2)_l,\ l = \overline{1, M}\} + \\ & q_2\mathbf{S}_2\boldsymbol{\sigma}] \otimes I_{T_{n-1}} + I_{WM} \otimes (Y_{n-1} + \hat{I}_{n-1}),\ n = \overline{2, N}, \end{aligned} \tag{2}$$

$$Q_{N+1,N+1} = (D_0 \oplus S) \otimes I_{T_N} + (1 - q_1 - q_2) I_W \otimes \operatorname{diag}\{(\mathbf{S}_2)_l,\ l = \overline{1, M}\} \otimes I_{T_N} + I_{WM} \otimes (Y_N + \hat{I}_N) + \sum_{r=1}^{R} \sum_{k=1}^{K_r} D_r^{(k)} \otimes I_{MT_N} + q_2 I_W \otimes \mathbf{S}_2 \boldsymbol{\sigma} \otimes I_{T_N}, \tag{3}$$

$$Q_{0,1} = \sum_{r=1}^{R} D_r^{(1)} \otimes \boldsymbol{\sigma}, \tag{4}$$

$$Q_{0,n} = \sum_{r=1}^{R} Z_n^r,\ n = \overline{2, \min\{N, K\}}, \tag{5}$$

where

$$Z_n^r = \begin{cases} D_r^{(n)} \otimes \boldsymbol{\sigma} \otimes (A_0(\mathbf{h}_r) \times A_1(\mathbf{h}_r) \cdots \times A_{n-2}(\mathbf{h}_r)), & \text{if } n \le K_r, \\ O, & \text{otherwise,} \end{cases}$$

$$Q_{0,N+1} = \sum_{r=1}^{R} \sum_{k=N+1}^{K_r} D_r^{(k)} \otimes \boldsymbol{\sigma} \otimes (A_0(\mathbf{h}_r) \times A_1(\mathbf{h}_r) \cdots \times A_{N-1}(\mathbf{h}_r)), \tag{6}$$

$$Q_{n,n+k} = \sum_{r=1}^{R} \tilde{Z}_n^{r,k},\ n = \overline{1, N},\ k = \overline{1, \min\{N - n, K\}},$$

where

$$\tilde{Z}_n^{r,k} = \begin{cases} D_r^{(k)} \otimes I_M \otimes A_{n-1}(\mathbf{h}_r) \times A_n(\mathbf{h}_r) \cdots \times A_{n+k-2}(\mathbf{h}_r), & \text{if } k \le K_r, \\ O, & \text{otherwise,} \end{cases}$$

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

$$Q_{1,0} = I_W \otimes (q_1 \mathbf{S}_2 + \mathbf{S}_1),$$

$$Q_{n,n-1} = I_W \otimes (q_1 \mathbf{S}_2 + \mathbf{S}_1) \boldsymbol{\sigma} \otimes E_{n-1}^{-} + I_{WM} \otimes L_{n-1}(\gamma),\ n = \overline{2, N+1}.$$

Proof. The proof of the theorem is based on the analysis of all possible transitions of the Markov chain $\xi_t,\ t \ge 0$, during the time interval of infinitesimally small length. The diagonal entries of the matrices $Q_{n,n}$, $n = \overline{0, N+1}$, are negative. The moduli of these elements define the total intensity of the exit of the Markov chain $\xi_t,\ t \ge 0$, from the corresponding state.

- In the case $n = 0$, the system is empty and the Markov chain $\xi_t,\ t \ge 0$, can only leave its state when the underlying process of the *BMMAP* arrival flow makes a transition from one state to another. The intensities of such transitions are defined by the modulus of the diagonal entries of the matrix D_0;
- In the case $n = 1$, the server is busy and the buffer is empty. In this case, the Markov chain $\xi_t,\ t \ge 0$, can leave its current state also due to a transition in the underlying process of service. The intensities of the underlying process of service exit from its states are given as the modulus of the diagonal entries of the matrix S and the total intensities of transitions of the service and arrival process in the case $n = 1$ are given as the modulus of the corresponding entries of the matrix $D_0 \oplus S$. However, not all the transitions of the service process lead to the change of the state of the chain ξ_t. There are possible situations when the service process transits from some state to the second absorbing state, a failure occurs and a customer who received service when a failure occurred, restarts the service from the same state. When such a situation occurs, the chain ξ_t does not exit from its state. The intensities of such transitions are given by

the diagonal entries of the matrix $I_W \otimes [(1-q_1-q_2)\mathrm{diag}\{(\mathbf{S}_2)_l,\ l=\overline{1,M}\}+q_2\mathbf{S}_2\boldsymbol{\sigma}]$. Thus, the total intensities of the leaving the states of the Markov chain $\xi_t,\ t\geq 0$, in the case $n=1$ are given by the modulus of the corresponding entries of the matrix defined by formula (1);

- In the case $n=\overline{2,N}$, there are $n-1$ customers in the buffer. Therefore, in contrast to the case $n=1$, the Markov chain $\xi_t,\ t\geq 0$, can also leave its state due to a transition of the process $\zeta_t^{(n-1)}$ that describes the dynamics of the types of the customers staying in the buffer. The intensities of transitions of the process $\zeta_t^{(n-1)}$ are given as the modulus of the diagonal entries of the matrix $\hat{I}_{n-1}$. Thus, the total intensities of the leaving the states of the Markov chain $\xi_t,\ t\geq 0$, in the case $n=\overline{2,N}$, are given by the modulus of the corresponding entries of matrix (2);
- In the case $n=N+1$, the Markov chain $\xi_t,\ t\geq 0$, can leave its state due to the same reasons as in the previous case. The reason why we separately consider this case is the following. When the buffer is full, the situation is possible when the underlying process of the $BMMAP$ makes a transition from one state to the same state with the generation of a batch of customers (transitions to the same state without a batch generation are not allowed in $BMMAP$). Since the buffer is full, the arriving batch will be lost. So, such a transition does not change the state of the Markov chain $\xi_t,\ t\geq 0$, and it is required to add to the negative diagonal entries of the matrix $(D_0\oplus S)\otimes I_{T_N}+(1-q_1-q_2)I_W\otimes \mathrm{diag}\{(\mathbf{S}_2)_l,\ l=\overline{1,M}\}\otimes I_{T_N}+I_{WM}\otimes\hat{I}_N+q_2I_W\otimes\mathbf{S}_2\boldsymbol{\sigma}\otimes I_{T_N}$ the positive diagonal entries of the matrix $\sum\limits_{r=1}^{R}\sum\limits_{k=1}^{K_r}D_r^{(k)}\otimes I_{MT_N}$. Thus, the total intensities of the exit from the states of the Markov chain $\xi_t,\ t\geq 0$, in the case $n=N+1$ are given by the modulus of the corresponding entries of matrix (3).

The non-diagonal entries of the matrices $Q_{n,n},\ n=\overline{0,N+1}$, are positive and define the intensities of the Markov chain $\xi_t,\ t\geq 0$, transitions that do not lead to the change the number of customers in the system n. These transitions are:

- The transitions of the $BMAP$ underlying process without generation of a batch (the intensities are defined by non-diagonal entries of the matrix D_0) in the case $n=\overline{0,N+1}$;
- The transitions of the service process to the second absorbing state which are accompanied with the restart of the service of a customer from the beginning from another state (they are defined by the non-diagonal entries of the matrix $I_Wq_2\mathbf{S}_2\boldsymbol{\sigma}\otimes I_{T_{n-1}}$) in the case $n=\overline{1,N+1}$;
- The transitions of the process $\zeta_t^{(n-1)}$ due to the change of the priority by some customer (the intensities of such transitions are defined by non-diagonal entries of the matrix $I_{WM}\otimes Y_{n-1}$) in the case $n=\overline{2,N+1}$;
- The transitions of the $BMAP$ underlying process with generation of a batch to the same state (the intensities are defined by non-diagonal entries of the matrix $\sum\limits_{r=1}^{R}\sum\limits_{k=1}^{K_r}D_r^{(k)}\otimes I_{MT_N}$) in the case $n=N+1$.

Taking into account all these reasonings, we obtain the form of the diagonal blocks $Q_{n,n},\ n=\overline{0,N+1}$, of the generator Q.

The entries of the matrices $Q_{n,n+k},\ n=\overline{0,N},\ k=\overline{1,\max\{K,N+1-n\}}$, are positive and define the intensities of the Markov chain $\xi_t,\ t\geq 0$, transitions that lead to the increase in the number of customers in the system from n to $n+k$. The entries of the matrix $Q_{0,1}$ define the intensity of the chain ξ_t transitions that lead to the increase in the number of customers in the system from 0 to 1. Such transitions occur when exactly one customer of any type arrives at the empty system. In this case, we have to establish the state of the service process according to the probabilistic vector $\boldsymbol{\sigma}$. Thus, the matrix $Q_{0,1}$ has form (4).

The entries of the matrices $Q_{0,n},\ n=\overline{2,\min\{N,K\}}$, define the intensity of the chain ξ_t transitions that lead to the increase in the number of customers in the system from 0 to n. Such transitions occur when exactly n customers of any type arrive at the empty system. In this case, except for the establishing the state of the underlying process of

service, we also have to establish the states of the process that describes the transitions of the number of different types of customers in the buffer using the matrices $A_0(\mathbf{h}_r) \times A_1(\mathbf{h}_r) \cdots \times A_{n-2}(\mathbf{h}_r)$, in the case of type-$r$ customers arrival. Note, the arrival of n type-r customers is not possible when n is greater than the maximal batch size K_r. For accounting such situations we introduce the matrices Z_n^r. Thus, the matrices $Q_{0,n}$, $n = \overline{2, \min\{N, K\}}$, have form (5).

The matrix $Q_{0,N+1}$ is not zero only when $N + 1 \leq K$. The entries of this matrix give the transitions of the chain ξ_t that lead to the arrival of more than N customers to the empty system. The explanation of form (6) of the matrix $Q_{0,N+1}$ is the same as in the previous case. The form of the matrices $Q_{n,n+k}$ and $Q_{n,N+1}$ can be explained similarly as above, only taking into account that there is no need to establish the state of the service process because the server is busy in these cases. Taking into account all these reasonings we obtain the form of the blocks $Q_{n,n+k}$, $n = \overline{0, N}$, of the generator Q.

The entries of the matrices $Q_{n,n-1}$, $n = \overline{1, N+1}$, are positive and define the intensities of the Markov chain ξ_t, $t \geq 0$, transitions that lead to the decrease in the number of customers in the system by one. In the case $n = 1$, such a decrease occurs when the service of a customer is successfully finished or terminated due to a failure occurrence with subsequent loss of the customer under the service. The intensities of such transitions are given by the entries of the matrix $I_W \otimes (q_1 \mathbf{S}_2 + \mathbf{S}_1)$. In the case $n = \overline{2, N+1}$, such a decrease also can occur due to the successful service or loss of a customer due to the failure. Since the buffer is not empty, in such a situation it is necessary to choose a customer with the highest priority from the buffer and establish the state of a new service process. Besides this, such a decrease can occur due to the customer loss from the buffer due to impatience. The intensities of such transitions are given by the matrix $I_{WM} \otimes L_{n-1}(\boldsymbol{\gamma})$. Taking into account all these reasonings, we obtain the form of the blocks $Q_{n,n-1}$, $n = \overline{1, N-1}$, of the generator Q. Since the customers are serviced only one by one, the service completion of two or more customers during a infinitesimally small time interval is impossible. Thus, $Q_{n,n-k}$, $n = \overline{1, N-1}$, $k = \overline{0, n-2}$, are zero matrices. □

The Markov chain ξ_t, $t \geq 0$, has a finite state space and is irreducible. This implies that, the steady-state probabilities of the system

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

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

exist for all values of the system parameters.

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

The vectors $\boldsymbol{\pi}_n$, $n = \overline{0, N+1}$, satisfy the following system of equilibrium (Chapman-Kolmogorov) equations:

$$(\boldsymbol{\pi}_0, \boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_{N+1}) Q = \mathbf{0}, \tag{7}$$

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

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

If the number of equations of system (7) is large, to solve it one should use an algorithm that takes into account the sparse structure of the generator Q. We propose the following Algorithm 1.

The idea of this algorithm is to substitute the system (7) with the system of equilibrium equations for the family of specially constructed censored Markov chain with varying censoring levels for the initial Markov chain ξ_t, $t \geq 0$. This idea is borrowed from the paper [17] where the analyzed Markov chain had an infinite state space. The presented above algorithm is essentially less memory consuming than the algorithm that is the direct adaptation of the algorithm from [17]. The latter algorithm has a step similar

to step 3 in the presented algorithm. However, in [17] the sequence of the matrices of size $WMT_{n-1} \times WMT_{n-1}$, $n = \overline{1, N+1}$, is recursively computed and stored while in the presented algorithm we recursively compute and store only the row vectors $\boldsymbol{\phi}_n$ of size WMT_{n-1}.

Algorithm 1: Computation of the probability vectors $\boldsymbol{\pi}_n$, $n = \overline{0, N+1}$

Step 1. Calculate the matrices $P_{i,n}$ using the following recursive formulas:

$$P_{i,N+1} = -Q_{i,N+1}(Q_{N+1,N+1})^{-1},\ i = \overline{0, N},$$

$$P_{i,n} = -(Q_{i,n} + P_{i,n+1}Q_{n+1,n})(Q_{n,n} + P_{n,n+1}Q_{n+1,n})^{-1},$$
$$i = \overline{0, n-1},\ n = N, N-1, \ldots, 1.$$

Step 2. Find the vector $\boldsymbol{\phi}_0$ as the only solution to the following system of linear algebraic equations:

$$\boldsymbol{\phi}_0(Q_{0,0} + P_{0,1}Q_{1,0}),\ \ \boldsymbol{\phi}_0\mathbf{e} = 1.$$

Step 3. Calculate the vectors $\boldsymbol{\phi}_n$, $n = \overline{1, N+1}$, as:

$$\boldsymbol{\phi}_n = \sum_{i=0}^{n-1} \boldsymbol{\phi}_i P_{i,n} = 0,\ n = \overline{1, N+1}.$$

Step 4. Find the constant $c = \left(\sum_{n=0}^{N+1} \boldsymbol{\phi}_n \mathbf{e} \right)^{-1}$.

Step 5. Find the vectors of the stationary distribution $\boldsymbol{\pi}_n$, $n = \overline{0, N+1}$, as

$$\boldsymbol{\pi}_n = c\boldsymbol{\phi}_n.$$

4. Performance Measures

When the stationary probabilities have been computed, we can find different system 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 $N_{buffer}^{(r)}$ of type-r, $r = \overline{1, R}$, customers in the buffer can be computed as

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

Here, the matrix $L_{n-1}(\mathbf{h}_r)$ is computed by the same formulas as the matrix $L_{n-1}(\boldsymbol{\gamma})$, with replacement of the vector $\boldsymbol{\gamma}$ of impatience rates by the stochastic vector $\mathbf{h}_r$.

The output rate of successfully serviced customers is

$$\lambda_{out} = \sum_{n=1}^{N+1} \boldsymbol{\pi}_n (I_W \otimes \mathbf{S}_1 \otimes I_{T_{n-1}})\mathbf{e}.$$

The output rate of customers who leave the system due to service failure is

$$\lambda_{fail} = q_1 \sum_{n=1}^{N+1} \boldsymbol{\pi}_n (I_W \otimes \mathbf{S}_2 \otimes I_{T_{n-1}})\mathbf{e}.$$

The output rate 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 as

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

The probability $P_{fail-loss}$ of loss of an arbitrary customer due to service failure is computed as

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

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

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

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

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

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 rate $\tilde{\lambda}^{(r)}$ of changing the types-l, $l = \overline{r+1,R}$, of a customer to type-r, $r = \overline{1,R-1}$, is computed as

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

The average rate $\hat{\lambda}^{(r)}$ of of changing the types-l, $l = \overline{1,r-1}$, of a customer to the type-r, $r = \overline{2,R}$, is computed as

$$\hat{\lambda}^{(r)} = \sum_{l=1}^{r-1} \beta_l N_{buffer}^{(l)} q_{l,r}.$$

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

$$P_{imp-loss}^{(r)} = \frac{\lambda_{imp}^{(r)}}{\lambda_r + \tilde{\lambda}^{(r)} - \hat{\lambda}^{(r)}}.$$

Here, we assume that $\tilde{\lambda}^{(R)} = 0$ and $\hat{\lambda}^{(1)} = 0$.

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

$$\begin{aligned} P_{ent-loss}^{(r)} =& \lambda_r^{-1} \Bigg[\boldsymbol{\pi}_0 \sum_{k=N+2}^{K_r} (k-(N+1)) D_r^{(k)} \mathbf{e} + \\ & \sum_{n=1}^{N+1} \sum_{k=N+2-n}^{K_r} \boldsymbol{\pi}_n (k-(N+1-n))(D_r^{(k)} \otimes I_{MT_{n-1}})\mathbf{e} \Bigg], r = \overline{1,R}. \end{aligned}$$

The probability of an arbitrary customer loss upon arrival is

$$P_{ent-loss} = \lambda^{-1} \sum_{r=1}^{R} \Bigg[\boldsymbol{\pi}_0 \sum_{k=N+2}^{K_r} (k-(N+1)) D_r^{(k)} \mathbf{e} + \sum_{n=1}^{N+1} \sum_{k=N+2-n}^{K_r} \boldsymbol{\pi}_n (k-(N+1-n)) (D_r^{(k)} \otimes I_{MT_{n-1}}) \mathbf{e} \Bigg].$$

Remark 1. *To control the accuracy of calculations, the following equality can be used*

$$P_{loss} = P_{ent-loss} + P_{imp-loss} + P_{fail-loss}.$$

5. Numerical Example

In this numerical experiment, we assume that the number of types of customers is $R = 3$.

The matrices D_0 and $D_r^{(k)}$, which define the $BMMAP$, are given by

$$D_0 = \begin{pmatrix} -0.529945 & 0 \\ 0 & -0.5315542 \end{pmatrix},$$

$$D_1^{(1)} = \begin{pmatrix} 0.00643918 & 0.0070831 \\ 0.006117 & 0.006439 \end{pmatrix}, D_1^{(2)} = \begin{pmatrix} 0.0128784 & 0.00998072 \\ 0.0122344 & 0.0135223 \end{pmatrix},$$

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

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

$$D_3^{(1)} = \begin{pmatrix} 0.0482939 & 0.0321959 \\ 0 & 0 \end{pmatrix}, D_3^{(2)} = \begin{pmatrix} 0.0643918 & 0.0006439 \\ 0.0321959 & 0 \end{pmatrix},$$

$$D_3^{(3)} = \begin{pmatrix} 0.016098 & 0 \\ 0.0321959 & 0 \end{pmatrix}.$$

These matrices provide the value of intensities λ_r, $r = \overline{1,R}$, such as $\lambda_1 = 0.322758$, $\lambda_2 = 0.467236$, $\lambda_3 = 0.210007$. The total customers arrival intensity is $\lambda = 1$.

We assume that the PHF distribution of the service time is defined by the vectors $\boldsymbol{\sigma} = (0.05, 0.95)$, $\boldsymbol{S}_1 = (0.29, 13.3)^T$, $\boldsymbol{S}_2 = (0.01, 1)^T$, the matrix $S = \begin{pmatrix} -0.5 & 0.2 \\ 0.7 & -15 \end{pmatrix}$, and the probabilities $q_1 = 0.2$, $q_2 = 0.3$.

The mean service time (successful or not) is $b_1 = 0.258152$.

The intensities and probabilities that define the changes of customers priorities are the following: $\alpha_2 = 0.1$, $\alpha_3 = 0.15$, $\beta_1 = 0.01$, $\beta_2 = 0.04$, $p_{2,1} = 1$, $p_{3,1} = p_{3,2} = 0.5$, $q_{1,2} = 0.6$, $q_{1,3} = 0.4$, $q_{2,3} = 1$. The intensities of impatience are $\gamma_1 = 0.02$, $\gamma_2 = 0.015$, $\gamma_3 = 0.01$.

Let us vary the values of the buffer capacity N over the interval $[1, 40]$ with step 1. Also, we vary the values of the total arrival intensity λ over the interval $[1, 3]$ with step 0.25. It is done using multiplication of the matrices D_0 and $D_r^{(k)}$ by the corresponding intensity λ. For example, the $BMMAP$ with matrices $1.25D_0$ and $1.25D_r^{(k)}$ has the total arrival intensity $\lambda = 1.25$.

Figures 2–5 illustrate the dependence of the average number N_{buffer} of customer in the buffer and average number $N_{buffer}^{(r)}$ of type-r, $r = \overline{1,R}$, customers in the buffer on the buffer capacity N and average total arrival rate λ.

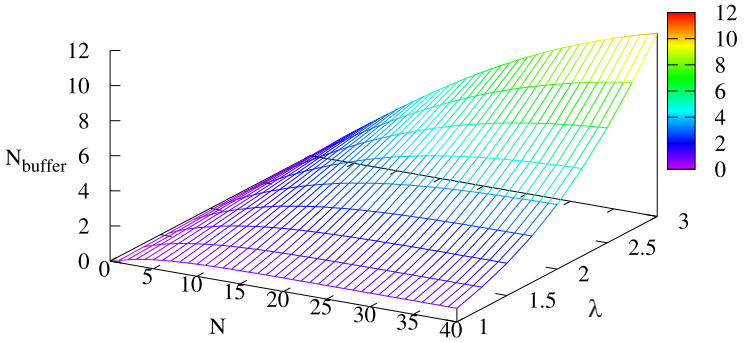

Figure 2. Dependence of N_{buffer} on N and λ.

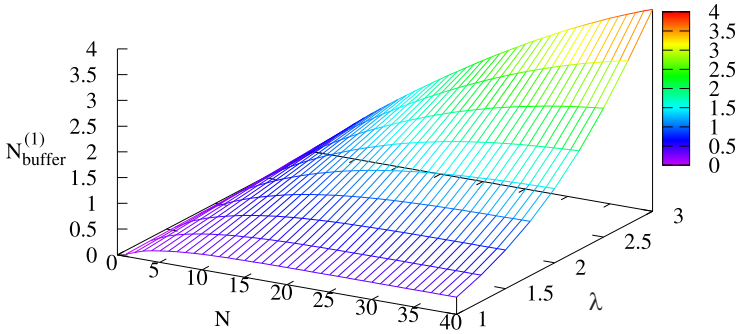

Figure 3. Dependence of $N^{(1)}_{buffer}$ on N and λ.

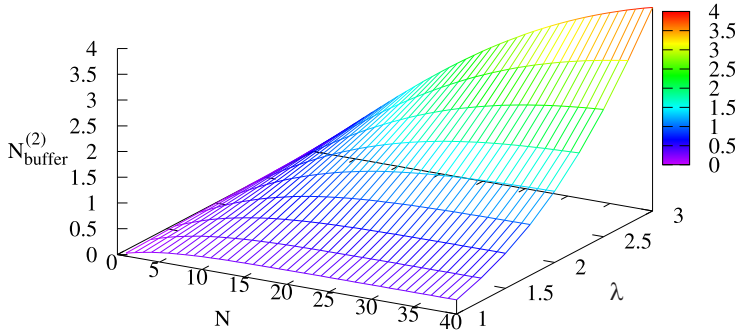

Figure 4. Dependence of $N^{(2)}_{buffer}$ on N and λ.

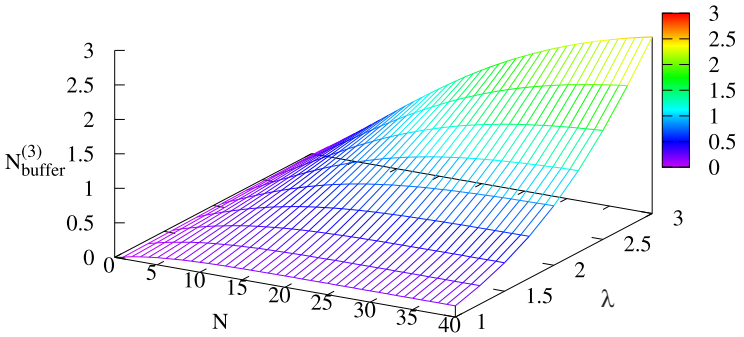

Figure 5. Dependence of $N^{(3)}_{buffer}$ on N and λ.

As it is seen from Figures 2–5, the average numbers of customers (total and of each type) in the buffer increase with the increase of the buffer capacity and the average arrival rate.

Figure 6 illustrates the dependence of the probability $P_{ent-loss}$ of an arbitrary customer loss upon arrival on the buffer capacity N and average total arrival rate λ.

The loss probability $P_{ent-loss}$ decreases with the increase of the buffer capacity N and increases with the increase of the arrival intensity λ. Our results allow us to estimate this intuitively clear dependence quantitatively.

Figure 7 shows the dependence of the probability $P_{fail-loss}$ of loss of an arbitrary customer due to service failure on the buffer capacity N and average total arrival rate λ.

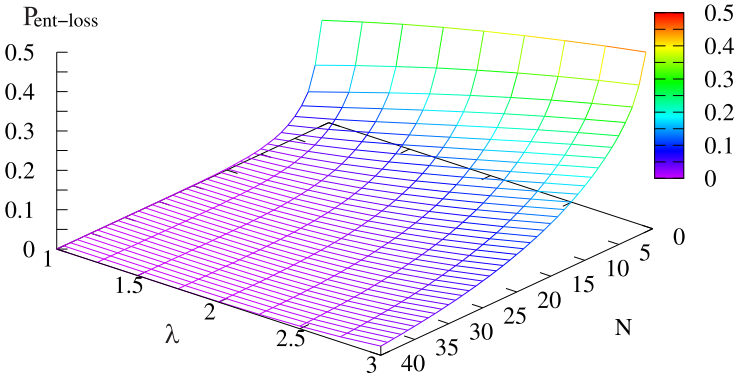

Figure 6. Dependence of the probability $P_{ent-loss}$ on N and λ.

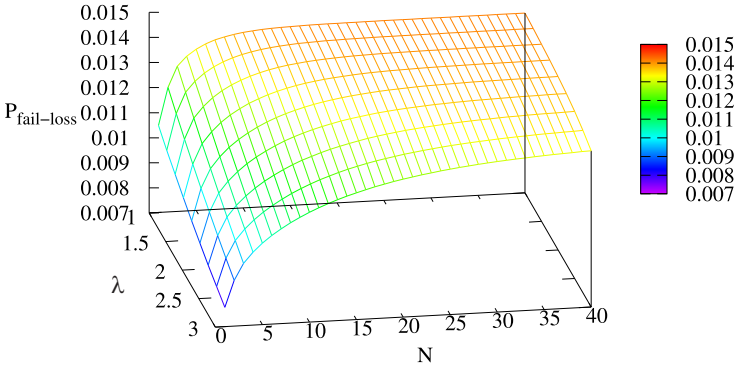

Figure 7. Dependence of the probability $P_{fail-loss}$ on N and λ.

The loss probability $P_{fail-loss}$ increases with the increase of the buffer capacity N and decreases with the increase of the arrival intensity λ. When the buffer capacity N increases and the arrival intensity λ decreases, the share of customers that succeed to reach the server grows, which implies the increase of the loss probability $P_{fail-loss}$.

Figure 8 illustrates the dependence of the probability $P_{imp-loss}$ of loss of an arbitrary customer due to impatience on the buffer capacity N and average total arrival rate λ.

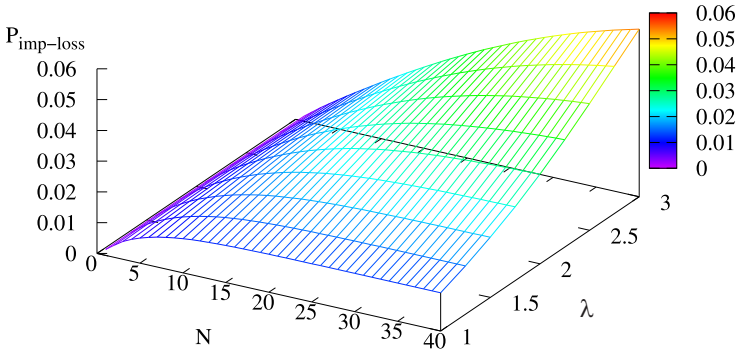

Figure 8. Dependence of the probability $P_{imp-loss}$ on N and λ.

The loss probability $P_{imp-loss}$ grows with the increase of the buffer capacity N. It can be explained as follows. With the growth of N, the number of customers staying in the buffer increases, and more customers leave the buffer due to impatience. The behavior of the loss probability $P_{imp-loss}$ with the increase of λ can be not monotonic. For example, for $N = 5$ for $\lambda = 1$ the loss probability $P_{imp-loss} = 0.008144564$, for $\lambda = 2.5$ the loss probability $P_{imp-loss} = 0.00967045$, and for $\lambda = 3$ the loss probability $P_{imp-loss} = 0.0095407$. This is because the increase of λ leads, on the one hand, to the increase in the number of customers in the buffer that has to increase the probability $P_{imp-loss}$, but on the other hand, the increase of λ implies the growth of the share of the customers that leave the system upon arrival and cannot be lost due to impatience.

Figure 9 illustrates the dependence of the loss probability of an arbitrary customer P_{loss} on the buffer capacity N and average total arrival rate λ.

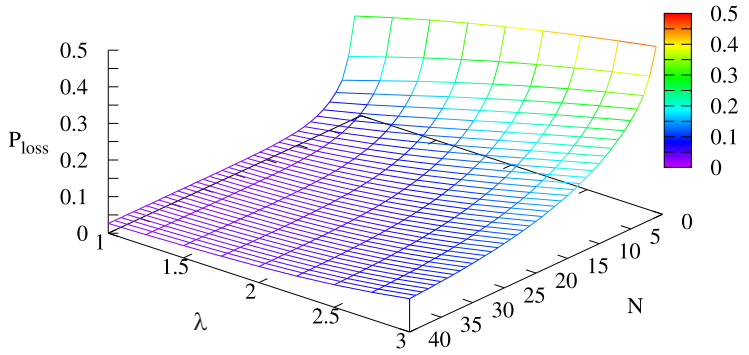

Figure 9. Dependence of the probability P_{loss} on N and λ.

As it is seen from Figure 9, the loss probability P_{loss} increases with the increase of the average total arrival rate λ and the decrease of the buffer capacity N.

Using the obtained results we can solve various optimization problems. For example, let the optimization problem be formulated as follows: it is required to determine the minimal buffer capacity N^* that guarantees the fulfillment of the inequality $P_{loss}(N^*) < 0.05$.

In the considered example, for the average total arrival rate $\lambda = 1$, the optimal value of the buffer capacity is $N^* = 8$ and the corresponding loss probability $P_{loss}(N^*)$ is equal to 0.04762911.

For the average total arrival rate $\lambda = 1.25$, the optimal value of the buffer capacity is $N^* = 11$ and the corresponding loss probability $P_{loss}(N^*)$ is equal to 0.047593474.

For the average total arrival rate $\lambda = 1.5$, the optimal value of the buffer capacity is $N^* = 15$ and the corresponding loss probability $P_{loss}(N^*)$ is equal to 0.04719013.

For the average total arrival rate $\lambda = 1.75$, the optimal value of the buffer capacity is $N^* = 19$ and the corresponding loss probability $P_{loss}(N^*)$ is equal to 0.04984426.

For the average total arrival rate $\lambda = 2$, the optimal value of the buffer capacity is $N^* = 27$ and the corresponding loss probability $P_{loss}(N^*)$ is equal to 0.049408936.

For the average total arrival rates $\lambda = 2.25$, 2.5, 2.75, and 3, it is impossible to determine the optimal buffer size only from the values presented on Figure 9 because $P_{loss}(40) > 0.05$ for all these arrival intensities. In the case of $\lambda = 3$, the optimal solution doesn't exist because the value of the probability $P_{imp-loss}$ of loss of an arbitrary customer due to impatience when $N = 40$ is equal to 0.05445345, and as we mentioned above, this probability grows with the increase of buffer capacity N. So, the loss probability P_{loss} in the case $\lambda = 3$ is greater than 0.05445345 for all $N \geq 40$.

In the cases $\lambda = 2.5$ and $\lambda = 2.75$, the probability $P_{imp-loss}$ of loss of a customer due to impatience when $N = 40$ is less than 0.05 (0.04119366 and 0.0478299, respectively). However, the sum of probabilities $P_{imp-loss}$ and $P_{fail-loss}$ is 0.05485347 in the case $\lambda = 2.5$, and 0.061301324 in the case $\lambda = 2.75$ and exceeds the value 0.05. Since the loss probability $P_{fail-loss}$ also increases with increase of N, the optimal solution doesn't exist in the case $\lambda = 2.5$ and $\lambda = 2.75$.

In the case of $\lambda = 2.25$, it is necessary to increase the buffer capacity to obtain the optimal solution. In the considered example, the optimal buffer capacity is $N^* = 52$, and the corresponding loss probability $P_{loss}(N^*)$ is equal to 0.049960257.

6. Conclusions

A single-server queueing model with a buffer of a finite capacity, heterogeneous correlated arrival process with the possibility of batch arrivals and non-pre-emptive priorities is analyzed. Customers are impatient with the impatience rate depending on the type of the customer. The server is unreliable. The priorities can be changed (increased or decreased) randomly in the Markov manner during the customer's stay in the buffer. The stationary behavior of the system having the listed features is analyzed via the analysis of the properly constructed Markov chain. The numerical results give some insight into the dependence of the main performance measures of the system on the total arrival rate and capacity of the buffer. The possibility of achieving the admissible value of an arbitrary customer loss probability via the proper choice of the buffer capacity is discussed.

As the directions for further research, generalizations of the model to the cases with a dependence of service time distribution on the type of a customer, possibility of the service pre-emption and the loss of the customers whose service is interrupted or their retrials deserve investigation in the future.

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

Funding: The publication was prepared with the support by the RUDN University Strategic Academic Leadership Program.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

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

References

1. 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]
2. Cao, P.; Xie, J. Optimal control of a multiclass queueing system when customers can change types. *Queueing Syst.* **2016**, *82*, 285–313. [CrossRef]
3. He, Q.-M.; Xie, J.; Zhao, X. Priority Queue with Customer Upgrades. *Nav. Res. Logist.* **2012**, *59*, 362–375. [CrossRef]
4. 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]
5. Fajardo, V.A.; Drekic, S. Waiting Time Distributions in the Preemptive Accumulating Priority Queue. *Methodol. Comput. Appl. Probab.* **2017**, *19*, 255–284. [CrossRef]
6. 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]
7. 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]
8. Stanford, D.A.; Taylor, P.; Ziedins, I. Waiting time distributions in the accumulating priority queue. *Queueing Syst.* **2014**, *77*, 297–330. [CrossRef]
9. Xie, O.; He, Q.-M.; Zhao, X. Stability of a priority queueing system with customer transfers. *Oper. Res. Lett.* **2008**, *36*, 705–709. [CrossRef]
10. 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]
11. 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]

12. Lee, S.K.; Dudin, S.; Dudina, O.; Kim, C.S.; Klimenok, V. A Priority Queue with Many Customer Types, Correlated Arrivals and Changing Priorities. *Mathematics* **2020**, *8*, 1292. [CrossRef]
13. Khalid, A.; Dudin, A.; Mushko, V. Novel queueing model for multimedia over downlink in 3.5 G wireless network. *J. Commun. Softw. Syst.* **2006**, *2*, 68–80.
14. Dudin, A.; Dudin, S. Analysis of a Priority Queue with Phase-Type Service and Failures. *Int. J. Stoch. Anal.* **2016**, *2016*, 9152701. [CrossRef]
15. 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]
16. Graham, A. *Kronecker Products and Matrix Calculus with Applications*; Horwood, E., Ed.; Courier Dover Publications: Cichester, UK, 1981.
17. Klimenok, V.I.; Dudin, A.N. Multi-dimensional asymptotically quasi-Toeplitz Markov chains and their application in queueing theory. *Queueing Syst.* **2006**, *54*, 245–259. [CrossRef]