mathematics

Article

Priority Multi-Server Queueing System with Heterogeneous Customers

Valentina Klimenok [1], Alexander Dudin [1] and Vladimir Vishnevsky [2,*]

[1] Department of Applied Mathematics and Computer Science, Belarusian State University, 220030 Minsk, Belarus; klimenok@bsu.by (V.K.); dudin@bsu.by (A.D.)

[2] Sciences and Closed Corporation Information and Networking Technologies, Institute of Control Sciences of Russian Academy, 119991 Moscow, Russia

* Correspondence: vishn@inbox.ru

Received: 2 August 2020; Accepted: 2 September 2020; Published: 4 September 2020

Abstract: In this paper, we analyze a multi-server queueing system with heterogeneous customers that arrive according to a marked Markovian arrival process. Customers of two types differ in priorities and parameters of phase type distribution of their service time. The queue under consideration can be used to model the processes of information transmission in telecommunication networks in which often the flow of information is the superposition of several types of flows with correlation of inter-arrival times within each flow and cross-correlation. We define the process of information transmission as the multi-dimensional Markov chain, derive the generator of this chain and compute its stationary distribution. Expressions for computation of various performance measures of the system, including the probabilities of loss of customers of different types, are presented. Output flow from the system is characterized. The presented numerical results confirm the high importance of account of correlation in the arrival process. The values of important performance measures for the systems with the correlated arrival process are essentially different from the corresponding values for the systems with the stationary Poisson arrival process. Measurements in many real world systems show poor approximation of real flows by such an arrival process. However, this process is still popular among the telecommunication engineers due to the evident existing gap between the needs of adequately modeling the real life systems and the current state of the theory of algorithmic methods of queueing theory.

Keywords: multi-server queueing system; heterogeneous customers; marked Markovian arrival process; priorities; loss probabilities

1. Introduction

One of the essential sections of the queuing theory is the theory of priority systems. In such systems, customers of different classes are assigned different categories of importance and service is carried out in accordance with a priority scheme. More important customers have an advantage in access to service compared to less important ones. Priority queueing models arise in many real world applications. In particular, the considered in this paper priority queueing model can be effectively used in various applications in telecommunication networks, where traffic prioritization is required, e.g., when using the IEEE 1588 synchronization protocol in cellular networks, DVB-T2 video transmission for synchronizing TV transmitters, in unmanned vehicle systems, in telemedicine applications, etc., it is necessary to ensure the guaranteed delivering of the highest priority packets. Similar models are used for voice and data transmission in multiprocessor switching nodes to ensure priority of voice traffic. The model has applications in other areas, including scheduling computations in multiprocessor systems, operation of medical institutions, etc. Various priority schemes are used in hospital emergency departments during sorting incoming patients according to the severity o, e.g.f the injury or disease, see, e.g., [1]. In the case of unreliable systems, a priority customer may be considered as equipment failure. Priority can be also set to maximize a company's profit or increase system utilization. For instance, an online store the manager can set a high priority for customers of big spenders in order to prevent their departure to other online resources, see [2]. In some telecommunication networks, the priority of a customer is determined by its owner through a Service Level Agreement (SLA), whereby certain customers have chosen to pay more so as to get high-priority access to some high-demand resource.

Another set of potential applications of the priority queueing model like considered in our paper is described in [3] as applications in customer service centers, see [4], airport security checkpoints [5], hospital emergency rooms [6], cloud computing systems [7] or processor management in certain computer operating systems [8].

There is an extensive literature on priority queues. An overview of early research works on priority queueing models can be found in monographs [9–13] and references therein. Mostly of these works focused on the queueing models with stationary Poisson inputs. However, the area of application of these models is currently being greatly narrowed, since flows in modern telecommunication networks do not possess memoryless property of stationary Poisson flow. They are, as a rule, correlated and heterogeneous. In the case of homogeneous customers, a good mathematical model of such flows well known in the literature as a Markovian arrival process (MAP), see, e.g., [14–16]. A MAP is the significant generalization of the stationary Poisson process to the case of correlated bursty traffic. Queuing systems with a MAP and priorities are discussed in the papers [17–26].

Markovian arrival processes are a suitable class of stochastic processes to represent correlated traffic in case when all customers are of the same type and only the sequence of inter-arrival times is of interest. However, systems with correlated flows of heterogeneous customers are also of great interest to applications. Such flows are well modeled by a marked Markovian arrival process ($MMAP$), see [27]. Using of a $MMAP$ allows to represent flows where inter-arrival times are correlated across customer classes and to achieve models of greater accuracy. However, very few results are known about priority queue with $MMAP$. We can only refer to the papers [1,25,26,28]. In [28] $MMAP/MAP/1$ queue with preemptive priorities is analysed, the moments of queue size is derived. The article [1] deals with a multi-server queue with a $MMAP$, preemptive priorities and priority upgrade. For this system, the authors found a condition for the existence of a stationary regime and the bounds for the lengths of the queues. The problem of concrete computation of the stationary distribution of the system states is not touched in that paper. This problem is considered in the recent papers [25,26]. However, those papers deal with single server priority queues. In [25], possibility of increase of the non-preemptive priority during customer stay in the buffer is analysed. In [26], the discipline of flexible providing of non-preemptive priorities is under study.

In this paper, we consider a priority $MMAP/PH_{1,2}/N/N$ queueing system with two types of customers. As it is pointed out above, although analysis of such multi-server systems is very important for practice, these systems did not get a proper portion of attention of queueing theorists. e.g., the authors of [3] note that "Despite the large number of systems that can be viewed as instances of a priority queue with multiple servers, the literature devoted to their theoretical analysis appears rather moderate as the inherent complexity of these queues hinders their analysis". It is also mentioned in [3] that "To the best of our knowledge, few exact results exist in the case of priority queues with multiple servers even under the simplest assumption of exponentially distributed service times". In paper [3], the authors analyze the model similar to the considered in our paper with two essential differences: (i) Several priority classes are suggested in [3] while we consider the case of two classes, (ii) arrival flow is assumed be the renewal with phase type distribution of mutually independent inter-arrival times while we assume that the inter-arrival times can be dependent. One more advantages of our paper consists of the fact that we provide exact analysis of the model while only the accurate approximate solution is proposed in [3].

In this paper, we make some step to provide the analysis of a multi-server priority system. We assume that customers of one of the types have the preemptive priority. The service times of customers of both types have phase type distribution (PH) with different parameters. Exact description of the model is given in Section 2. The most important performance measures of the considered system are probabilities of loss of customers of different types. To calculate these probabilities, we, first, described the process of the system operation by a multi-dimensional Markov chain. This description is given in Section 3. The explicit form of the generator of this Markov is given in that section along with the short explanation of the form of the blocks of this generator. In Section 4, we briefly note how to compute stationary distribution of this Markov chain. We do not go into details of computations and give just an advice how to compute this distribution. It is worth to mention that although the state space of the Markov chain is finite, in situation when the number of servers N and (or) state spaces of the underlying processes of arrival and service process are large, computation of this distribution is far from the trivial. Using the stationary distribution been calculated, in Section 4 we also derive loss probabilities associated with the system and a number of other characteristics. Numerical results are provided in Section 5.

2. Model Description

We consider an N -server queueing system without a buffer. Customers of two types arrive to the system according to a $MMAP$. The arrivals in the $MMAP$ is directed by the underlying process $\nu_t,\ t \geq 0$, which is an irreducible continuous time Markov chain with the state space $\{0, 1, ..., W\}$. In case of two types of customers, the $MMAP$ is completely defined by the state space of underlying process and $(W + 1) \times (W + 1)$ matrices $D_k,\ k = 0, 1, 2$, or their generating function $D(z) = D_0 + D_1 z + D_2 z^2$. The entries of the matrix D_k give the rate of transitions of the process $\nu_t,\ t \geq 0$, which are accompanied by generating a customer of type $k, k = 1, 2$. Non-diagonal entries of the matrix D_0 describe the rate of transitions of the process $\nu_t,\ t \geq 0$, which are not accompanied by generating a customer. Diagonal entries of the matrix D_0 are negative and such that the matrix $D(1) = D_0 + D_1 + D_2$ is an infinitesimal generator of the chain $\nu_t,\ t \geq 0$. The fundamental rate of arrivals of type-k customers is calculated as $\lambda_k = \boldsymbol{\theta}\, D_k \mathbf{e}$, $k = 1$, 2, where $\boldsymbol{\theta}$ is a row vector of the steady state probabilities of the underlying process ν_t. The vector $\boldsymbol{\theta}$ is the unique solution to the system $\boldsymbol{\theta} D(1) = \mathbf{0},\ \boldsymbol{\theta}\mathbf{e} = 1$. Here and in the sequal $\mathbf{e}$ is a column vector consisting of 1's, and $\mathbf{0}$ is a row vector consisting of 0's. The total fundamental rate of arrivals is $\lambda = \lambda_1 + \lambda_2$. The variance of inter-arrival times of customers of type k is calculated as

$$v^{(k)} = \frac{2\boldsymbol{\theta}(-D_0 - D_{\bar{k}})^{-1}\mathbf{e}}{\lambda_k} - \left(\frac{1}{\lambda_k}\right)^2,\ \ \bar{k} \neq k,\ k, \bar{k} = 1, 2.$$

The coefficient of correlation of lengths of two successive inter-arrival times of k-type of customers is calculated as

$$c_k = \left[\frac{\boldsymbol{\theta}(D_0 + D_{\bar{k}})^{-1}}{\lambda_k} D_k (D_0 + D_{\bar{k}})^{-1}\mathbf{e} - \left(\frac{1}{\lambda_k}\right)^2\right](v^{(k)})^{-1},\ \bar{k} \neq k,\ k, \bar{k} = 1, 2.$$

More details about a $MMAP$ can be found in [27].

The service time of the k-type customer has PH type distribution with irreducible representation $(\boldsymbol{\beta}_k, S_k)$. Here $\boldsymbol{\beta}_k$ is a row vector of size M_k, and S_k is a square matrix of size M_k. Thus, the specified service time is interpreted as the time during which an underlying Markov chain $m_t^{(k)}, t \geq 0$, with state space $\{1, \ldots, M_k, M_k + 1\}$ will reach the only absorbing state $M_k + 1$. The transition rates of the chain $m_t^{(k)}, t \geq 0$, within the space of transient states $\{1, \ldots, M_k\}$ are defined by the sub-generator S_k, and the rates of transitions to the absorbing state are defined by the vector $\mathbf{S}_0^{(k)} = -S_k \mathbf{e}$. At the time the service starts, the state of the process $m_t^{(k)},\ t \geq 0$, is selected from the state space $\{1, \ldots, M_k\}$ according to the probability row vector $\boldsymbol{\beta}_k$. The service rate are calculated as $\mu_k = -(\boldsymbol{\beta}_k S_k^{-1}\mathbf{e})^{-1}$. More information about the PH type distribution can be found, e.g., in [17,29].

We assume that customers of type 1 have the preemptive priority. If a priority customer arrives to the system when all servers are busy and there are servers occupied with non-priority customers, then the arriving priority customer crowds out one of these customers (which is lost) and takes his/her place on the server. If all servers are occupied with priority customers, then an arriving priority customer is lost. If a non-priority customer arrives at the system when all server are busy, he/she leaves the system forever.

3. Process of the System States

Let at the time t,

- n_t be the number of busy servers , $n_t = \overline{0,N}$;
- r_t be the number of servers serving type 1 customers, $r_t = \overline{0,n_t}$;
- ν_t be the state of the underlying process $\nu_t = \overline{0,W}$;
- $m_t^{(j,k)}$ be the state of underlying process of service on jth server servicing type k customer, $m_t^{(j,1)} = \overline{1,r_t}, m_t^{(j,2)} = \overline{1,n_t - r_t}$. We assume the following dynamical enumeration of the busy servers. The servers, which provide service to customers of type 2, are located after servers serving customers of type 1. In addition, we assume that the servers serving customers of the k-th type are numbered in the order of their occupation, i.e. the server that starts the service is numbered by the maximum number among all servers engaged in servicing customers of this type. When some server finishes the service, the corresponding renumbering of servers occurs.

The operation of the queue under consideration is described by the Markov chain

$$\xi_t = \{n_t, r_t, \nu_t, m_t^{(1,1)}, m_t^{(2,1)}, \ldots, m_t^{(r_t,1)}, m_t^{(1,2)}, m_t^{(2,2)}, \ldots, m_t^{(n_t-r_t,2)}\}$$

with the state space

$$\Omega = \{(n, r, \nu, m^{(j,1)}, m^{(l,2)}),\ n = \overline{0,N}, r = \overline{0,n},\ \nu = \overline{0,W},\ m^{(j,1)} = \overline{1,M_1}, j = \overline{1,r},$$

$$m^{(l,1)} = \overline{1,M_2}, l = \overline{1,n-r}\}.$$

It can be calculated that cardinality of the set Ω is equal to $K = (W+1)(1 + \sum\limits_{n=1}^{N} \sum\limits_{l=0}^{n} M_1^l M_2^{n-l})$.

Introduce the following notation:

$\mathbf{e}_n$ is a column vector of size n, consisting of 1's;

I (O) is an identity (zero) matrix of appropriate dimension. When needed we will identify the dimension of this matrix with suffix;

$diag\,\{A_l,\ l = \overline{1,L}\,\}$ is a diagonal matrix with diagonal blocks A_l;

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

$diag^+\{A_l, l = \overline{0,L}\}$ is an over-diagonal matrix with the over-diagonal blocks A_l;

$\otimes$ and $\oplus$ are the symbols of the Kronecker product and sum of matrices, see [30];

$A^{\otimes l} = \underbrace{A \otimes \ldots \otimes A}_{l},\ l \geq 1,\ A^{\otimes 0} = 1;$

$A^{\oplus l} = \sum\limits_{m=0}^{l-1} I_{n^m} \otimes A \otimes I_{n^{l-m-1}},\ l \geq 1$, for the matrix A having n rows;

$\bar{W} = W + 1$;

Let us arrange the states of the chain ξ_t in the lexicographic order and form the matrices $Q_{n,n'},\ n, n' = \overline{0,N}$, consisting of the rates of the chain transition from the states corresponding to the value n of the first component to the states corresponding to the value n' of this component. Then the infinitesimal generator of the chain is defined by the following theorem.

Theorem 1. *The infinitesimal generator of the Markov chain ξ_t has the following block structure:*

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

where

$$Q_{n,n} = diag\{D_0 \oplus S_1^{\oplus r} \oplus S_2^{\oplus n-r},\ r = \overline{0,n}\},\ n = \overline{0,N-1},$$

$$Q_{N,N} = diag\{(D_0 + D_2)) \oplus S_1^{\oplus r} \oplus S_2^{\oplus N-r},\ r = \overline{0,N-1},\ D(1) \oplus S_1^{\oplus N}\}$$

$$+diag^{+}\{D_1 \otimes I_{M_1^r} \otimes \mathbf{e}_{M_2}\boldsymbol{\beta}_1 \otimes I_{M_2^{N-r-1}}, r = \overline{0,N-1}\},$$

$$Q_{n,n-1} = \left(\begin{array}{c} diag\{I_{\bar{W}} \otimes I_{M_1^r} \otimes (\boldsymbol{S}_0^{(2)})^{\oplus n-r},\ r = \overline{0,n-1}\} \\ \hline O_{\bar{W}M_1^n \times \bar{W}\sum\limits_{r=0}^{n-2} M_1^r M_2^{n-r-1}} \ \mid\ I_{\bar{W}} \otimes (\boldsymbol{S}_0^{(1)})^{\oplus n} \end{array}\right)$$

$$+diag^{-}\{I_{\bar{W}} \otimes (\boldsymbol{S}_0^{(1)})^{\oplus r} \otimes I_{M_2^{n-r}}\ r = \overline{1,n}\}, n = \overline{1,N},$$

$$Q_{n,n+1} = \left(\begin{array}{c} diag\{D_2 \otimes I_{M_1^r} \otimes I_{M_2^{n-r}}\boldsymbol{\beta}_2, r = \overline{0,n}\} \ \mid\ O_{\bar{W}\sum\limits_{r=0}^{n} M_1^r M_2^{n-r} \times \bar{W}M_1^{n+1}} \end{array} \right)$$

$$+ \left(\begin{array}{c} O_{\bar{W}\sum\limits_{r=0}^{n-1} M_1^r M_2^{n-r} \times \bar{W}\sum\limits_{r=0}^{n+1} M_1^r M_2^{n-r+1}} \\ \hline O_{\bar{W}M_1^n \times \bar{W}\sum\limits_{r=0}^{n} M_1^r M_2^{n-r+1}} \mid D_1 \otimes I_{M_1^n} \otimes \boldsymbol{\beta}_1 \end{array}\right)$$

$$+ \left(\begin{array}{c} diag^{+}\{D_1 \otimes I_{M_1^r} \otimes \boldsymbol{\beta}_1 \otimes I_{M_2^{n-r}},\ r = \overline{0,n-1}\} \mid O_{\bar{W}\sum\limits_{r=0}^{n} M_1^r M_2^{n-r} \times \bar{W}M_1^{n+1}} \end{array} \right),$$

$$n = \overline{1,N-1}.$$

Proof. The tri-block diagonal form of the generator is easily explained by the evident fact that the customers arrive to the considered system and depart from this system only one-by-one. Therefore, all blocks $Q_{n,n'}$ of the generator are equal to zero matrices if $|n - n'| > 1$. Before to immediately prove the expressions for the nonzero blocks of the generator, let us rewrite some blocks in the less compact but more transparent for explanations form:

$$Q_{N,N} =$$

$$\begin{pmatrix} \bar{D} \oplus S_2^{\oplus N} & D_1 \otimes \mathbf{e}_{M_2}\boldsymbol{\beta}_1 \otimes I_{M_2^{N-1}} & O & \dots & O \\ O & \bar{D} \oplus S_1 \oplus S_2^{\oplus N-1} & D_1 \otimes I_{M_1} \otimes \mathbf{e}_{M_2}\boldsymbol{\beta}_1 \otimes I_{M_2^{N-2}} & \dots & O \\ O & O & \bar{D} \oplus S_1^{\oplus 2} \oplus S_2^{\oplus N-2} & \dots & O \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ O & O & O & \dots & D_1 \otimes I_{M_1^{N-1}} \otimes \mathbf{e}_{M_2}\boldsymbol{\beta}_1 \\ O & O & O & \dots & D(1) \oplus S_1^{\oplus N} \end{pmatrix},$$

where $\bar{D} = D_0 + D_1$,

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

$$\begin{pmatrix} I_{\bar{W}} \otimes (\boldsymbol{S}_0^{(2)})^{\oplus n} & O & O & \dots & O \\ I_{\bar{W}} \otimes \boldsymbol{S}_0^{(1)} \otimes I_{M_2^{n-1}} & I_{\bar{W}M_1} \otimes (\boldsymbol{S}_0^{(2)})^{\oplus n-1} & O & \dots & O \\ O & I_{\bar{W}} \otimes (\boldsymbol{S}_0^{(1)})^{\oplus 2} \otimes I_{M_2^{n-2}} & I_{\bar{W}M_1^2} \otimes (\boldsymbol{S}_0^{(2)})^{\oplus n-2} & \dots & O \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ O & O & O & \dots & I_{\bar{W}M_1^{n-1}} \otimes \boldsymbol{S}_0^{(2)} \\ O & O & O & \dots & I_{\bar{W}} \otimes (\boldsymbol{S}_0^{(1)})^{\oplus n} \end{pmatrix},$$

$$n = \overline{1, N},$$

$$Q_{n,n+1} =$$

$$\begin{pmatrix} D_2 \otimes I_{M_2^n} \otimes \boldsymbol{\beta}_2 & D_1 \otimes \boldsymbol{\beta}_1 \otimes I_{M_2^n} & O & \dots & O \\ O & D_2 \otimes I_{M_1 M_2^{n-1}} \otimes \boldsymbol{\beta}_2 & D_1 \otimes I_{M_1} \otimes \boldsymbol{\beta}_1 \otimes I_{M_2^{n-1}} & \dots & O \\ O & O & D_2 \otimes I_{M_1^2 M_2^{n-2}} \otimes \boldsymbol{\beta}_2 & \dots & O \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ O & O & O & \dots & D_1 \otimes I_{M_1^n} \otimes \boldsymbol{\beta}_1 \end{pmatrix},$$

$$n = \overline{0, N-1}.$$

Now, to prove the presented forms of the non-zero blocks $Q_{n,n'}$ of the generator, we analyze transitions of the Markov chain $\xi_t, t \geq 0$, during the infinitesimal length interval.

These blocks for various values of n and n' have the following meaning:

- The non-diagonal entries of the blocks $Q_{n,n}, n = \overline{0, N}$, define the transition rates of the chain $\xi_t, t \geq 0$, which do not lead to the change in the number n of busy servers. The diagonal entries of the blocks $Q_{n,n}$ define the departure rates of the chain $\xi_t, t \geq 0$, from the corresponding states. If $n < N$ and the number of priority customers in the service is r, transitions, which do not lead to the change of the number n of busy servers, occur either when the underlying process of the $MMAP$ makes an idle transition (i.e., a transition without generation of any customers) or a phase of service time of one of r priority customers is changed or a phase of service time of one of $n - r$ non-priority customers is changed. The corresponding transition rates are described by the matrix $D_0 \oplus S_1^{\oplus r} \oplus S_2^{\oplus n-r}$, $r = \overline{0, n}$, $n = \overline{0, N-1}$. It is worth to mention here that the operations of the Kronecker product and sum of matrices are very useful for description of transition rates or transition probabilities of the multi-dimensional random processes with independent Markovian components.

In the case $n = N$ and $r < N$ the number of busy servers and the number of priority customers in the service do not change when $MMAP$ makes an idle transition or a non-priority customer arrives (this customer is lost). The corresponding transition rates are described by the matrix $(D_0 + D_2) \oplus S_1^{\oplus r} \oplus S_2^{\oplus N-r}$. If in this case a priority customer arrives, he/she pushes a non-priority customer from the service (which is lost) and takes his/her place on the server. In this case the number of priority customers in the service becomes equal to $r + 1$. The corresponding transition rates are described by the matrix $D_1 \otimes I_{M_1^r} \otimes \mathbf{e}_{M_2} \boldsymbol{\beta}_1 \otimes I_{M_2^{N-r}}$. If at an arrival time all servers are occupied by priority customers, i.e., $r = N$, an arriving customer, regardless of its priority, is lost. The corresponding transition rates are described by the matrix $D(1) \oplus S_1^{\oplus N}$.

- The blocks $Q_{n,n-1}, n = \overline{1, N}$, define the transition rates of the chain $\xi_t, t \geq 0$, which lead to a decrease in the number of busy servers from n to $n - 1$. If the number of priority customers in the service is equal to r, such a transition occurs when the service of one of priority customers ends (the corresponding transition rates of the chain $\xi_t, t \geq 0$, are described by the matrix $I_{\bar{W}} \otimes (\boldsymbol{S}_0^{(1)})^{\oplus r} \otimes I_{M_2^{n-r}}$) or when the service of one of non- priority customers ends (the corresponding transition rates are described by the matrix $I_{\bar{W}} \otimes I_{M_1^r} \otimes (\boldsymbol{S}_0^{(2)})^{\oplus n-r}$).

- The blocks $Q_{n,n+1}, n = \overline{0, N-1}$, define the transition rates of the chain $\xi_t, t \geq 0$, which lead to an increase in the number of busy servers from n to $n+1$. Such an increase occurs if a non-priority customer arrives (the corresponding transition rates are defined by the matrix $D_2 \otimes I_{M_1^r} \otimes I_{M_2^{n-r}} \otimes \boldsymbol{\beta}_2$ or a priority customer arrives (the corresponding transition rates are defined by the matrix $D_1 \otimes I_{M_1^r} \otimes \boldsymbol{\beta}_1 \otimes I_{M_2^{n-r}}$).

This completes the proof of the theorem. □

4. Stationary Distribution. Performance Measures

The Markov chain ξ_t is irreducible and admits the values in a finite state space. Therefore, a unique stationary distribution of this chain exists for any values of the system parameters. Let $\mathbf{p}$ be the row vector of the steady state (stationary) probabilities of the states of the chain enumerated in the lexicographic order. It is well known that the vector $\mathbf{p}$ is defined as the unique solution of the Chapman-Kolmogorov (equilibrium or balance) equations

$$\mathbf{p}Q = \mathbf{0}, \;\; \mathbf{pe} = 1.$$

This system can be solved by any of the well-known methods for solving the finite system of linear algebraic equations. However, in the case of a large dimension of this system, the solution of this system can be not trivial due to existing restrictions on the computer memory and computation speed. Therefore, for solution of this system it is advisable to use a special stable algorithm proposed in [31] and based on the idea of substituting this system of equations by an alternative system derived via consideration of a sequence of specially constructed so called censored Markov chains.

As the result of computations, we obtain the partitioned vector $\mathbf{p} = (\mathbf{p}_0, \mathbf{p}_1, \ldots, \mathbf{p}_N)$ where $\mathbf{p}_n$ is a row vector of steady state probabilities corresponding to the value n of the first component of the Markov chain, $n = \overline{0, N}$. Having the vectors $\mathbf{p}_n$ be calculated, we are able to calculate a number of stationary performance measures of the queue under consideration.

- The distribution of the number of busy servers at an arbitrary time $p_n = \mathbf{p}_n\mathbf{e}$, $n = \overline{0, N}$.
- The average number of busy servers $N_{busy} = \sum\limits_{n=1}^{N} np_n$.
- The distribution of the number of servers providing service to priority customers

$$q_r^{(1)} = \sum_{n=r}^{N} \mathbf{p}_n \mathbf{u}(r, n-r),\; r = \overline{0, N}, \tag{1}$$

where $\mathbf{u}(r, n-r) = \begin{pmatrix} \mathbf{0}^T_{\bar{W} \sum\limits_{l=0}^{r-1} M_1^l M_2^{n-l}} \\ \mathbf{e}_{\bar{W} M_1^r M_2^{n-r}} \\ \mathbf{0}^T_{\bar{W} \sum\limits_{l=r+1}^{n} M_1^l M_2^{n-l}} \end{pmatrix}$.

To make clear Formula (1), we note that, multiplying the vector $\boldsymbol{p}_n$ by the vector $\mathbf{u}(r, n-r)$, we select and sum up the entries of the probability vector $\boldsymbol{p}_n$ which correspond to the states with n busy servers of which r servers are busy with priority customers. Summing the results over n, we obtain the probability $q_r^{(1)}$.

- The average number of servers providing service to priority customers $N_{busy}^{(1)} = \sum\limits_{r=1}^{N} rq_r^{(1)}$.
- The distribution of the number of servers providing service to non-priority customers

$$q_m^{(2)} = \sum_{n=m}^{N} \mathbf{p}_n \mathbf{u}(n-m, m),\; m = \overline{0, N}. \tag{2}$$

Formula (2) is explained similarly to Formula (1).

- The average number of servers providing service to non-priority customers $N_{busy}^{(2)} = \sum\limits_{m=1}^{N} mq_m^{(2)}$. Evidently, $N_{busy} = N_{busy}^{(1)} + N_{busy}^{(2)}$ and this relation can be used for control of accuracy of computations.
- The probability that a priority customer will be lost

$$P_{loss,1} = \frac{1}{\lambda_1}\mathbf{p}_N \begin{pmatrix} O_{\bar{W}\sum\limits_{r=0}^{N-1} M_1^r M_2^{N-r} \times \bar{W}} \\ I_{\bar{W}} \otimes \mathbf{e}_{M_1^N} \end{pmatrix} D_1\mathbf{e}. \qquad (3)$$

The brief explanation of Formula (3) can be done as follows. A priority customer will be lost if at the moment of his/her arrival all servers are occupied with priority customers. The νth entry of the vector $\mathbf{p}_N \begin{pmatrix} O_{\bar{W}\sum\limits_{r=0}^{N-1} M_1^r M_2^{N-r} \times \bar{W}} \\ I_{\bar{W}} \otimes \mathbf{e}_{M_1^N} \end{pmatrix}$ gives the probability that an arbitrary time all server are busy with priority customers and the underlying process of the $MMAP$ is in the state $\nu, \nu = \overline{0, W}$. Multiplying this vector by $\frac{D_1\mathbf{e}}{\lambda_1}$, we obtain the probability that at the moment of the priority customer arrival all servers are busy. In this case, the priority customers is lost.

- The probability that a non-priority customer will be lost due to lack of free servers at the moment of his/her arrival

$$P_{loss,2}^{input} = \frac{1}{\lambda_2}\mathbf{p}_N \begin{pmatrix} I_{\bar{W}} \otimes \mathbf{e}_{M_1^0 M_2^N} \\ I_{\bar{W}} \otimes \mathbf{e}_{M_1^1 M_2^{N-1}} \\ \vdots \\ I_{\bar{W}} \otimes \mathbf{e}_{M_1^N M_2^0} \end{pmatrix} D_2\mathbf{e}. \qquad (4)$$

Formula (4) is explained similarly to Formula (3).

- The probability that an arriving priority customer pushes out a non-priority customer from the server

$$P_{loss,2}^{serv} = \frac{1}{\lambda_1}\mathbf{p}_N \begin{pmatrix} I_{\bar{W}} \otimes \mathbf{e}_{M_1^0 M_2^N} \\ I_{\bar{W}} \otimes \mathbf{e}_{M_1^1 M_2^{N-1}} \\ \vdots \\ I_{\bar{W}} \otimes \mathbf{e}_{M_1^{N-1} M_2^1} \\ I_{\bar{W}} \otimes \mathbf{0}_{M_1^N M_2^0} \end{pmatrix} D_1\mathbf{e}. \qquad (5)$$

The brief explanation of Formula (5) is as follows. The probability $P_{loss,2}^{serv}$ is calculated by considering the situation at the moment of an arrival of a priority customer which meets all servers busy and at least one server occupied with a non-priority customer.

The νth entry of the vector $\mathbf{p}_N \begin{pmatrix} I_{\bar{W}} \otimes \mathbf{e}_{M_1^0 M_2^N} \\ I_{\bar{W}} \otimes \mathbf{e}_{M_1^1 M_2^{N-1}} \\ \vdots \\ I_{\bar{W}} \otimes \mathbf{e}_{M_1^{N-1} M_2^1} \\ I_{\bar{W}} \otimes \mathbf{0}_{M_1^N M_2^0} \end{pmatrix}$ gives the probability that at an arbitrary time all servers are busy, at least one server is occupied with a non-priority customer and the underlying process of the $MMAP$ is in the state $\nu, \nu = \overline{0, W}$. Multiplying this vector by $\frac{D_1\mathbf{e}}{\lambda_1}$, we obtain the probability that at the moment of a priority customer arrival all servers are busy and at least one server is occupied with a non-priority customer. In this case the priority customer pushes out a non-priority customer from the server and takes his/her place.

- The probability that an arbitrary customer arriving to the system will be lost due to the lack of free servers
$$P_{loss} = \frac{\lambda_1 P_{loss,1} + \lambda_2 P^{input}_{loss,2}}{\lambda}. \tag{6}$$
The numerator in the right hand side of Formula (6) is the rate of lost customers of two types and the denominator is the total input rate. The probability P_{loss} is calculated as the ratio of these rates.

5. Numerical Experiments

In this section, we present the results of numerical experiments that allow us to estimate the effect of the input rate λ and correlation in the $MMAP$ on the system performance measures.

We consider three $MMAP$s with the same arrival rates of customers of both types but different coefficients of correlation. These $MMAP$s are defined by the matrix D_0, D_1, D_2 as follows. For each $MMAP$, to get the matrices D_1, D_2, we first define a certain matrix D and then split it into the matrices D_1, D_2 in the proportion $D_1 = 0.7D$, $D_2 = 0.3D$.

The first $MMAP$ is the superposition of two stationary Poisson flows. In this case the matrices D_0, D are defined as follows:
$$D_0 = -6.124137, \ D = 6.124137.$$

For this $MMAP$, the coefficients of variation of inter-arrival times are $c^{(1)}_{var} = c^{(2)}_{var} = 1$, and the coefficients of correlation of inter-arrival times are $c_1 = c_2 = 0$.

The second $MMAP$ is defined by the matrices
$$D_0 = \begin{pmatrix} -8.281261 & 0 \\ 0 & -0.268743 \end{pmatrix}, \ D = \begin{pmatrix} 8.226134 & 0.055127 \\ 0.149638 & 0.119104 \end{pmatrix}.$$

For this $MMAP$, the coefficients of variation of inter-arrival times are $c^{(1)}_{var} = 1.693996$, $c^{(2)}_{var} = 3.417903$, and the coefficients of correlation of inter-arrival times are $c_1 = 0.023423$, $c_2 = 0.187824$.

The third $MMAP$ is defined by the matrices
$$D_0 = \begin{pmatrix} -29.668039 & 0.003450 \\ 0.006900 & -0.952137 \end{pmatrix}, \ D = \begin{pmatrix} 29.323061 & 0.341527 \\ 0.068995 & 0.876242 \end{pmatrix}.$$

For this $MMAP$, the coefficients of variation of inter-arrival times are $c^{(1)}_{var} = 2.394561$, $c^{(2)}_{var} = 3.087863$, and the coefficients of correlation of inter-arrival times are $c_1 = 0.205982$, $c_2 = 0.402641$.

The number of servers $N = 8$.

The service time of a priority customer has Erlang distribution defined by the vector $\boldsymbol{\beta}_1 = \begin{pmatrix} 1 & 0 \end{pmatrix}$ and the matrix $S_1 = \begin{pmatrix} -40 & 40 \\ 0 & -40 \end{pmatrix}$. The service time of a non-priority customer has Erlang distribution defined by the vector $\boldsymbol{\beta}_1 = \begin{pmatrix} 1 & 0 \end{pmatrix}$ and the matrix $S_1 = \begin{pmatrix} -10 & 10 \\ 0 & -10 \end{pmatrix}$.

Experiment 1. In the experiment, we investigate the behavior of loss probabilities associated with the system under consideration. We consider the probabilities $P_{loss,1}$, $P^{input}_{loss,2}$ and $P^{serv}_{loss,2}$ as functions of the input rate λ and coefficients of correlation c_1 and c_2.

Figure 1 depicts the dependence of the probability that a priority customer will be lost, $P_{loss,1}$, on the input rate λ and coefficients of correlation c_1 and c_2.

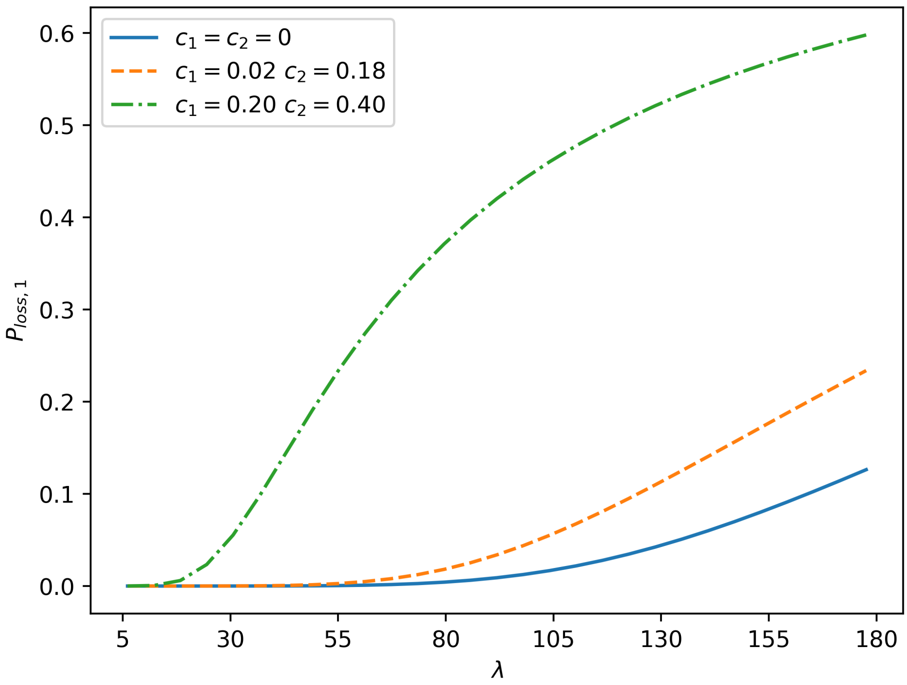

Figure 1. The dependence of the probability $P_{loss,1}$ on the input rate λ for the marked Markovian arrival processes ($MMAP$s) with different coefficients of correlation.

As expected, the value of $P_{loss,1}$ increases with increasing λ. Under the same value of λ, this probability is essentially greater for the larger coefficients of correlation in the $MMAP$. This effect is easily explained intuitively. The positive correlation in the arrival process causes fluctuation of the instantaneous arrival rate. Periods of time when customers arrive frequently (and likely a lot of customers is lost due to the business of all servers) alternate with the periods when customers arrive rarely (and likely starvation of the servers occurs).

Figure 2 shows the dependence of the probability of losing non-priority customers due to occupancy of all servers, $P_{loss,2}^{input}$, on the input rate λ and coefficients of correlation c_1 and c_2.

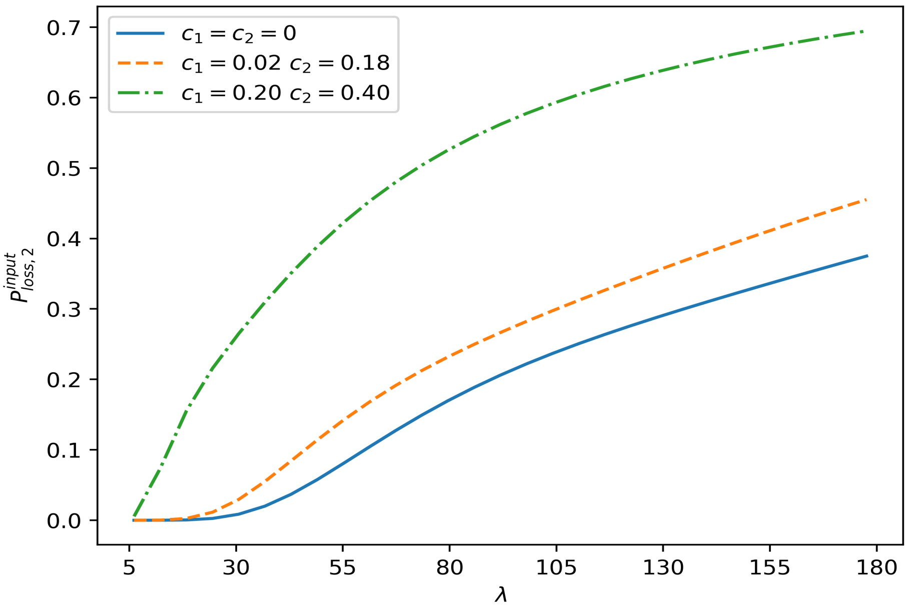

Figure 2. The dependence of the probability $P_{loss,2}^{input}$ on the input rate λ for the $MMAP$s with different coefficients of correlation.

Comparing the Figures 1 and 2, we see that the behavior of the curves is similar, but with the same λ, the probability $P^{input}_{loss,2}$ is greater than the probability $P_{loss,1}$. This is because priority customers are blocked only when all servers are occupied with priority customers, while non-priority customers are blocked if all servers are busy with any customers, priority and non-priority.

Figure 3 shows the dependence of the probability that an arriving priority customer will push out a non-priority customer from the server, $P^{serv}_{loss,2}$, on the input rate λ and coefficients of correlation c_1 and c_2.

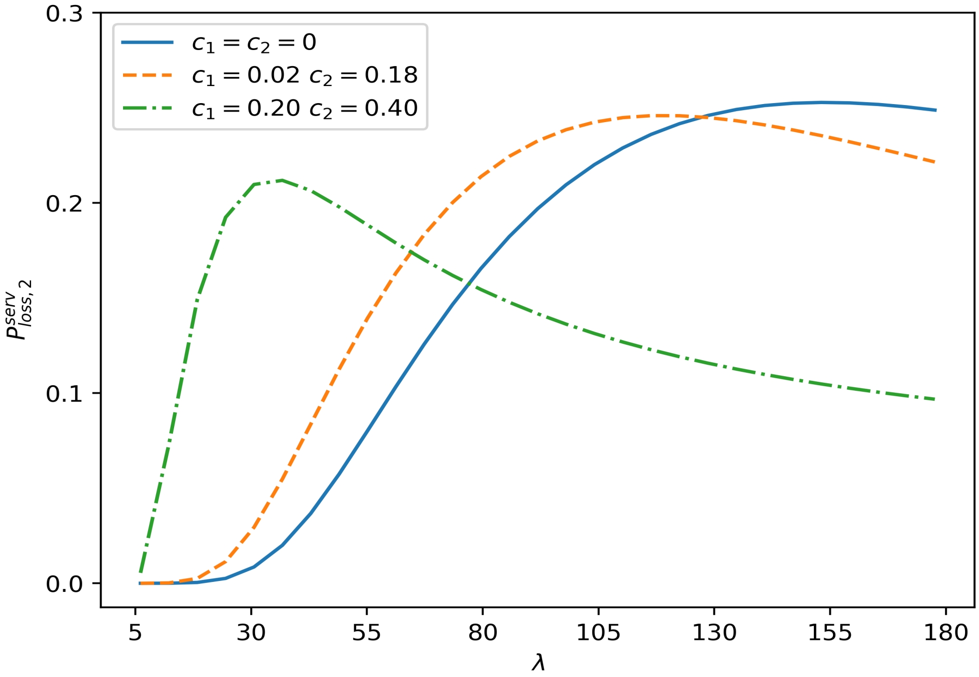

Figure 3. The dependence of the probability $P^{serv}_{loss,2}$ on the input rate λ for the $MMAP$s with different coefficients of correlation.

It is seen from the figure that the curves for different $MMAP$s first increase, and then decrease. The reason for this behavior of the curves lies in the following. When the input rate is relatively small, an arriving priority customer often finds at least one free server and it does not have a need to push out the non-priority customer from the service. As the input rate increases, the system becomes more crowded, and priority customers are forced to push out a non-priority customer to get service. Therefore, the probability $P^{serv}_{loss,2}$ increases and reaches maximum at some point λ_{max}. The further decreasing of this probability is explained by the fact that with λ increasing the most servers become busy with priority customers. Then the number of non-priority customers in the service decreases (they are mostly lost upon arrival to the system) and the probability that a priority customer removes a non-priority customer from the service decreases.

Experiment 2. In this experiment, we investigate the mean number of servers providing service to the priority and non-priority customers as functions of the input rate λ and coefficients of correlation c_1 and c_2.

Figure 4 shows the dependence of the mean number of servers providing service to priority customers, $N_{busy,1}$, on the input rate λ and coefficients of correlation c_1 and c_2.

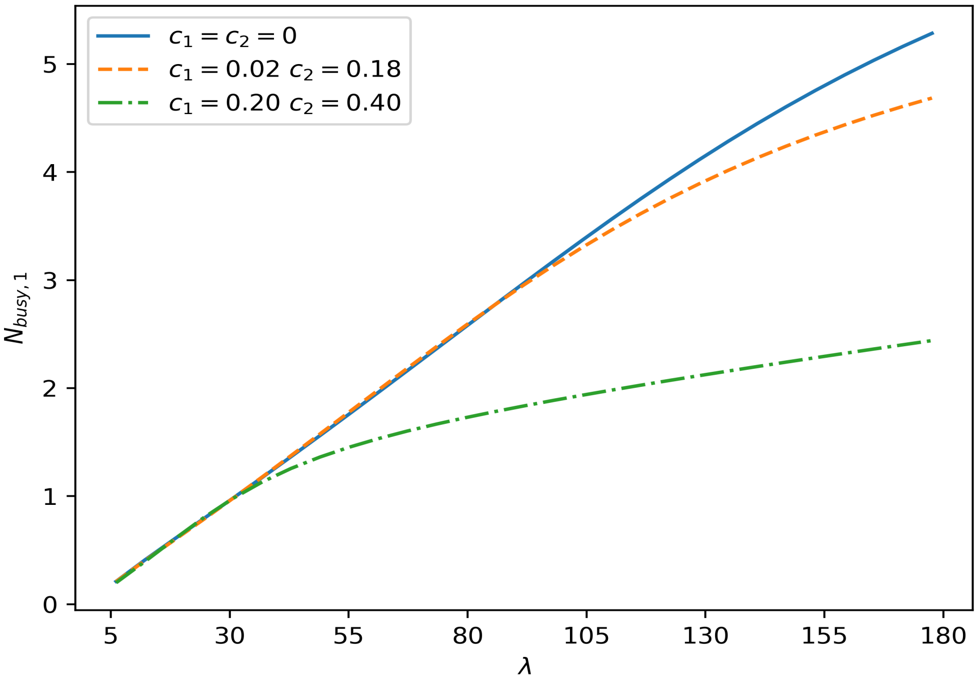

Figure 4. Dependence of the mean number $N_{busy,1}$ of servers providing service to priority customers on the input rate λ for the $MMAP$s with different coefficients of correlation.

It is expectable, that the value of $N_{busy,1}$ increases with λ increasing. More interesting is that, for the same values of λ, the mean number of servers servicing priority customers decreases with increasing correlation in the input flow. Such a relation between the values of $N_{busy,1}$ for $MMAP$s with different correlations is explained by the mentioned above fact that, with less correlation, customers arrive more uniformly, which guarantees more uniform occupation of servers and an increase in the mean number of occupied servers.

Figure 5 depicts the mean number of servers providing service to non-priority customers, $N_{busy,2}$, as a function of the input rate λ and coefficients of correlation c_1 and c_2.

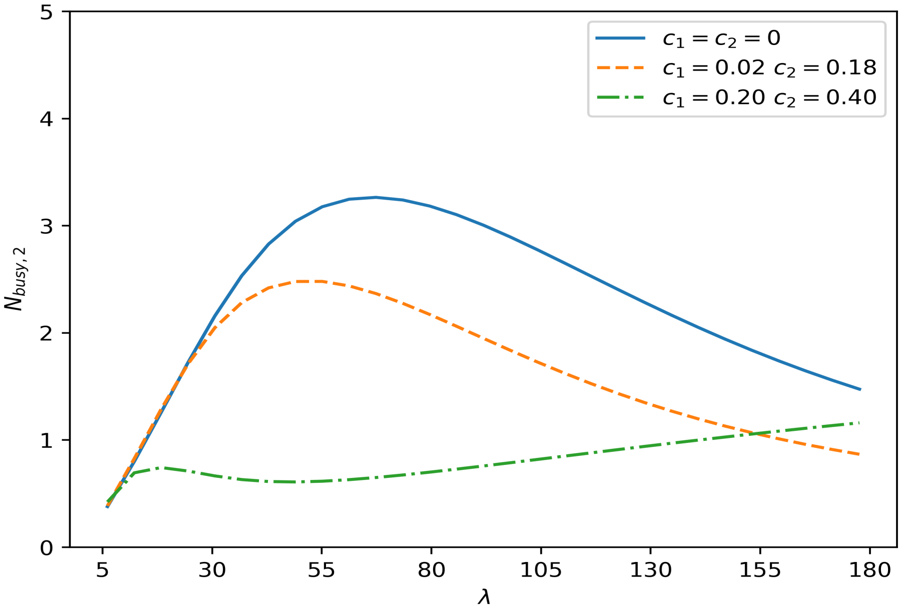

Figure 5. Dependence of the mean number $N_{busy,2}$ of servers providing service to non-priority customers on the input rate λ for the $MMAP$s with different coefficients of correlation.

It is seen from the figure, that the curves for different *MMAP*s first increase, and then decrease. Such a behavior of the curves can be explained by the mechanism of occupation of servers by non-priority customers: when the input rate is relatively small, an arriving non-priority customer often finds at least one free server and occupies it. This customer has a good chance of not being pushed out of service by a priority customer. Therefore, in a certain area, the mean number of servers servicing non-priority customers increases with λ increasing. When the input rate increases further, the system becomes more loaded, and priority customers push out non-priority customers from the servers. Thus, most servers become busy with priority customers and the mean number of servers servicing non-priority customer decreases.

6. Output Flow

Quite often, service to customers is provided not by one set of servers but by the series of such sets. This implies the necessity to consider not a separate queueing system, but a tandem or a network of queues. In such a case, it is very important to investigate the output flow from each system. The output flow from the considered queueing system is characterized as follows.

The output flow from the system under consideration is a *MMAP*. The underlying process of this *MMAP* is the Markov chain ξ_t which describes the operation of the system, i.e.,

$$\xi_t = \{n_t, r_t, \nu_t, m_t^{(1,1)}, m_t^{(2,1)}, \dots, m_t^{(r_t,1)}, m_t^{(1,2)}, m_t^{(2,2)}, \dots, m_t^{(n_t-r_t,2)}\},\ t \geq 0.$$

Let us enumerate the states of the chain ξ_t in the lexicographic order. Then the output flow is defined by the following theorem.

Theorem 2. *The output flow from the system under study is a MMAP that is defined by the matrix $D_0^{(output)}$, $D_1^{(output)}$, $D_2^{(output)}$, which are calculated by the formulas*

$$D_0^{(output)} = \begin{pmatrix} Q_{0,0} & Q_{0,1} & O & \dots & O & O \\ O & Q_{1,1} & Q_{1,2} & \dots & O & O \\ O & O & Q_{2,2} & \dots & O & O \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ O & O & O & \dots & Q_{N-1,N-1} & Q_{N-1,N} \\ O & O & O & \dots & O & Q_{N,N} \end{pmatrix},$$

where the matrices $Q_{n,n}, n = \overline{0,N}$, $Q_{n,n+1}, n = \overline{0,N-1}$, are defined in Theorem 1,

$$D_1^{(output)} = diag^{-}\{\mathcal{S}_{0,1}^{(n)},\ n = \overline{1,N}\},$$

where

$$\mathcal{S}_{0,1}^{(n)} = \begin{pmatrix} O & O & \dots & O \\ I_{\bar{W}} \otimes \mathbf{S}_0^{(1)} \otimes I_{M_2^{n-1}} & O & \dots & O \\ O & I_{\bar{W}} \otimes (\mathbf{S}_0^{(1)})^{\oplus 2} \otimes I_{M_2^{n-2}} & \dots & O \\ \vdots & \vdots & \ddots & \vdots \\ O & O & \dots & I_{\bar{W}} \otimes (\mathbf{S}_0^{(1)})^{\oplus n} \end{pmatrix},$$

$$D_2^{(output)} = diag^{-}\{\mathcal{S}_{0,2}^{(n)},\ n = \overline{1,N}\},$$

where

$$\mathcal{S}_{0,2}^{(n)} =$$

$$\begin{pmatrix} I_{\bar{W}} \otimes (\boldsymbol{S}_0^{(2)})^{\oplus n} & O & O & \dots & O \\ O & I_{\bar{W}M_1} \otimes (\boldsymbol{S}_0^{(2)})^{\oplus n-1} & O & \dots & O \\ O & O & I_{\bar{W}M_1^2} \otimes (\boldsymbol{S}_0^{(2)})^{\oplus n-2} & \dots & O \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ O & O & O & \dots & I_{\bar{W}M_1^{n-1}} \otimes \boldsymbol{S}_0^{(2)} \\ O & O & O & \dots & O \end{pmatrix}.$$

Proof. The underlying process ξ_t of the output $MMAP$ makes idle transitions when:

(i) The underlying process ν_t of the input $MMAP$ makes transitions without a customer arrival or the phase of service processes $m_t^{(k)}, k = 1,2$, changes in the set of transient states. If the number of busy servers is n, the rates of corresponding transitions are described by the block $Q_{n,n}$ of the matrix $D_0^{(output)}$.

(ii) The process ν_t makes a transition accompanied by arrival a customer of type 1 or 2. If the number of busy servers is n, the rates of corresponding transition are described by the block $Q_{n,n+1}$ of the matrix $D_0^{(output)}$.

The underlying process ξ_t of the output $MMAP$ makes transitions, which are accompanied by generation of type k customer, when one of servers servicing type k customer finishes the service, $k = 1,2$. If the number of servers servicing type k customer is n, the rates of corresponding transitions are described by the block $\mathcal{S}_{0,k}^{(n)}$ of the matrix $D_k^{(output)}$, $k = 1,2$. □

Corollary 1. *The output rate of type-k customers is computed by*

$$\mu_k = \mathbf{p} D_k^{(output)} \mathbf{e},\ k = 1,2,$$

where $\mathbf{p}$ *is the solution of the Chapman-Kolmogorov equations derived above.*

The loss probability of type-k customers is computed by

$$P_{loss,k} = 1 - \frac{\mu_k}{\lambda_k},\ k = 1,2.$$

The loss probability of type-2 customer loss due pushing out of service is computed by

$$P_{loss,2}^{push} = P_{loss,2} - P_{loss,2}^{input}.$$

7. Conclusions

In this paper, we investigated the multi-server priority queueing system with correlated flow of two types of customers. Such kind of systems practically is not investigated in the existing literature. The type of a customer defines its priority and distribution of the required service time. The system is analysed under quite general assumptions about the arrival and service processes. We calculated the stationary distribution of system states and the main performance measures including the probability of losses due to the lack of free servers at an arrival moment and due to forcing out of service a non-priority customer with priority ones. We conducted numerical experiments that showed the influence of the mean arrival rate and the effect of correlation in the input flow on the system performance measures. It is evidently seen that the use of the superposition of the stationary Poisson processes as a model of arrival flow leads to the significantly redundant optimism in prediction of the values of the main performance indicators of the system. This is unacceptable in real life applications and justifies the necessity of the provided mathematical analysis of the system.

The used methodology of analysis (primarily the methodology for constructing the structured generator of multi-dimensional Markov chain) looks to be suitable for extension to the systems with more than two priority classes. The results seem to be also extendable to the systems with finite or infinite buffers. The results can be used to model modern telecommunication networks where the flows of information may be essentially heterogeneous with respect to the required bandwidth, importance for the system and tolerance to the losses and (or) delay and jitter.

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

Funding: This work has been supported by grant No 19-29-06043 of the Russian Foundation for Fundamental Research.

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

References

1. He, Q.M.; Xie, J.; Zhao, X. Priority queue with customer upgrades. *Nav. Res. Logist.* **2012**, *59*, 362–375. [CrossRef]
2. McWherter, D.; Schroeder, B.; Ailamaki, N.; Harchol-Balter, M. Priority mechanisms for OLTP and transactional web applications. In Proceedings of the 20th International Conference on Data Engineering (ICDE 2004), Boston, MA, USA, 2 April 2004; pp. 535–546.
3. Brandwajn, A.; Begin, T. Multi-server preemptive priority queue with general arrivals and service times *Perform. Eval.* **2017**, *115*, 150–164.
4. Gans, N.; Koole, G.; Mandelbaum, A. Telephone call centers: A tutorial and literature review. *Manuf. Serv. Oper. Manag.* **2002**, *5*, 79–141. [CrossRef]
5. De Lange, R.; Samoilovich, I.; van der Rhee, B. Virtual queuing at airport security lanes. *Eur. J. Oper. Res.* **2013**, *225*, 153–165. [CrossRef]
6. Lin, D.; Patrick, J.; Labeau, F. Estimating the waiting time of multi-priority emergency patients with downstream blocking. *Healthc. Manag. Sci.* **2014**, *17*, 88–99. [CrossRef] [PubMed]
7. Ellens, W.; Akkerboom, J.; Litjens, R.; van den Berg, H. Performance of cloud computing centers with multiple priority classes. In Proceedings of the 2012 IEEE Fifth International Conference on Cloud Computing, Honolulu, HI, USA, 24–29 June 2012; pp. 245–252.
8. Stallings, W. *Operating Systems Internals and Design Principles*, 4th ed.; Prentice Hall: Englewood Cliffs, NJ, USA, 2004.
9. Gnedenko, B.V.; Danielyan, E.A.; Dimitrov, B.N.; Klimov, G.P.; Matvejev, V.F. *Priority Queueing Systems*; Moscow State University: Moscow, Russian, 1973. (In Russian)
10. Matveev, V.F.; Ushakov, V.G. *Queueing Systems*; Moscow State University: Moscow, Russian, 1984. (In Russian)
11. Miller, R. Priority queues. *Ann. Math. Stat.* **1960**, *31*, 86–1032. [CrossRef]
12. Kleinrock, L. *Queueing Systems Volume II: Computer Applications*; John Wiley and Sons: New York, NY, USA, 1976.
13. Takagi, H. *Queueing analysis: A Foundation of Performance Evaluation, Volume 1: Vacation and Priority Systems, Part 1*; North-Holland: Amsterdam, The Netherlands, 1991.
14. Lucantoni, D. New results on the single server queue with a batch Markovian arrival process *Commun. Stat. Stoch. Model.* **1991**, *7*, 1–46.
15. Chakravarthy, S.R. The batch Markovian arrival process: A review and future work. In *Advances in Probability Theory and Stochastic Processes*; AIP Conference Proceedings: Melville, NY, USA, 2001; pp. 21–49.
16. Dudin, A.N.; Klimenok, V.I.; Vishnevsky, V.M. *The Theory of Queuing Systems with Correlated Flows*; Springer: Berlin/Heidelberg, Germany, 2019.
17. Bocharov, P.P.; D'Apice, C.; Pechinkin, A.V.; Salerno, S. *Queueing Theory*; Utrecht-Boston, VSP: Boston, MA, USA, 2004.
18. Choi, B.D.; Hwang, G.U. The $MAP, M/G_1, G_2/1$ queue with preemptive priority *J. Appl. Math. Stoch. Anal.* **1997**, *10*, 407–421. [CrossRef]
19. Machihara, F. A bridge between preemptive and nonpreemptive queueing models. *Perform. Eval.* **1995**, *23*, 93–106. [CrossRef]

20. Takine, T.; Sengupta, B. A single server queue with service interruptions. *Queueing Syst.* **1997**, *26*, 285–300. [CrossRef]
21. Choi, B.D.; Shin, B.C.; Choi, K.B.; Han, D.H., Jang, J.S. Priority queue with two state Markov modulated arrivals. In Proceedings of the ICC/SUPERCOMM '96-International Conference on Communications, Dallas, TX, USA, 23–27 June 1996; pp. 1055–1059.
22. Krishnamoorthy, A.; Divya, V. $(M, MAP)/(PH, PH)/1$ queue with Nonpremptive Priority, Working Interruption and Protection. *Reliab. Theory Appl.* **2018** *13*, 14–34.
23. Sun, B.; Lee, M.H.; Dudin, A.N.; Dudin, S.A. $MAP + MAP/M_2/N/\infty$ queueing system with absolute priority and reservation of servers. *Math. Probl. Eng.* **2014**, *2014*, 813150. [CrossRef]
24. Sun, B.; Lee, M.H.; Dudin, S.A.; Dudin, A.N. Analysis of multiserver queueing system with opportunistic occupation and reservation of servers. *Math. Probl. Eng.* **2014**, *2014*, 178108. [CrossRef]
25. 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]
26. Dudin, S.; Dudina, O.; Samouylov, K.; Dudin, A. Improvement of fairness of non-preemptive priorities in transmission of heterogeneous traffic. *Mathematics* **2020**, *8*, 929. [CrossRef]
27. He, Q.M. Queues with marked calls. *Adv. Appl. Probab.* **1996**, *28*, 567–587. [CrossRef]
28. Horvath, G. Efficient analysis of the queue length moments of the $MMAP/MAP/1$ preemptive priority queue. *Perform. Eval.* **2012**, *69*, 684–700. [CrossRef]
29. Neuts, M. *Matrix-Geometric Solutions in Stochastic Models*; The Johns Hopkins University Press: Baltimore, MD, USA, 1981.
30. Graham, A. *Kronecker Products and Matrix Calculus with Applications*; Ellis Horwood: Chichester, UK, 1981.
31. Klimenok, V.I.; Kim, C.S.; Orlovsky, D.S.; Dudin, A.N. Lack of invariant property of Erlang $BMAP/PH/N/0$ model. *Queueing Syst.* **2005**, *49*, 187–213. [CrossRef]