Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2014, Article ID 813150, 15 pages
http://dx.doi.org/10.1155/2014/813150

Research Article

$MAP + MAP/M_2/N/\infty$ Queueing System with Absolute Priority and Reservation of Servers

Bin Sun,[1,2] Moon Ho Lee,[3] Alexander N. Dudin,[4] and Sergey A. Dudin[4]

[1] *School of Economics and Management, Inner Mongolia University of Science and Technology, 014010 Baotou, China*
[2] *Inner Mongolia Industry Informatization and Innovation Research Center, Inner Mongolia University of Science and Technology, 014010 Baotou, China*
[3] *Institute of Information and Communication, Chonbuk National University, Jeonju 561-765, Republic of Korea*
[4] *Department of Applied Mathematics and Computer Science, Belarusian State University, 4 Nezavisimosti Avenue, 220030 Minsk, Belarus*

Correspondence should be addressed to Moon Ho Lee; moonho@jbnu.ac.kr

Received 6 May 2014; Revised 6 August 2014; Accepted 7 August 2014; Published 30 November 2014

Academic Editor: Joao B. R. Do Val

We consider a multiserver queueing system with an infinite buffer and two types of customers. The flow of customers is described by two Markovian arrival processes (*MAP*s). Type 1 customers have absolute priority over type 2 customers. If the arriving type 1 customer encounters all servers busy, but some of them provide service to type 2 customers, service of one type 2 customer is terminated and type 1 customer occupies the released server. To avoid too frequent termination of service of type 2 customers, we suggest reservation of some number of servers for type 1 customers. Type 2 customers, who do not succeed to get a server upon arrival or are knocked out from a server, join the buffer or leave the system forever. During a waiting period in the buffer, type 2 customers can be impatient and may leave the system forever. The ergodicity condition of the system is derived in an analytically tractable form. The stationary distribution of the system states and the main performance measures are calculated. The Laplace-Stieltjes transform of the waiting time distribution of an arbitrary type 2 customer is derived. Numerical examples are presented. The problem of the optimal channel reservation is numerically solved.

1. Introduction

Queueing theory is the well recognized mathematical tool for solving the problems of design, capacity planning, performance evaluation, and optimization of many real life objects, especially in telecommunications, manufacturing, computer engineering, and so forth. The relevant literature is huge. The most part of the literature is devoted to queueing systems with homogeneous customers while it is a quite frequent real life situation when arriving customers have different importance to the system and various requirements to the quality of their service. So, an important part of queueing literature is devoted to so called priority queues. In such queues, customers of different types are arranged into several classes enumerated, for example, in descending order of their value (economical, social, etc.) for the system and customers from different classes have different treatment in the system. Customers from the classes having higher priority have preferences to others in access to the servers, if they are available or picking-up from the queue to service.

The existing literature concerning the priority queues is also huge. So, to essentially reduce the list of the relevant references, here we will cite only papers devoted to priority queues with a Markovian arrival process (*MAP*); see, for example, [1, 2]. Such an arrival process is the significant generalization of the stationary Poisson process and is very popular descriptor of the flows of customers in modern real life systems now. In contrast to the stationary Poisson process, which allows fitting of only the average value of

interarrival times, the *MAP* allows fitting of also the variance and high order moments of the distribution of interarrival times and correlation of successive interarrival times. The single-server priority queues with the *MAP*s or a bit more general marked Markovian arrival process (*MMAP*) (see, e.g., [3]) were considered, for examples, in [4–8].

There are different kinds of priorities provided to important customers. The most well known are preemptive and nonpreemptive priorities. Nonpreemptive priority suggests that an arriving high priority customer cannot interrupt service currently provided to a low priority customer. Nonpreemptive priority plays a role only when a server finishes service and the next customers should start processing at this server. Preemptive priority suggests that an arriving high priority customer interrupts service currently provided to a low priority customer. This low priority customer leaves the server. It can leave the system without service permanently or try to get service later when a server will become available. Its service time may be started from the point of the interruption, from the last check point before the interruption, or from the early beginning with the same or different distribution of the service time.

The most popular variant of priority queues assumes existence of only two types of customers. Note that the important special cases of priority queues with a preemptive priority are queues with server breakdowns or interruptions. The breakdowns and interruptions can be interpreted as high level customers whose arrival causes termination of service of a usual (low level) customer. Short review of the recent papers related to this subject can be found in [9] where a multiserver queue with quite involved mechanisms of servers breakdowns and repair is analyzed with emphasis to evaluation of survivability of the system. In [10–12], multiserver queues with nonpreemptive priorities are investigated. Model in [12] assumes self-generation of priorities. In [10], multiserver queue with nonpreemptive priority and impatient customers is investigated. Model in [11] assumes that a decision about admission of a nonpriority customer to the first station of a tandem queue is based on information about the current number of customers at the second station of a tandem.

In this paper, we consider a multiserver queue with a preemptive priority. The most close papers in literature are [13, 14]. It is supposed in [13] that there is a finite buffer for the nonpriority customers whose service was interrupted. In [14], it is assumed that the nonpriority customers whose service was interrupted go to orbit and retry for the service later on as the priority customers.

It is evident that the preemptive priority is much better comparing to the nonpreemptive priority from the point of view of high priority customers. However, the preemptive priority is much worse comparing to the nonpreemptive one from the point of view of low priority customers and from the point of view of the system resources utilization. Some work already done by a server is wasted when the service interruption occurs. As a trade-off, the discipline with the nonpreemptive priority can be used in combination with reservation of some servers exclusively for the service of priority customers. Queueing systems with nonpreemptive priority and reservation of some servers were considered, for example, in [15–17]. In [15], such a system was used for analysis and optimization of the work of the cell of a mobile communication network. The models considered in [16, 17] are the dual tandem queues.

The obvious aim of the proposed combination of the reservation with the non-preemptive priority in [15–17] was the desire to give more advantage to high priority customers comparing to an usual nonpreemptive priority discipline. In our present paper, we propose the combination of a reservation with a *preemptive* priority. Because the preemptive priority itself gives too much advantage to high priority customers, they do not need an additional reservation of the servers. So, motivation of the discipline considered in this paper is to improve quality of low priority customers by means of decreasing the frequency of service interruptions. This is important because interruptions may be quite offensive for low priority customers and also lead to the waste of the system resources.

Related model was analyzed in our recent paper [18]. In that paper, it is assumed that the arriving nonpriority customers are not registered by the system manager and the nonpriority customers who cannot enter the service immediately upon arrival or whose service was interrupted go to some virtual place called as an orbit and retry for the service later on. In this paper, we consider another type of arrival process and also assume that there exists registration of nonpriority customers and they are placed in an infinite buffer if they cannot enter the service immediately upon arrival or are interrupted. This assumption allows us to provide a detailed analysis of waiting time distribution for nonpriority customers while such an analysis is impossible for the system with retrials.

The results of analysis presented in our paper can be used for enhancement of operation of many real life systems. Let us briefly mention three of them.

(i) Cognitive radio systems; see, for example, [19–24]: in these systems, the high priority is assigned to the primary, licensed customers and the low priority is assigned to the secondary, unlicensed customers. The secondary customers may occupy the free servers (channels or subchannels), but the service of some secondary customers is terminated if a primary customer arrives and does not see a free server.

(ii) Contact and call centers; see, for example, [25–34]: in these systems, the high priority is assigned to the more important customers or voice requests and the low priority is assigned to the less important customers, e-mail requests, and work relating to providing a call-back option.

(iii) Different technical, manufacturing, service systems where, to avoid possible starvation and increase a profit, the servers provide the service to some background or external customers in absence of the primary customers.

The rest of the paper is organized as follows. In Section 2, the mathematical model is described. Multidimensional Markov chain defining behavior of the system is introduced

in Section 3. The infinitesimal generator of this Markov chain is written down there. The ergodicity condition and the stationary distribution of the system states are analyzed in Section 4. The expressions for the main system performance measures are given in Section 5. The Laplace-Stieltjes transform of the waiting time distribution of an arbitrary type 2 customer is derived in Section 6. Section 7 contains some numerical illustrations. Finally, Section 8 concludes the paper.

2. Mathematical Model

We consider an N-server queueing model with an infinite buffer and two types of customers. Type 1 customers arrive according to the Markovian arrival flow MAP_1. The MAP_1 is defined by the underlying process $\nu_t, t \geq 0$, which is an irreducible continuous-time Markov chain with the state space $\{0, 1, \ldots, W\}$. Arrivals occur only at the epochs of jumps in the underlying process $\nu_t, t \geq 0$. The intensities of transitions of the process $\nu_t, t \geq 0$, which are accompanied (not accompanied) by the arrival of a customer, are defined by the square matrix $D_1(D_0)$ of size $\overline{W} = W + 1$. The matrix $D = D_0 + D_1$ is an infinitesimal generator of the process $\nu_t, t \geq 0$. The stationary distribution vector $\boldsymbol{\chi}$ of this process satisfies the system of equations $\boldsymbol{\chi} D = \mathbf{0}$, $\boldsymbol{\chi}\mathbf{e} = 1$. Here and throughout this paper, $\mathbf{0}$ is a zero row vector, and $\mathbf{e}$ denotes a unit column vector.

Such arrival process was introduced as a versatile Markovian point process (*VMPP*) by M. F. Neuts in the 70s. The original development of *VMPP* contained extensive notations; however, these notations were simplified greatly in [2] and ever since this process bears the name Markovian arrival process. The class of *MAP*s includes many input flows considered previously, such as stationary Poisson (M), Erlangian (E_k), Hyper-Markovian (HM), Phase-Type (PH), and Markov Modulated Poisson Process (*MMPP*). Generally speaking, the *MAP* is correlated, so it is ideal to model correlated and/or bursty traffic in modern telecommunication networks. For more information about the *MAP*, its properties, and special cases, see [1, 2]. Discussion of applicability of the *MAP* for description of real life information flows in modern telecommunication networks can be found in [35, 36]. Possibilities of fitting the real world arrival process by means of the *MAP* are discussed, for example, in [37].

The average intensity λ_1 (fundamental rate) of the MAP_1 is defined by $\lambda_1 = \boldsymbol{\chi} D_1 \mathbf{e}$. The coefficient of variation c_{var} of intervals between customer arrivals is calculated as $c_{\text{var}} = 2\lambda_1 \boldsymbol{\chi}(-D_0)^{-1}\mathbf{e} - 1$, and the coefficient of correlation c_{cor} of successive intervals between arrivals is given as $c_{\text{cor}} = (\lambda_1 \boldsymbol{\chi}(-D_0)^{-1} D_1 (-D_0)^{-1}\mathbf{e} - 1)/c_{\text{var}}^2$.

Type 2 customers arrive to the system according to the MAP_2. The MAP_2 is defined by the underlying process $\zeta_t, t \geq 0$, with the finite state space $\{0, 1, \ldots, Z\}$, and described by the square matrices H_0 and H_1 of size $\overline{Z} = Z + 1$.

The average intensity λ_2 of arrival of type 2 customers is given by $\lambda_2 = \boldsymbol{\theta} H_1 \mathbf{e}$; there the vector $\boldsymbol{\theta}$ is the stationary distribution vector of the process $\zeta_t, t \geq 0$ and is defined as the unique solution to the system $\boldsymbol{\theta}(H_0 + H_1) = \mathbf{0}, \boldsymbol{\theta}\mathbf{e} = 1$.

We assume that type 1 customers have absolute priority over type 2 customers. If there is a free server during a type 1 customer arrival epoch, this customer receives service immediately. If all servers are busy during a type 1 customer arrival epoch and there are type 2 customers receiving service, the service process of one type 2 customer is terminated and type 1 customer occupies the released server. The type 2 customer who was knocked out from the system goes to the buffer in the tail of the queue with probability p and leaves the system with the complimentary probability $1 - p$. If all servers are occupied by type 1 customers during a type 1 customer arrival epoch, this customer leaves the system forever.

We assume that some parameter (threshold) M is fixed, $0 < M \leq N$. Type 2 customers are admitted to the system if the number of busy servers is less than M during type 2 customer arrival epoch. If the number of busy servers is greater than $M-1$ during an arbitrary type 2 customer arrival epoch, this customer goes to the buffer with probability q and with the complimentary probability leaves (balks) the system.

Additionally, we assume that type 2 customers can be impatient and leave the buffer after an exponentially distributed time described by parameter α, $\alpha > 0$, due to lack of service. If we assume that type 2 customers are patient, we set $\alpha = 0$.

The service time of type r customers has an exponential distribution with the parameter $\mu_r, r = 1, 2$.

3. Process of System States

Let $i_t, i_t \geq 0$, be the number of type 2 customers in the buffer, let $n_t, n_t = \overline{0, N}$, be the number of busy servers, let $l_t, l_t = \overline{0, \min\{n_t, M\}}$, be the number of type 2 customers in service, let $\nu_t, \nu_t = \overline{0, W}$, be the state of the directing process of the MAP_1, and let $\zeta_t, \zeta_t = \overline{0, Z}$, be the state of the directing process of the MAP_2 at the epoch $t, t \geq 0$. Here the notation like $k = \overline{0, K}$ means that the variable k takes integer values from the set $\{0, 1, 2, \ldots, K\}$.

The behavior of the system under study can be described in terms of the regular irreducible continuous-time Markov chain $\xi_t = \{i_t, n_t, l_t, \nu_t, \zeta_t\}, t \geq 0$.

Let us introduce the following notation:

(i) I is the identity matrix, and O is a zero matrix of appropriate dimension. If the dimension of a matrix or a vector is not clear from context, it is indicated as the suffix;

(ii) $\oplus$ and $\otimes$ indicate the Kronecker sum and product of matrices, respectively; see, for example, [38];

(iii) $\text{diag}\{A_1, \ldots, A_l\}$ is the diagonal matrix with the diagonal entries or blocks $A_1, \ldots, A_l$;

(iv) $\underline{C}_l = \text{diag}\{0, 1, \ldots, l\}$, $\overline{C}_l = \text{diag}\{l, l-1, \ldots, 0\}, l = \overline{0, M}$;

(v) $\widetilde{C}_l = \text{diag}\{l, l-1, \ldots, l-M+1, l-M\}$, $l = \overline{M, N}$;

(vi) $E_n^+, \widehat{E}_n^+, n = \overline{0, M-1}$, are the matrices of size $(n+1) \times (n+2)$ with all zero entries except the entries $(E_n^+)_{l,l+1}, l = \overline{0, n+1}$, and $(\widehat{E}_n^+)_{l,l}$, $l = \overline{0, n+1}$, which are equal to 1;

(vii) $E_n^-, \widehat{E}_n^-, n = \overline{1, M}$, are the matrices of size $(n+1) \times n$ with all zero entries except the entries $(E_n^-)_{l,l}$, $l = \overline{0, n}$, and $(\widehat{E}_n^-)_{l,l-1}$, $l = \overline{1, n+1}$, which are equal to 1;

(viii) E^- is the square matrix of size $M+1$ with all zero entries except the entries $(E^-)_{l,l-1}, l = \overline{1, M}$, which are equal to 1;

(ix) $\widehat{I}_l, l = \overline{M+1, N-M+1}$, are the square matrices of size l with all zero entries except the entry $(\widehat{I}_l)_{0,0}$ which is equal to 1.

Let us enumerate the states of the Markov chain ξ_t in the direct lexicographic order of the components (i, n, l, ν, ζ) and refer to the pair (i, n) as a macrostate.

Let Q be the generator of the Markov chain $\xi_t, t \geq 0$, consisting of the blocks $Q_{i,j}$, which, in turn, consist of the matrices $(Q_{i,j})_{n,n'}$ of the transition rates of the Markov chain ξ_t from the macrostate (i, n) to the macrostate $(j, n'), n, n' = \overline{0, N}$. The diagonal entries of the matrices $Q_{i,i}$ are negative, and the moduli of the diagonal entries of these matrices define the total intensities of leaving the corresponding state of the Markov chain $\xi_t, t \geq 0$.

Lemma 1. *The infinitesimal generator Q of the Markov chain $\xi_t, t \geq 0$, has the block-three-diagonal structure*

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

The nonzero blocks $Q_{i,j}, i, j \geq 0$, have the following form:

$$Q_{0,0} = \begin{pmatrix} A_0^{(0)} & B^{(0)} & O & \dots & O & O \\ F^{(1)} & A_0^{(1)} & B^{(1)} & \dots & O & O \\ \vdots & \vdots & \ddots & \ddots & \vdots & \vdots \\ O & O & O & \ddots & A_0^{(N-1)} & B^{(N-1)} \\ O & O & O & \dots & F^{(N)} & A_0^{(N)} \end{pmatrix} + I_{(M+1)(N-M/2+1)} \otimes \left(D_0 \oplus H_0\right),$$

$$Q_{i,i} = \begin{pmatrix} A_i^{(M)} & B^{(M)} & O & \dots & O & O \\ F^{(M+1)} & A_i^{(M+1)} & B^{(M+1)} & \dots & O & O \\ \vdots & \vdots & \ddots & \ddots & \vdots & \vdots \\ O & O & O & \ddots & A_i^{(N-1)} & B^{(N-1)} \\ O & O & O & \dots & F^{(N)} & A_i^{(N)} \end{pmatrix} + I_{(N-M+1)(M+1)} \otimes \left(D_0 \oplus H_0\right), \quad i \geq 1,$$

$$Q_{0,1} = \begin{pmatrix} O_{(M(M+1)/2)\overline{W}\,\overline{Z} \times (N-M+1)(M+1)\overline{W}\,\overline{Z}} \\ Q^+ \end{pmatrix},$$

$$Q^+ = \operatorname{diag}\left\{ \underbrace{qI_{(M+1)\overline{W}} \otimes H_1, \dots, qI_{(M+1)\overline{W}} \otimes H_1}_{N-M}, \; pE^- \otimes D_1 \otimes I_{\overline{Z}} + qI_{(M+1)\overline{W}} \otimes H_1 \right\},$$

$$Q_{i,i+1} = Q^+, \quad i \geq 1,$$

$$Q_{1,0} = \left(O_{(N-M+1)(M+1)\overline{W}\,\overline{Z} \times (M(M+1)/2)\overline{W}\,\overline{Z}} \;\; Q^-\right),$$

$$Q^- = \operatorname{diag}\left\{\alpha I_{M+1} + \mu_1 \overline{C}_M E^+ + \mu_2 C_M, \alpha I_{M+1}, \dots, \alpha I_{M+1}\right\} \otimes I_{\overline{W}\,\overline{Z}},$$

$$Q_{i,i-1} = \operatorname{diag}\left\{i\alpha I_{M+1} + \mu_1 \overline{C}_M E^+ + \mu_2 C_M, i\alpha I_{M+1}, \dots, i\alpha I_{M+1}\right\} \otimes I_{\overline{W}\,\overline{Z}}, \quad i > 1, \tag{2}$$

where

$$A_i^{(n)} = \begin{cases} -\left(\mu_1 \overline{C}_n + \mu_2 C_M\right) \otimes I_{\overline{W}\,\overline{Z}}, & 0 \leq n < M, \\ -\left(\mu_1 \overline{C}_n + \mu_2 C_M + i\alpha I_{M+1}\right) \otimes I_{\overline{W}\,\overline{Z}} \\ \quad + (1-q)\, I_{(M+1)\overline{W}} \otimes H_1, & M \leq n < N, \\ -\left(\mu_1 \overline{C}_n + \mu_2 C_M + i\alpha I_{M+1}\right) \otimes I_{\overline{W}\,\overline{Z}} \\ + \left(\widehat{I}_{M+1} + (1-p)\, E^-\right) \otimes D_1 \otimes I_{\overline{Z}} \\ \quad + (1-q)\, I_{(M+1)\overline{W}} \otimes H_1, & n = N, i \geq 0; \end{cases}$$

$$B^{(n)} = \begin{cases} E_n^+ \otimes I_{\overline{W}} \otimes H_1 \\ \quad + \widehat{E}_n^+ \otimes D_1 \otimes I_{\overline{Z}}, & 0 \leq n < M, \\ I_{M+1} \otimes D_1 \otimes I_{\overline{Z}}, & M \leq n < N; \end{cases}$$

$$F^{(n)} = \begin{cases} \left(\mu_1 \overline{C}_n E_n^- + \mu_2 C_n \widehat{E}_n^-\right) \otimes I_{\overline{W}\,\overline{Z}}, & 0 < n \leq M, \\ \left(\mu_1 \overline{C}_n + \mu_2 C_M E^-\right) \otimes I_{\overline{W}\,\overline{Z}}, & M < n \leq N. \end{cases} \tag{3}$$

Proof of the lemma is implemented by the analysis of the Markov chain $\xi_t, t \geq 0$, transitions during the interval of infinitesimal length, and further combining the corresponding transition intensities in block matrix form.

4. Ergodicity Condition and Computation of the Stationary Probabilities

Further, we will separately consider the following two cases.

(1) Let us assume that $\alpha \neq 0$; that is, customers in the buffer are impatient. In this case, it is possible to verify that the following limits exist:

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

$$Y_1 = \lim_{i \to \infty} R_i^{-1} Q_{i,i} + I, \tag{4}$$

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

where R_i is a diagonal matrix with diagonal entries defined as the moduli of the corresponding diagonal entries of the matrix $Q_{i,i}$, $i \geq 0$.

The matrix R_i is the block-diagonal matrix with the diagonal blocks $T_i^{(n)}$, $n = \overline{M, N}$, $i \geq 0$, defined as follows:

$$T_i^{(n)} = \begin{cases} \left(\mu_1 \widetilde{C}_n + \mu_2 C_M + i\alpha I_{M+1}\right) \otimes I_{\overline{W}\,\overline{Z}} \\ \quad + I_{M+1} \otimes \left(\Lambda_0 \oplus \Sigma_0\right) \\ \quad - \left(1-q\right) I_{(M+1)\overline{W}} \otimes \Sigma_1, & n = \overline{M, N-1}; \\ \left(\mu_1 \widetilde{C}_N + \mu_2 C_M + i\alpha I_{M+1}\right) \otimes I_{\overline{W}} \\ \quad - \widehat{I}_{M+1} \otimes \Sigma_1 \otimes I_{\overline{Z}} \\ \quad + I_{M+1} \otimes \left(\Lambda_0 \oplus \Sigma_0\right) \\ \quad - \left(1-q\right) I_{(M+1)\overline{W}} \otimes \Sigma_1, & n = N, \end{cases} \tag{5}$$

where $\Lambda_0, \Lambda_1, \Sigma_0$, and Σ_1 are diagonal matrices with diagonal entries defined by the diagonal entries of the matrices $-D_0$, $D_1, -H_0$, and H_1, respectively.

The matrices Y_0, Y_1, and Y_2 have the following form:

$$Y_0 = I, \qquad Y_1 = O, \qquad Y_2 = O; \tag{6}$$

so, their sum is the stochastic matrix.

According to the definition given in [39], the Markov chain $\xi_t, t \geq 0$, belongs to the class of so called continuous-time asymptotically quasi-Toeplitz Markov chains (*AQTMC*).

As follows from [39], the sufficient condition for the ergodicity of the *AQTMC* is the following condition:

$$\mathbf{y} Y_0 \mathbf{e} > \mathbf{y} Y_2 \mathbf{e}, \tag{7}$$

where the row-vector $\mathbf{y}$ is the unique solution to the following system of linear algebraic equations:

$$\mathbf{y}\left(Y_0 + Y_1 + Y_2\right) = \mathbf{y}, \qquad \mathbf{y}\mathbf{e} = 1. \tag{8}$$

Taking into account explicit values (6) of the matrices Y_0, Y_1, and Y_2, inequality (7) can be rewritten as $\mathbf{y}Y_0\mathbf{e} > 0$, and system (8) is rewritten in the form $\mathbf{y}Y_0 = \mathbf{y}, \mathbf{y}\mathbf{e} = 1$. Hence, $\mathbf{y}Y_0\mathbf{e} = \mathbf{y}\mathbf{e} = 1 > 0$; so, inequality (7) holds true for each set of the system parameters. So, the Markov chain under consideration is ergodic and the queueing system under study is stable for any set of the system parameters.

(2) Let us assume now that $\alpha = 0$. In this case, the blocks of the generator have the following form:

$$\begin{aligned} Q_{i,i} &= Q_1 \\ &= \begin{pmatrix} A_0^{(M)} & B^{(M)} & O & \dots & O & O \\ F^{(M+1)} & A_0^{(M+1)} & B^{(M+1)} & \dots & O & O \\ \vdots & \vdots & \ddots & \ddots & \vdots & \vdots \\ O & O & O & \ddots & A_0^{(N-1)} & B^{(N-1)} \\ O & O & O & \dots & F^{(N)} & A_0^{(N)} \end{pmatrix} \\ &\quad + I_{(N-M+1)(M+1)} \otimes \left(D_0 \oplus H_0\right), \quad i > 0, \\ Q_{i,i+1} &= Q_2 = Q^+, \quad i \geq 1, \\ Q_{i,i-1} &= Q_0 \\ &= \operatorname{diag}\left\{\mu_1 \overline{C}_M E^+ + \mu_2 C_M, O_{M+1}, \dots, O_{M+1}\right\} \otimes I_{\overline{W}\,\overline{Z}}, \\ &\quad i > 1. \end{aligned} \tag{9}$$

Thus, the blocks of the generator do not depend on the variable i when $i > 1$ and the Markov chain $\xi_t, t \geq 0$, belongs to the class of continuous-time quasi-Toeplitz Markov chains (*QTMC*) or $M/G/1$-type Markov chains; see [40].

As follows from [40], the necessary and sufficient condition for the ergodicity of the *QTMC* is the fulfillment of the following inequality:

$$\mathbf{x} Q_0 \mathbf{e} > \mathbf{x} Q_2 \mathbf{e}, \tag{10}$$

where the vector $\mathbf{x}$ is the unique solution to the system

$$\mathbf{x}\left(Q_0 + Q_1 + Q_2\right) = \mathbf{0}, \qquad \mathbf{x}\mathbf{e} = 1. \tag{11}$$

It is easy to verify that system (11) can be rewritten in the form

$$\begin{aligned} \mathbf{0} &= \mathbf{x}\left(Q_0 + Q_1 + Q_2\right) \\ &= \mathbf{x}\left[\begin{pmatrix} I_{M+1} \otimes D_0 & I_{M+1} \otimes D_1 & \dots & O & O \\ O & I_{M+1} \otimes D_0 & \dots & O & O \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ O & O & \dots & I_{M+1} \otimes D_0 & I_{M+1} \otimes D_1 \\ O & O & \dots & O & \left(E^- + \widehat{I}\right) \otimes D_1 + I_{M+1} \otimes D_0 \end{pmatrix} \otimes I_{\overline{Z}} \right. \\ &\quad + I_{(N-M+1)(M+1)\overline{W}} \otimes H(1) \\ &\quad \left. + \begin{pmatrix} -\mu_1 \widetilde{C}_M \left(I - E^+\right) & O & O & \dots & O & O \\ \mu_1 \widetilde{C}_{M+1} + \mu_2 C_M E^- & -\left(\mu_1 \widetilde{C}_{M+1} + \mu_2 C_M\right) & O & \dots & O & O \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ O & O & O & \dots & -\left(\mu_1 \widetilde{C}_{N-1} + \mu_2 C_M\right) & O \\ O & O & O & \dots & \mu_1 \widetilde{C}_N + \mu_2 C_M E^- & -\left(\mu_1 \widetilde{C}_N + \mu_2 C_M\right) \end{pmatrix} \otimes I_{\overline{W}\,\overline{Z}} \right], \\ \mathbf{x}\mathbf{e} &= 1. \end{aligned} \tag{12}$$

By postmultiplying the left and right hand side of (12) by $\mathbf{e}_{(N-M+1)(M+1)\overline{W}} \otimes I_{\overline{Z}}$ we obtain that

$$\mathbf{x}\left(\mathbf{e}_{(N-M+1)(M+1)\overline{W}} \otimes H(1)\right) = \mathbf{0}, \qquad \mathbf{x}\mathbf{e} = 1. \tag{13}$$

By postmultiplying the left and right hand side of (12) by $\mathbf{e}_{(N-M+1)(M+1)} \otimes I_{\overline{W}} \otimes \mathbf{e}_{\overline{Z}}$ we obtain that

$$\mathbf{x}\left(\mathbf{e}_{(N-M+1)(M+1)} \otimes D(1) \otimes \mathbf{e}_{\overline{Z}}\right) = \mathbf{0}, \qquad \mathbf{x}\mathbf{e} = 1. \tag{14}$$

It follows from (13) and (14) that the vector $\mathbf{x}$ can be represented in the form

$$\mathbf{x} = \mathbf{z} \otimes \boldsymbol{\chi} \otimes \boldsymbol{\theta}, \tag{15}$$

where $\mathbf{z}$ is a stochastic vector of size $(N - M + 1)(M + 1)$, $\boldsymbol{\chi}$ is the invariant probability vector of the underlying Markov chain of the MAP_1, and $\boldsymbol{\theta}$ is the invariant probability vector of the underlying Markov chain of the MAP_2.

Substituting the vector $\mathbf{x}$ form (15) into (12), postmultiplying this equation by $I_{(N-M+1)(M+1)} \otimes \mathbf{e}_{\overline{W}\,\overline{Z}}$, and taking into account that $\boldsymbol{\chi}D_1\mathbf{e} = \lambda_1$, $\boldsymbol{\chi}D_0\mathbf{e} = -\boldsymbol{\chi}D_1\mathbf{e} = -\lambda_1, \boldsymbol{\theta}H(1) = \mathbf{0}, \boldsymbol{\theta}\mathbf{e} = 1$, and $\boldsymbol{\chi}\mathbf{e} = 1$, we obtain that the vector $\mathbf{z}$ is the unique solution to the system

$$\mathbf{z}\Omega = \mathbf{0}, \qquad \mathbf{z}\mathbf{e} = 1, \tag{16}$$

where

$$\Omega = \begin{pmatrix} \widetilde{A}^{(M)} & \lambda_1 I_{M+1} & O & \dots & O & O & O \\ \mu_1 \widetilde{C}_{M+1} + \mu_2 C_M E^- & \widetilde{A}^{(M+1)} & \lambda_1 I_{M+1} & \dots & O & O & O \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ O & O & O & \dots & \mu_1 \widetilde{C}_{N-1} + \mu_2 C_M E^- & \widetilde{A}^{(N-1)} & \lambda_1 I_{M+1} \\ O & O & O & \dots & O & \mu_1 \widetilde{C}_N + \mu_2 C_M E^- & \widetilde{A}^{(N)} \end{pmatrix}, \tag{17}$$

$$\widetilde{A}^{(n)} = \begin{cases} -\mu_1 \widetilde{C}_M (I - E^+) - \lambda_1 I_{M+1}, & n = M, M < N; \\ -\mu_1 \widetilde{C}_M (I - E^+) - \lambda_1 \left(I - E^- - \widehat{I}\right), & n = M = N, \\ -\mu_1 \widetilde{C}_n - \mu_2 C_M - \lambda_1 I_{M+1}, & M < n < N, \\ -\mu_1 \widetilde{C}_N - \mu_2 C_M - \lambda_1 \left(I - E^- - \widehat{I}_{M+1}\right), & n = N > M. \end{cases}$$

It easy to verify that the matrix Ω is the generator of two-dimensional Markov chain $\{n_t, l_t\}, t \geq 0$, defining the number of busy servers $n_t, n_t = \overline{M, N}$, and the number of servers occupied by type 2 customers $l_t, l_t = \overline{0, M}$, at the moment t in the case when the system is overloaded. It is evident that when the system is overloaded the number of busy servers n_t can vary in the interval $[M, N]$. It follows from (16) that the vector $\mathbf{z} = (\mathbf{z}_M, \dots, \mathbf{z}_N), \mathbf{z}_n = (\mathbf{z}(n, 0), \dots, \mathbf{z}(n, M)), n = \overline{M, N}$, defines the joint stationary distribution of the number of busy servers and the number of servers occupied by type 2 customers when the system is overloaded.

Taking this into account, substituting the vector $\mathbf{x}$ form (15) into inequality (10) and performing some algebra, we obtain the following inequality:

$$\mathbf{z}_M \left(\mu_1 \overline{C}_M + \mu_2 C_M\right) \mathbf{e} > q\lambda_2 + p\lambda_1 \mathbf{z}_N \widehat{\mathbf{e}}, \tag{18}$$

where $\widehat{\mathbf{e}}$ is the vector of size $M + 1$ with all unit entries except the entry $(\widehat{\mathbf{e}})_0$ which is equal to zero.

Thus, we have proved the following assertion.

Theorem 2. *If $\alpha \neq 0$, the Markov chain $\xi_t, t \geq 0$, is ergodic for any set of the system parameters. If $\alpha = 0$, the Markov chain $\xi_t, t \geq 0$, is ergodic, if and only if inequality* (18) *holds true where the vector $\mathbf{z}$ is defined as the solution to system* (16).

Remark 3. Condition (18) is intuitively clear and can be interpreted as follows. If the system is overloaded and type 2 customers are patient ($\alpha = 0$), then type 2 customer can leave the buffer only if the number of busy servers becomes less than M. The components $\mathbf{z}(M, l), l = \overline{0, M}$, of the vector $\mathbf{z}_M$ define the probability that during an arbitrary epoch the number of busy servers is M and there are l type 2 customers who receive service. So, the left hand side of (18) defines the intensity of the service completions when M servers are busy (the intensity of customers leaving the buffer). The right hand side of (18) defines the total intensity of type 2 customers arrival into the buffer. When the system is overloaded, all admitted type 2 customers (the intensity of this flow is $q\lambda_2$) and type 2 customers whose service was terminated by arrival of type 1 customers (the intensity of this flow is $p\lambda_1 \mathbf{z}_N \widehat{\mathbf{e}}$) arrive into the buffer. Thus, ergodicity condition (18) requires that, in the situation when the system is overloaded, the intensity of type 2 customers arriving to the buffer is less than the intensity of type 2 customers leaving the buffer.

Further, we assume that Markov chain $\xi_t, t \geq 0$, is ergodic. Then the following limits (stationary probabilities) exist:

$$\pi(i, n, l, \nu, \zeta) = \lim_{t \to \infty} P\{i_t = i, n_t = n, l_t = l, \nu_t = \nu, \zeta_t = \zeta\},$$
$$i \geq 0, \quad n = \overline{(1 - \delta_{i,0})\, M, N},$$
$$l = \overline{0, \min\{n, M\}}, \quad \nu = \overline{0, W}, \quad \zeta = \overline{0, Z}. \tag{19}$$

Here $\delta_{i,0}$ indicates the Kronecker delta.

Let us form the row vectors $\boldsymbol{\pi}_i$ of these probabilities as follows:

$$\begin{aligned}
&\boldsymbol{\pi}(i,n,l,\nu)\\
&\quad=(\pi(i,n,l,\nu,0),\pi(i,n,l,\nu,1),\ldots,\pi(i,n,l,\nu,Z)),\\
&\qquad i\geq 0,\quad n=\overline{(1-\delta_{i,0})M,N},\\
&\qquad l=\overline{0,\min\{n,M\}},\quad \nu=\overline{0,W},\\
&\boldsymbol{\pi}(i,n,l)=(\boldsymbol{\pi}(i,n,l,0),\boldsymbol{\pi}(i,n,l,1),\ldots,\boldsymbol{\pi}(i,n,l,W)),\\
&\qquad i\geq 0,\quad n=\overline{(1-\delta_{i,0})M,N},\\
&\qquad l=\overline{0,\min\{n,M\}},\\
&\boldsymbol{\pi}(i,n)=(\boldsymbol{\pi}(i,n,0),\boldsymbol{\pi}(i,n,1),\ldots,\boldsymbol{\pi}(i,n,\min\{n,M\})),\\
&\qquad i\geq 0,\quad n=\overline{(1-\delta_{i,0})M,N},\\
&\boldsymbol{\pi}_i=(\boldsymbol{\pi}(i,0),\boldsymbol{\pi}(i,1),\ldots,\boldsymbol{\pi}(i,N)),\quad i\geq 0.
\end{aligned}\tag{20}$$

It is well known that the probability vectors $\boldsymbol{\pi}_i$, $i \geq 0$, satisfy the following system of linear algebraic equations:

$$(\boldsymbol{\pi}_0,\boldsymbol{\pi}_1,\ldots)Q=\mathbf{0},\qquad (\boldsymbol{\pi}_0,\boldsymbol{\pi}_1,\ldots)\mathbf{e}=1.\tag{21}$$

System (21) is infinite, so it can not be solved on a computer by standard methods. To compute the probability vectors $\boldsymbol{\pi}_i, i \geq 0$, in both the cases, $\alpha > 0$ and $\alpha = 0$, the numerically stable algorithms presented in [39] can be used. The idea of these algorithms is as follows. Instead of solution of the system (21), another infinite system of linear algebraic equations for the probability vectors $\boldsymbol{\pi}_i$, $i \geq 0$, is derived by means of sequential constructions of so called censored Markov chains (see, e.g., [41]) for the initial Markov chain ξ_t, $t \geq 0$, with different levels of censoring. This leads to numerically stable procedure for computation of the vectors $\boldsymbol{\pi}_i$, $i \geq 0$.

5. Performance Measures

Having computed the vectors of the stationary probabilities $\boldsymbol{\pi}_i, i \geq 0$, it is possible to compute a variety of the system performance measures.

The average number of customers in the system is

$$L=\sum_{n=1}^{N}n\boldsymbol{\pi}(0,n)\mathbf{e}+\sum_{i=1}^{\infty}\sum_{n=M}^{N}(i+n)\boldsymbol{\pi}(i,n)\mathbf{e}.\tag{22}$$

The average number of customers in the buffer is

$$N^{\text{buffer}}=\sum_{i=1}^{\infty}i\boldsymbol{\pi}_i\mathbf{e}.\tag{23}$$

The average number of busy servers is

$$N^{\text{server}}=\sum_{i=0}^{\infty}\sum_{n=1}^{N}n\boldsymbol{\pi}(i,n)\mathbf{e}.\tag{24}$$

The average number of busy servers providing service to type 1 customers is

$$\begin{aligned}
N^{\text{server}-1}&=\sum_{n=1}^{N}\sum_{l=0}^{\min\{n,M\}}(n-l)\boldsymbol{\pi}(0,n,l)\mathbf{e}\\
&\quad+\sum_{i=1}^{\infty}\sum_{n=M}^{N}\sum_{l=0}^{M}(n-l)\boldsymbol{\pi}(i,n,l)\mathbf{e}.
\end{aligned}\tag{25}$$

The average number of busy servers providing service to type 2 customers is

$$\begin{aligned}
N^{\text{server}-2}&=\sum_{n=1}^{N}\sum_{l=1}^{\min\{n,M\}}l\boldsymbol{\pi}(0,n,l)\mathbf{e}\\
&\quad+\sum_{i=1}^{\infty}\sum_{n=M}^{N}\sum_{l=1}^{M}l\boldsymbol{\pi}(i,n,l)\mathbf{e}=N^{\text{server}}-N^{\text{server}-1}.
\end{aligned}\tag{26}$$

The intensity of output flow of type 1 customers is

$$\lambda_{\text{out}}^{(1)}=\mu_1 N^{\text{server}-1}.\tag{27}$$

The intensity of output flow of type 2 customers is

$$\lambda_2^{\text{out}}=\mu_2 N^{\text{server}-2}.\tag{28}$$

The intensity of output flow of customers is

$$\lambda_{\text{out}}=\lambda_1^{\text{out}}+\lambda_2^{\text{out}}.\tag{29}$$

The loss probability of type 1 customers is

$$P_1^{\text{loss}}=\lambda_1^{-1}\sum_{i=0}^{\infty}\boldsymbol{\pi}(i,N,0)\left(D_1\otimes I_{\overline{Z}}\right)\mathbf{e}=1-\frac{\lambda_1^{\text{out}}}{\lambda_1}.\tag{30}$$

The loss probability of type 2 customers is

$$P_2^{\text{loss}}=1-\frac{\lambda_2^{\text{out}}}{\lambda_2}.\tag{31}$$

The loss probability of an arbitrary customer is

$$P^{\text{loss}}=1-\frac{\lambda_{\text{out}}}{\lambda_1+\lambda_2}.\tag{32}$$

The probability of type 2 customer loss at the entrance to the system is

$$P^{\text{ent-loss}} = (1-q)\lambda_2^{-1}\sum_{i=0}^{\infty}\sum_{n=M}^{N}\boldsymbol{\pi}(i,n)\left(I_{(M+1)\overline{W}}\otimes H_1\right)\mathbf{e}. \quad (33)$$

The probability that an arbitrary type 2 customer will be forced to terminate its service and go to the buffer is

$$P^{\text{knock-out-to-buffer}} = p\lambda_1^{-1}\sum_{i=0}^{\infty}\sum_{l=1}^{M}\boldsymbol{\pi}(i,N,l)\left(D_1\otimes I_{\overline{Z}}\right)\mathbf{e}. \quad (34)$$

The probability that an arbitrary type 2 customer will be forced to terminate its service and leave the system is

$$P^{\text{knock-out-loss}} = (1-p)\lambda_1^{-1}\sum_{i=0}^{\infty}\sum_{l=1}^{M}\boldsymbol{\pi}(i,N,l)\left(D_1\otimes I_{\overline{Z}}\right)\mathbf{e}. \quad (35)$$

The probability that an arbitrary type 2 customer will leave the buffer due to impatience is

$$P^{\text{imp-loss}} = P_2^{\text{loss}} - P^{\text{ent-loss}} - P^{\text{knock-out-loss}}. \quad (36)$$

6. Distribution of the Waiting Time of an Arbitrary Type 2 Customer

Type 1 customers having preemptive priority do not wait for service. So, we analyze only the distribution of waiting time of an arbitrary type 2 customer. Speaking about the arbitrary type 2 customer, we do not distinguish type 2 customers arriving to the system from outside and customers arriving to the buffer due to the service force termination.

Let $V(x)$ be the distribution function of the waiting time of an arbitrary type 2 customer in the system and let $v(s) = \int_0^{\infty} e^{-sx}dV(x)$, Re $s > 0$, be its Laplace-Stieltjes transform (LST).

To derive an expression for the LST $v(s)$, we use the method of collective marks; see, for example, [42, 43]. Let us tag an arbitrary type 2 customer and keep track of its staying in the system. According to the idea of the method of collective marks, $v(s)$ has the meaning of the probability that no catastrophe from some virtual stationary Poisson flow of catastrophes with the intensity s arrives during the waiting time of the tagged type 2 customer.

Let $v(s,r,n,l,\nu)$ be the probability that no catastrophe arrive during the rest of the tagged customer waiting time conditioned on the fact that, at the given moment, the position of the tagged customer in the buffer is equal to $r, r \geq 1$, the number of busy servers is $n, n = \overline{M,N}$, the number of busy servers providing service to type 2 customers is $l, l = \overline{0,M}$, and the state of the process $\nu_t, t \geq 0$, is ν.

Let us enumerate the probabilities $v(s,r,n,l,\nu)$ in the lexicographic order of the components n, l, and ν and form the following column vectors:

$$\begin{aligned}
\mathbf{v}(s,r,n,l) &= (v(s,r,n,l,0),\dots,v(s,r,n,l,W))^T, \quad l=\overline{0,M},\\
\mathbf{v}(s,r,n) &= (\mathbf{v}(s,r,n,0),\dots,\mathbf{v}(s,r,n,M))^T, \quad n=\overline{M,N},\\
\mathbf{v}(s,r) &= (\mathbf{v}(s,r,M),\dots,\mathbf{v}(s,r,N))^T, \quad r\geq 1.
\end{aligned} \quad (37)$$

Let us also introduce the following notation:

$$V_{r,r} = \begin{pmatrix}
\overline{A}_r^{(M)} & I_{M+1}\otimes D_1 & O & \dots & O & O\\
\overline{F}^{(M+1)} & \overline{A}_r^{(M+1)} & I_{M+1}\otimes D_1 & \dots & O & O\\
\vdots & \vdots & \ddots & \ddots & \vdots & \vdots\\
O & O & O & \ddots & \overline{A}_r^{(N-1)} & I_{M+1}\otimes D_1\\
O & O & O & \dots & \overline{F}^{(N)} & \overline{A}_r^{(N)}
\end{pmatrix} + I_{(N-M+1)(M+1)}\otimes D_0, \quad r\geq 1, \quad (38)$$

$$\begin{aligned}
V_{r,r-1} &= \widehat{I}_{N-M+1}\otimes\left(\mu_1\overline{C}_M E^{+} + \mu_2 C_M\right)\otimes I_{\overline{W}}\\
&\quad + (r-1)\alpha I_{(N-M+1)(M+1)\overline{W}}, \quad r\geq 1,
\end{aligned}$$

where

$$\overline{A}_r^{(n)} = \begin{cases}
-\left(\mu_1\widetilde{C}_n + \mu_2 C_M + r\alpha I_{M+1}\right)\otimes I_{\overline{W}},\\
\qquad M\leq n<N, r\geq 0,\\
-\left(\mu_1\widetilde{C}_n + \mu_2 C_M + r\alpha I_{M+1}\right)\otimes I_{\overline{W}}\\
\quad +\left(E^{-} + \widehat{I}_{M+1}\right)\otimes D_1,\\
\qquad n=N, r\geq 0;
\end{cases} \quad (39)$$

$$\overline{F}^{(n)} = \left(\mu_1\widetilde{C}_n + \mu_2 C_M E^{-}\right)\otimes I_{\overline{W}}, \quad M<n\leq N. \quad (40)$$

Lemma 4. *The vectors* $\mathbf{v}(s,r)$, $r \geq 1$, *can be recursively calculated as follows:*

$$\begin{aligned}
\mathbf{v}(s,1) &= \left(sI - V_{1,1}\right)^{-1}\\
&\quad\times\left[\alpha I + \widehat{I}_{N-M+1}\otimes\left(\mu_1\widetilde{C}_M + \mu_2 C_M\right)\otimes I_{\overline{W}}\right]\mathbf{e},\\
\mathbf{v}(s,r) &= \left(sI - V_{r,r}\right)^{-1}\left(V_{r,r-1}\mathbf{v}(s,r-1) + \alpha\mathbf{e}\right), \quad r>1.
\end{aligned} \quad (41)$$

Proof. Based on a probabilistic sense of the *LST*, we obtain the following system of equations for calculation of the vectors $\mathbf{v}(s,r,n,l)$:

$$\begin{aligned}
&\mathbf{v}(s,r,n,l)\\
&\quad=\left[(s+r\alpha+(n-l)\mu_1+l\mu_2)I_{\overline{W}}-D_0\right]^{-1}\\
&\qquad\times\Big(\left(1-\delta_{n,N}\right)D_1\mathbf{v}(s,r,n+1,l)\\
&\qquad\quad+(r-1)\alpha\mathbf{v}(s,r-1,n,l)\\
&\qquad\quad+\delta_{n,N}\left(\delta_{l,0}D_1\mathbf{v}(s,r,n,l)\right.\\
&\qquad\qquad\left.+\left(1-\delta_{l,0}\right)D_1\mathbf{v}(s,r,n,l-1)\right)\\
&\qquad\quad+\left(1-\delta_{r,1}\right)\delta_{n,M}\left(l\mu_2\mathbf{v}(s,r-1,n,l)\right.\\
&\qquad\qquad\left.+(n-l)\mu_1\mathbf{v}(s,r-1,n,l+1)\right)\\
&\qquad\quad+\delta_{r,1}\delta_{n,M}\left(l\mu_2+(n-l)\mu_1\right)\mathbf{e}+\alpha\mathbf{e}\\
&\qquad\quad+\left(1-\delta_{n,M}\right)\left(l\mu_2\mathbf{v}(s,r,n-1,l-1)\right.\\
&\qquad\qquad\left.+(n-l)\mu_1\mathbf{v}(s,r,n-1,l)\right)\Big),\\
&\qquad l=\overline{0,M},\quad n=\overline{M,N},\quad r\geq 1.
\end{aligned}\tag{42}$$

System (42) can be rewritten in the following matrix form:

$$\begin{aligned}
&\left(-sI+V_{r,r}\right)\mathbf{v}(s,r)+\left(1-\delta_{r,1}\right)V_{r,r-1}\mathbf{v}(s,r-1)\\
&\quad+\alpha\mathbf{e}+\delta_{r,1}\left(\widehat{I}_{N-M+1}\otimes\left(\mu_2C_M+\mu_1\widetilde{C}_M\right)\otimes I_{\overline{W}}\right)\mathbf{e}=0,\\
&\qquad\qquad r\geq 1,
\end{aligned}\tag{43}$$

which can be further transformed to form (41). Lemma is proved. □

Theorem 5. *The LST $v(s)$ of the distribution of the waiting time of an arbitrary type 2 customer in the system is calculated as follows:*

$$\begin{aligned}
&v(s)\\
&=\widetilde{\lambda}^{-1}\left(\sum_{n=0}^{M-1}\boldsymbol{\pi}(0,n)\left(I_{(n+1)\overline{W}}\otimes H_1\right)\mathbf{e}\right.\\
&\qquad+\left(1-q\right)\sum_{i=0}^{\infty}\sum_{n=M}^{N}\boldsymbol{\pi}(i,n)\left(I_{(M+1)\overline{W}}\otimes H_1\right)\mathbf{e}\\
&\qquad+q\sum_{i=0}^{\infty}\sum_{n=M}^{N}\sum_{l=0}^{M}\boldsymbol{\pi}(i,n,l)\left(I_{\overline{W}}\otimes H_1\mathbf{e}_{\overline{Z}}\right)\mathbf{v}(s,i+1,n,l)\\
&\qquad\left.+\ p\sum_{i=0}^{\infty}\sum_{l=1}^{M}\boldsymbol{\pi}(i,N,l)\left(D_1\otimes\mathbf{e}_{\overline{Z}}\right)\mathbf{v}(s,i+1,N,l-1)\right),
\end{aligned}\tag{44}$$

where

$$\widetilde{\lambda}=\lambda_2+p\sum_{i=0}^{\infty}\sum_{l=1}^{M}\boldsymbol{\pi}(i,N,l)\left(D_1\otimes I_{\overline{Z}}\right)\mathbf{e}.\tag{45}$$

Proof. The proof follows from the law of total probability and a probabilistic sense of the *LST*s.

The following situations are possible during the arrival epoch of the tagged type 2 customer.

(i) The number of busy servers is less than M and the tagged customer immediately receives service. The probability of this event is $\widetilde{\lambda}^{-1}\sum_{n=0}^{M-1}\boldsymbol{\pi}(0,n)(I_{(n+1)\overline{W}}\otimes H_1)\mathbf{e}$. In this case, the probability of no catastrophe arrival during the waiting time is equal to one.

(ii) The number of busy servers is greater than M and the customer decides to balk the system. The probability of this event is $\widetilde{\lambda}^{-1}(1-q)\sum_{i=0}^{\infty}\sum_{n=M}^{N}\boldsymbol{\pi}(i,n)(I_{(M+1)\overline{W}}\otimes H_1)\mathbf{e}$. In this case, the probability of no catastrophe arrival during the waiting time of the tagged type 2 customer is also equal to one.

(iii) The number of busy servers is greater than M and the customer decides to join the buffer. The probability of this event is $\widetilde{\lambda}^{-1}q\sum_{i=0}^{\infty}\sum_{n=M}^{N}\sum_{l=0}^{M}\boldsymbol{\pi}(i,n,l)(I_{\overline{W}}\otimes H_1\mathbf{e}_{\overline{Z}})$. In this case, the probability of no catastrophe arrival during the waiting time of the tagged type 2 customer under the fixed values of the components i,n, and l is equal to $\mathbf{v}(s,i+1,n,l)$.

(iv) The tagged customer arrives to the buffer after the termination of its service by arriving type 1 customer. The probability of this event is $\widetilde{\lambda}^{-1}p\sum_{i=0}^{\infty}\sum_{l=1}^{M}\boldsymbol{\pi}(i,N,l)(D_1\otimes\mathbf{e}_{\overline{Z}})$. In this case, the probability of no one catastrophe arrival during the waiting time of the tagged type 2 customer under the fixed values of the components i,n,l is equal to $\mathbf{v}(s,i+1,N,l-1)$.

Using the law of total probability, now it is possible to easily verify the validity of the statement of the theorem. □

Corollary 6. *The average waiting time V^{wait} of an arbitrary type 2*

$$\begin{aligned}
&V^{wait}\\
&=-\widetilde{\lambda}^{-1}\left(p\sum_{i=0}^{\infty}\sum_{l=1}^{M}\boldsymbol{\pi}(i,N,l)\left(D_1\otimes\mathbf{e}_{\overline{Z}}\right)\right.\\
&\qquad\times\mathbf{v}'(s,i+1,N,l-1)|_{s=0}\\
&\qquad+q\sum_{i=0}^{\infty}\sum_{n=M}^{N}\sum_{l=0}^{M}\boldsymbol{\pi}(i,n,l)\left(I_{\overline{W}}\otimes H_1\mathbf{e}_{\overline{Z}}\right)\\
&\qquad\left.\times\mathbf{v}'(s,i+1,n,l)|_{s=0}\right).
\end{aligned}\tag{46}$$

Here, the column vectors $\mathbf{v}'(s,r,n,l)|_{s=0}$ are defined as blocks of the vectors $\mathbf{v}'(s,r)|_{s=0}, r\geq 1$, which can be calculated as

$$\begin{aligned}
&\mathbf{v}'(s,1)\Big|_{s=0}\\
&\quad=-\left(V_{1,1}\right)^{-2}\left[\alpha I+\widehat{I}_{N-M+1}\otimes\left(\mu_2C_M+\mu_1\widetilde{C}_M\right)\otimes I_{\overline{W}}\right]\mathbf{e},\\
&\mathbf{v}'(s,r)\Big|_{s=0}\\
&\quad=\left(V_{r,r}\right)^{-1}\left(\mathbf{e}-V_{r,r-1}\mathbf{v}'(s,r-1)\right),\quad r>1.
\end{aligned}\tag{47}$$

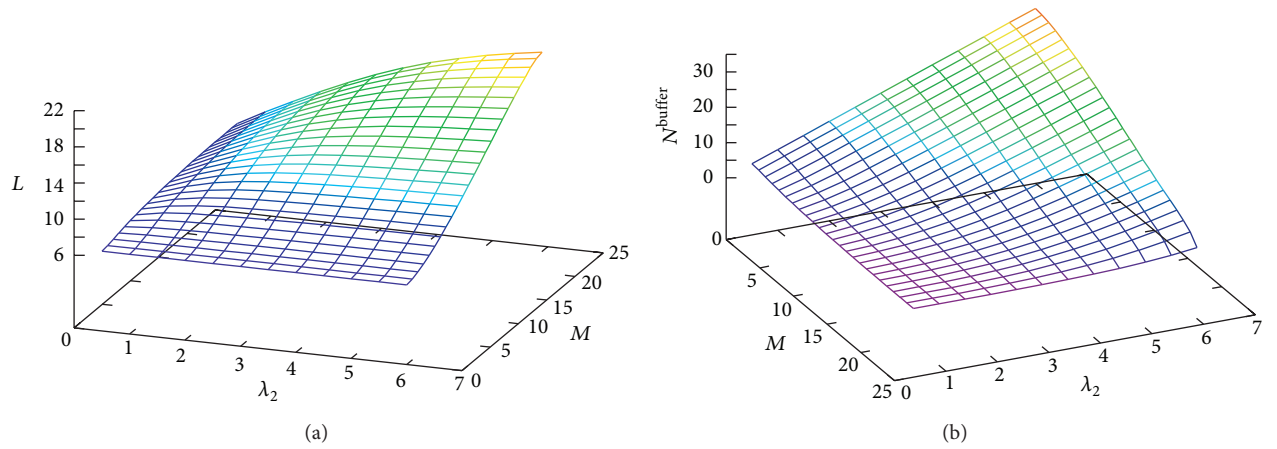


Figure 1: Dependence of L and N^{buffer} on M and λ_2.

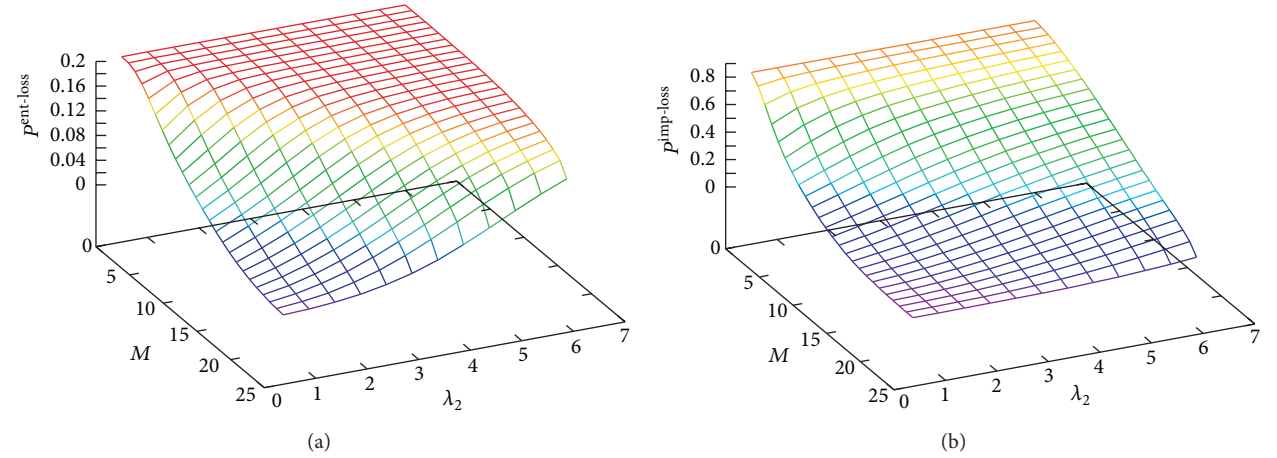


Figure 2: Dependence of $P^{\text{ent-loss}}$ and $P^{\text{imp-loss}}$ on M and λ_2.

7. Numerical Examples and Optimization Problem

The purposes of this section are to demonstrate the feasibility of the proposed algorithms for computation of the key performance measures of the system, to give an example of numerical solution of optimization problem and to bring out some qualitative aspects of the considered queue.

Let the arrival process of type 1 customers be defined by the matrices

$$D_0 = \begin{pmatrix} -3.64163 & 0.10758 \\ 0.04921 & -0.31828 \end{pmatrix},$$
$$D_1 = \begin{pmatrix} 3.45660 & 0.07745 \\ 0.06276 & 0.20631 \end{pmatrix}. \tag{48}$$

The fundamental rate of this arrival process is $\lambda_1 = 1.5$, the coefficient of correlation of successive interarrival times is $c_{\text{cor}} = 0.25$, the coefficient of variation of interarrival times is $c_{\text{var}} = 5.4$.

Basically, the arrival process of type 2 customers is defined by the matrices

$$H_0 = \begin{pmatrix} -0.6759 & 0 \\ 0 & -0.02193 \end{pmatrix},$$
$$H_1 = \begin{pmatrix} 0.67141 & 0.00449 \\ 0.01222 & 0.00971 \end{pmatrix}. \tag{49}$$

The fundamental rate of this arrival process is $\lambda_2 = 0.5$, the coefficient of correlation of successive interarrival times is $c_{\text{cor}} = 0.2$, and the coefficient of variation of interarrival times is $c_{\text{var}} = 12.34$.

The rest of the system parameters are given by

$$N = 24, \quad \alpha = 0.15, \quad q = 0.8, \quad p = 0.1, \quad \mu_1 = 0.22, \quad \mu_2 = 0.3. \tag{50}$$

To illustrate behavior of the key performance measures of the system, we will vary the threshold M of admission strategy (recall that a type 2 customer is admitted to service only if the number of busy servers at its arrival instant is less

than M) in the range $M \in [1;\ 24]$ and the intensity λ_2 of the arrival process of type 2 customers in the range $\lambda_2 \in [0.5;\ 6]$. Variation of the intensity λ_2 is easily implemented by means of multiplying the matrices H_0 and H_1 by the corresponding factor.

Three-dimensional Figure 1 illustrates the dependence of the average number of customers in the system L and the average number of customers in the buffer N^{buffer} on M and λ_2.

It is quite natural that both L and N^{buffer} increase when the intensity λ_2 grows: the grow of the arrival rate under the fixed services rates causes presence of more customers in the system and in the buffer. But the increase of M oppositely affects the values L and N^{buffer}. When M grows, the value of L increases because the growth of M decreases the probability of type 2 customers balking (increases the rate of the flow of admitted customers). The value of N^{buffer} decreases when M grows because the growth of M increases the percentage of type 2 customers that succeed to get access to service immediately upon arrival, without visiting the buffer. However, when M becomes too close to N, many type 2 customers are compelled to terminate service due to type 1 customer arrival and the value of N^{buffer} stops decreasing.

Figure 2 illustrates the dependence of the probability $P^{\text{ent-loss}}$ of the type 2 customer loss at the entrance to the system and the probability $P^{\text{imp-loss}}$ that an arbitrary type 2 customer leaves the buffer due to impatience on M and λ_2.

It can be seen that the value of $P^{\text{ent-loss}}$ sharply grows when M decreases and λ_2 increases. It is obvious because the decrease of M and the growth of λ_2 imply congestion in the system for type 2 customers. When M is small or λ_2 is large, type 2 customers practically have the chance to start service immediately upon arrival. In this situation, they balk with probability $1 - q = 0.2$ what explains the presence of almost plate part on the surface in Figure 2 at level 0.2. Behavior of $P^{\text{imp-loss}}$ correlates with behavior of N^{buffer} in Figure 1 what is easily understandable because the rate of customers departure from the system due to impatience positively correlates with the number of customers in the buffer.

Figure 3 illustrates the dependence of the probability $P^{\text{knock-out-loss}}$ that an arbitrary type 2 customer will be forced to terminate service and will leave the system and the probability P_2^{loss} that an arbitrary type 2 customer is lost on M and λ_2.

The plot illustrating behavior of the probability $P^{\text{knock-out-loss}}$ well illustrates the reasonability of reservation of some part of servers exclusively for service of priority customers. When the parameter M approaches the value N, that is, the number of reserved servers decreases, the probability $P^{\text{knock-out-loss}}$ drastically increases, especially when the arrival rate λ_2 is not very small. The plot illustrating behavior of the probability P_2^{loss} is quite clear taking in mind that this probability is the sum of the probabilities $P^{\text{ent-loss}}$, $P^{\text{knock-out-loss}}$, and $P^{\text{imp-loss}}$ illustrated in the previous figures.

Figure 4 illustrates the dependence of the intensity of output flow of type 2 customers λ_2^{out} and the average waiting time V^{wait} of an arbitrary type 2 customer on M and λ_2.

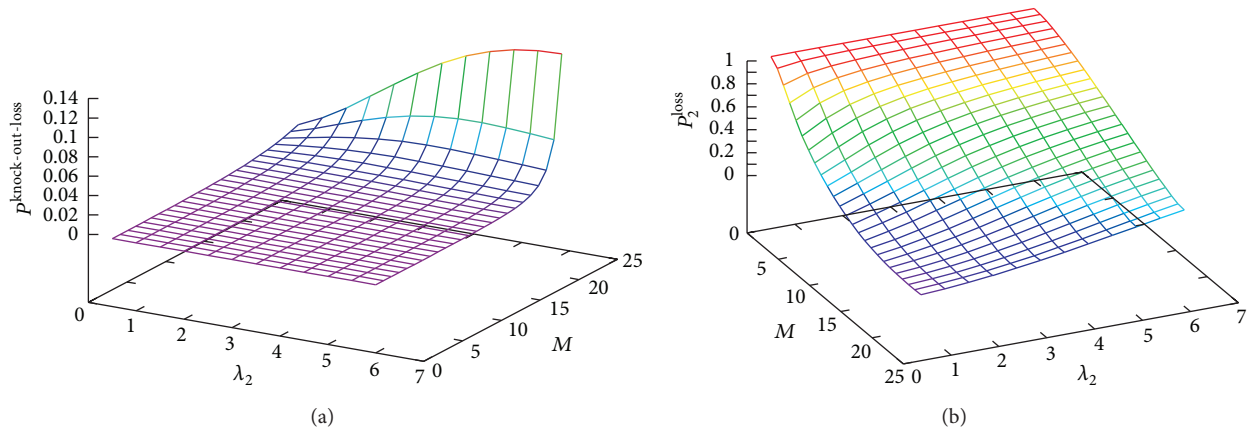
Behavior of λ_2^{out} is quite clear taking into account the illustrated above behavior of P_2^{loss} and formula $P_2^{\text{loss}} = 1 - (\lambda_2^{\text{out}}/\lambda_2)$. Behavior of V^{wait} is similar to behavior of N^{buffer} in Figure 1. However, V^{wait} does not grow so quickly for small M and large λ_2 as the value N^{buffer} grows. This is explained by the fact that it is accounted for in calculation of the average waiting time V^{wait} of an arbitrary type 2 customer that, by definition, some such customers do not visit the buffer at all (they are lost at the entrance to the system or are lucky to receive service without visiting the buffer) and, so, have zero waiting time.

Let us consider optimization problem. First of all, we have to formulate a cost criterion in terms of which the quality of system operation will be evaluated. Because type 1 customers have preemptive priority, the quality of their service is completely defined by the value of P_1^{loss} which does not depend on the arrival and service processes of type 2 customers. P_1^{loss} is constant for any values of the control parameter M. So, cost criterion should include only characteristics of service of type 2 customers. Because the goal of the operation of almost any queueing system is to provide maximal profit to the system owner and this profit is proportional to the number of customers, who got service in the system, it is reasonable to include the output rate λ_2^{out} as the main component of the cost criterion. But if the cost criterion will include only the output rate λ_2^{out}, solution of optimization problem seems be trivial $M = N$; that is, no server reservation should be assumed. Any arriving customer should be admitted to the system if ergodicity condition for the system is fulfilled. But, as we see from the presented numerical results, the absence of servers reservation implies high values of the probabilities of losses of type 2 customers, in particular, the probability of force termination of service provided to type 2 customers. Such a termination can be quite offensive and frustrating for type 2 customers. Also the absence of servers reservation causes long average waiting time, especially waiting time of customers who really wait for beginning of service.

As the result of these considerations, we conclude that the reasonable form of the cost criterion (profit obtained during a unit of time) is the following one:

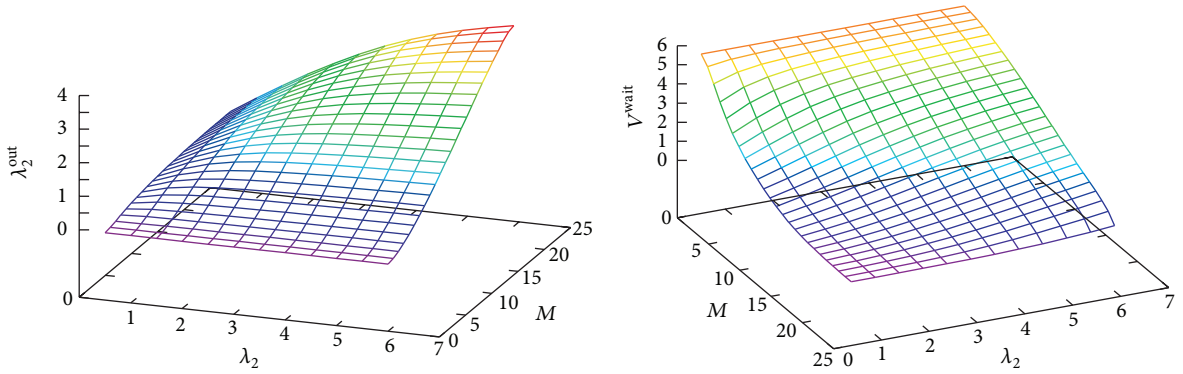
$$\begin{aligned} E(M, \lambda_2) = {} & a\lambda_2^{\text{out}} \\ & - \lambda_2 \left(c_1 P^{\text{ent-loss}} + c_2 P^{\text{imp-loss}} + c_3 P^{\text{knock-out-loss}} \right) \\ & - c_4 V^{\text{wait}}, \end{aligned} \tag{51}$$

where the performance indices λ_2^{out}, $P^{\text{ent-loss}}$, $P^{\text{imp-loss}}$, $P^{\text{knock-out-loss}}$, and V^{wait} were introduced above, a is the average profit earned by service of one type 2 customer, and c_1, c_2, c_3, c_4, are the corresponding charged for customer loss and waiting. Our aim is to find the optimal value M^* of the parameter M that provides the maximal value of the cost criterion $E(M, \lambda_2)$. It should be noted that maximization of $E(M, \lambda_2)$ has to be done with respect to M for a fixed value



(a) (b)

Figure 3: Dependence of $P^{\text{knock-out-loss}}$ and P_2^{loss} on M and λ_2.



(a) (b)

Figure 4: Dependence λ_2^{out} and V^{wait} on M and λ_2.

of λ_2. We included λ_2 to notation $E(M, \lambda_2)$ just to stress that the optimal value M^* may be different for various values of λ_2.

Let us fix the same values of parameters of the system as above and the values of cost coefficients are fixed as follows:

$$a = 10, \quad c_1 = 5, \quad c_2 = 3, \quad c_3 = 20, \quad c_4 = 3. \tag{52}$$

Figure 5 illustrates the dependence of the cost criterion $E(M, \lambda_2)$ on M and λ_2.

Presented table gives a bit more information about the behavior of the cost criterion $E(M, \lambda_2)$ on M. For various λ_2, the table gives the optimal value M^* of the parameter M, optimal value $E(M^*, \lambda_2)$ of the cost criterion, value $E(N, \lambda_2)$ of the cost criterion for the system without servers reservation, the difference $E(M^*, \lambda_2) - E(N, \lambda_2)$ (the profit provided by means of the optimal reservation), and the ratio $((E(M^*, \lambda_2) - E(N, \lambda_2))/E(N, \lambda_2)) \times 100\%$ of the relative profit provided by means of the optimal reservation comparing to a strategy of customers access without servers reservation.

It is seen from Figure 5 and Table 1 that the profit and the relative profit provided by means of the optimal reservation comparing to a strategy of customers access without servers reservation can be significant.

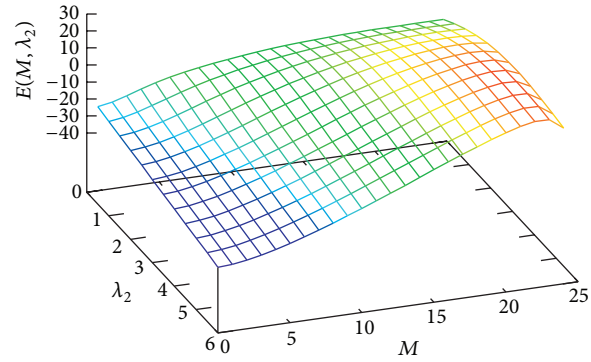


Figure 5

8. Conclusion

We considered a multiserver queuing system with an infinite buffer and two types of customers one of which has preemptive priority over another. To reduce the frequency of forced termination of service of type 2 customers, we offered the threshold strategy of access of type 2 customers. Access of such customers is denied if the number of busy servers at the customer arrival moment is not less than the preassigned

Table 1: Information about the optimal values of the threshold, cost criterion, and profit in comparison to the system without admission control for various intensities λ_2.

λ_2	M^*	$E(M^*, \lambda_2)$	$E(N, \lambda_2)$	$E(M^*, \lambda_2) - E(N, \lambda_2)$	$\frac{E(M^*, \lambda_2) - E(N, \lambda_2)}{E(N, \lambda_2)} \times 100\%$
0.5	23	4.76106	4.74152	0.01954	0.4121%
1	22	9.24359	9.1041	0.13949	1.5321%
1.5	22	13.3147	12.9048	0.4099	3.1763%
2	22	16.8746	16.0199	0.8547	5.3352%
2.5	22	19.8463	18.3934	1.4529	7.899%
3	22	22.1576	20.013	2.1446	10.716%
3.5	22	23.752	20.8855	2.8665	13.7248%
4	22	24.6281	21.0436	3.5845	17.0336%
4.5	22	24.8576	20.5675	4.2901	20.8586%
5	22	24.562	19.5818	4.9802	25.4327%
5.5	22	23.8748	18.2299	5.6449	30.9650%
6	22	22.9148	16.6467	6.2681	37.6537%

threshold value. For the fixed value of the threshold, we described behavior of the system by the multidimensional Markov chain. For this chain, the ergodicity condition is derived; the stationary distribution of the system states and the main performance measures are calculated. The Laplace-Stieltjes transform of the waiting time distribution of an arbitrary type 2 customer is derived. Numerical examples confirming the advantage of the proposed access strategy under properly chosen value of the threshold are presented.

Presented analysis may be extended to the cases of the *MMAP* (*marked Markovian arrival process*) of customers arrival (what allows to take into account possible correlation within arrivals of priority and nonpriority customers) and the *BMAP* (*batch Markovian arrival process*) of nonpriority customers, possibility to have an additional finite buffer for priority customers, and so forth.

In this paper we assume that the service time distribution for both types of customers is exponential. This is rather restrictive assumption. It is worth to note that in many practical situations available information about the service time is only the mean service time. In this case, there is no chance to estimate parameters of the service time distribution and assumption about exponential service time distribution is quite natural because it greatly simplifies the mathematical analysis. If the detailed statistics about the service time is available, more general distribution of service time can be suggested. The presented analysis could be more or less easily extended to the case of much more general phase type distribution (see [40]), which can be used to approximate many other distributions. However, in this case we meet the following difficulty. If service time distribution is exponential, it does not matter (from the point of view of the distribution of the number of customers in the system) in which busy server the service of nonpriority customer is interrupted. In case of phase type distribution, in situation when the service in some server should be terminated it is necessary to fix the rule of a choice of the concrete busy server in which service will be terminated; for example, the phase of the service at this server should belong to some fixed group of phases. The question of justification of a fixed rule is quite complicated. If this question will be answered, the problem of analysis of the corresponding queueing model can be solved by analogy with presented above analysis, but the dimension of the blocks of generator of the Markov chain, which describes behavior of the system, catastrophically increases with increase of the number of phases and (or) servers due to necessity of taking into account the current phase of service in each busy server or the number of servers providing currently the service at each phase. Solution of this problem may be simplified by means of a proper use of methodology by D. Lucantoni and V. Ramaswami by analogy with, for example, [44] or more involved methodology from [45].

In this paper we assumed that the service time distribution for both types of customers is exponential. This is rather restrictive assumption. But it is worth to note that in many practical situations available information about the service time is only the mean service time. In this case, there is no chance to estimate the parameters of the service time distribution and the assumption about an exponential service time distribution is quite natural because it greatly simplifies the mathematical analysis. If the detailed statistics about the service time is available, more general distributions of service time can be suggested. The presented analysis could be more or less easily extended to the case of much more general phase type distribution (see [40]) which can be used to approximate many other distributions. However, in this case we meet the following difficulty. If the service time distribution is exponential, it does not matter in which busy server service of nonpriority customer is interrupted. In the case of phase type distribution, in the situation when service in some server should be terminated it is necessary to fix the rule of a choice of the concrete busy server in which service will be terminated; for example, the phase of service at this server should belong to some fixed group of phases. The question of justification of the fixed rule is quite complicated. If this question will be answered, the problem of analysis of the corresponding queueing model can be solved by analogy with presented above analysis, but the dimension of the blocks of

generator of the Markov chain, which describes the behavior of the system, catastrophically increases with increase of the number of phases and (or) servers due to necessity of taking into account the current phase of service in each busy server or the number of servers providing currently service at each phase. Solution of this problem may be simplified by means of a proper use of methodology by D. Lucantoni and V. Ramaswami by analogy with, for example, [44] or more involved methodology from [45].

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgment

This work was supported by MEST 2012-002512, NRF, Korea.

References

[1] S. R. Chakravarthy, "The batch Markovian arriv al process: a review and future work," in *Advances in Probability Theory and Stochastic Processes*, pp. 21–49, Notable Publications, Englewood Cliffs, NJ, USA, 2001.

[2] D. M. Lucantoni, "New results on the single server queue with a batch Markovian arrival process," *Communications in Statistics. Stochastic Models*, vol. 7, no. 1, pp. 1–46, 1991.

[3] Q. M. He, "Queues with marked customers," *Advances in Applied Probability*, vol. 28, no. 2, pp. 567–587, 1996.

[4] P. P. Bocharov, C. D'Apice, A. V. Pechinkin, and S. Salerno, *Queueing Theory*, Utrecht-Boston, VSP, Boston, Mass, USA, 2004.

[5] B. D. Choi and G. U. Hwang, "The MAP,M/G1,G2/1 queue with preemptive priority," *Journal of Applied Mathematics and Stochastic Analysis*, vol. 10, no. 4, pp. 407–421, 1997.

[6] G. Horváth, "Efficient analysis of the queue length moments of the *MMAP/MAP*/1 preemptive priority queue," *Performance Evaluation*, vol. 69, no. 12, pp. 684–700, 2012.

[7] F. Machihara, "A bridge between preemptive and non-preemptive queueing models," *Performance Evaluation*, vol. 23, no. 2, pp. 93–106, 1995.

[8] T. Takine and B. Sengupta, "A single server queue with service interruptions," *Queueing Systems*, vol. 26, no. 3-4, pp. 285–300, 1997.

[9] K. Al-Begain, A. Dudin, V. Klimenok, and S. Dudin, "Generalized survivability analysis of systems with propagated failures," *Computers and Mathematics with Applications*, vol. 64, no. 12, pp. 3777–3791, 2012.

[10] S. Dudin, C. Kim, and O. Dudina, "*MMAP/M/N* queueing system with impatient heterogeneous customers as a model of a contact center," *Computers & Operations Research*, vol. 40, no. 7, pp. 1790–1803, 2013.

[11] C. Kim and S. Dudin, "Priority tandem queueing model with admission control," *Computers and Industrial Engineering*, vol. 61, no. 1, pp. 131–140, 2011.

[12] A. Krishnamoorthy, S. Babu, and V. C. Narayanan, "MAP/(PH/PH)/c queue with self-generation of priorities and non-preemptive service," *Stochastic Analysis and Applications*, vol. 26, no. 6, pp. 1250–1266, 2008.

[13] H. Qing and S. R. Chakravarthy, "Analytical and simulation modeling of a multi-server queue with Markovian arrivals and priority services," *Simulation Modelling Practice and Theory*, vol. 28, pp. 12–26, 2012.

[14] M. Senthil Kumar, S. R. Chakravarthy, and R. Arumuganathan, "Preemptive resume priority retrial queue with two classes of MAP arrivals," *Applied Mathematical Sciences*, vol. 7, no. 49–52, pp. 2569–2589, 2013.

[15] C. S. Kim, V. Klimenok, and A. Dudin, "Optimization of guard channel policy in cellular mobile networks with account of retrials," *Computers and Operation Research*, vol. 43, pp. 181–190, 2014.

[16] C. S. Kim, V. Klimenok, and O. Taramin, "A tandem retrial queueing system with two Markovian flows and reservation of channels," *Computers and Operations Research*, vol. 37, no. 7, pp. 1238–1246, 2010.

[17] V. Klimenok and R. Savko, "A retrial tandem queue with two types of customers and reservation of channels," *Communications in Computer and Information Science*, vol. 356, pp. 105–114, 2013.

[18] B. Sun, M. H. Lee, S. A. Dudin, and A. N. Dudin, "Analysis of multiserver queueing system with opportunistic occupation and reservation of servers," *Mathematical Problems in Engineering*, vol. 2014, Article ID 178108, 13 pages, 2014.

[19] I. F. Akyildiz, W. Lee, M. C. Vuran, and S. Mohanty, "NeXt generation/dynamic spectrum access/cognitive radio wireless networks: a survey," *Computer Networks*, vol. 50, no. 13, pp. 2127–2159, 2006.

[20] S. Chen, A. M. Wyglinski, S. Pagadarai, R. Vuyyuru, and O. Altintas, "Feasibility analysis of vehicular dynamic spectrum access via queueing theory model," *IEEE Communications Magazine*, vol. 49, no. 11, pp. 156–163, 2011.

[21] Y. Konishi, H. Masuyama, S. Kasahara, and Y. Takahashi, "Performance analysis of dynamic spectrum handoff scheme with variable bandwidth demand of secondary users for cognitive radio networks," *Wireless Networks*, vol. 19, no. 5, pp. 607–617, 2013.

[22] J. Mitola III and G. Q. Maguire Jr., "Cognitive radio: making software radios more personal," *IEEE Personal Communications*, vol. 6, no. 4, pp. 13–18, 1999.

[23] S. Zahed, I. Awan, and A. Cullen, "Analytical modeling for spectrum handoff decision in cognitive radio networks," *Simulation Modelling Practice and Theory*, vol. 38, pp. 98–114, 2013.

[24] X. A. Zhu, L. A. Shen, and T. S. Yum, "Analysis of cognitive radio spectrum access with optimal channel reservation," *IEEE Communications Letters*, vol. 11, no. 4, pp. 304–306, 2007.

[25] Z. Aksin, M. Armony, and V. Mehrotra, "The modern call center: a multi-disciplinary perspective on operations management research," *Production and Operations Management*, vol. 16, no. 6, pp. 665–688, 2007.

[26] O. Garnett, A. Mandelbaum, and M. Reiman, "Designing a call center with impatient customers," *Manufacturing & Service Operations Management*, vol. 4, no. 3, pp. 208–227, 2002.

[27] F. Iravani and B. Balcioglu, "On priority queues with impatient customers," *Queueing Systems*, vol. 58, no. 4, pp. 239–260, 2008.

[28] C. Kim, S. Dudin, O. Taramin, and J. Baek, "Queueing system *MAP—PH*—N—N+R with impatient heterogeneous customers as a model of call center," *Applied Mathematical Modelling*, vol. 37, no. 3, pp. 958–976, 2013.

[29] O. Jouini, Y. Dallery, and Z. Akşin, "Queueing models for full-flexible multi-class call centers with real-time anticipated

delays," *International Journal of Production Economics*, vol. 120, no. 2, pp. 389–399, 2009.

[30] O. Jouini, A. Pot, G. Koole, and Y. Dallery, "Online scheduling policies for multiclass call centers with impatient customers," *European Journal of Operational Research*, vol. 207, no. 1, pp. 258–268, 2010.

[31] P. Khudyakov, P. D. Feigin, and A. Mandelbaum, "Designing a call center with an IVR (Interactive Voice Response)," *Queueing Systems*, vol. 66, no. 3, pp. 215–237, 2010.

[32] C. Kim, A. Dudin, S. Dudin, and O. Dudina, "Tandem queueing system with impatient customers as a model of call center with Interactive Voice Response," *Performance Evaluation*, vol. 70, no. 6, pp. 440–453, 2013.

[33] C. Kim, O. Dudina, A. Dudin, and S. Dudin, "Queueing system MAP/M/N as a model of call center with call-back option," in *Analytical and Stochastic Modeling Techniques and Applications*, vol. 7314 of *Lecture Notes in Computer Science*, pp. 1–15, 2012.

[34] J. E. Nah and S. Kim, "Workforce planning and deployment for a hospital reservation call center with abandonment cost and multiple tasks," *Computers and Industrial Engineering*, vol. 65, pp. 297–309, 2013.

[35] D. P. Heyman and D. Lucantoni, "Modelling multip le IP traffic streams with rate limits," *IEEE/ACM Transactions on Networking*, vol. 11, no. 6, pp. 948–958, 2003.

[36] A. Klemm, C. Lindemann, and M. Lohmann, "Modeling IP traffic using the batch Markovian arrival process," *Performance Evaluation*, vol. 54, no. 2, pp. 149–173, 2003.

[37] P. Buchholz, P. Kemper, and J. Kriege, "Multi-class Markovian arrival processes and their parameter fitting," *Performance Evaluation*, vol. 67, no. 11, pp. 1092–1106, 2010.

[38] A. Graham, *Kronecker Products and Matrix Calculus: With Applications*, Ellis Horwood, Cichester, UK, 1981.

[39] V. I. Klimenok and A. N. Dudin, "Multi-dimensional asymptotically quasi-Toeplitz Markov chains and their application in queueing theory," *Queueing Systems*, vol. 54, no. 4, pp. 245–259, 2006.

[40] M. Neuts, *Matrix-Geometric Solutions in Stochastic Models—An Algorithmic Approach*, The Johns Hopkins University Press, Baltimore, Md, USA, 1981.

[41] J. G. Kemeni, J. L. Snell, and A. W. Knapp, *Denumerable Markov Chains*, Van Nostrand, New York, NY, USA, 1966.

[42] H. Kesten and J. Runnenburg, *Priority in Waiting Line Problems*, Mathematisch Centrum, Amsterdam, The Netherlands, 1956.

[43] D. van Dantzig, "Chaines de Markof dans les ensembles abstraits et applications aux processus avec regions absorbantes et au probleme des boucles," *Annales de l'Institut Henri Poincare*, vol. 14, pp. 145–199, 1955.

[44] C. S. Kim, V. V. Mushko, and A. N. Dudin, "Computation of the steady state distribution for multi-server retrial queues with phase type service process," *Annals of Operations Research*, vol. 201, no. 1, pp. 307–323, 2012.

[45] C. S. Kim, A. Dudin, O. Dudina, and S. Dudin, "Tandem qu eueing system with infinite and finite intermediate buffers and gene ralized phase-type service time distribution," *European Journal of Operational Research*, vol. 235, pp. 170–179, 2014.