Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2014, Article ID 178108, 13 pages
http://dx.doi.org/10.1155/2014/178108

Research Article

Analysis of Multiserver Queueing System with Opportunistic Occupation and Reservation of Servers

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

[1] *School of Economics and Management, Inner Mongolia University of Science and Technology, Baotou 014010, China*
[2] *Inner Mongolia Industry Informatization and Innovation Research Center, Inner Mongolia University of Science and Technology, Baotou 014010, 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 3 February 2014; Revised 8 April 2014; Accepted 8 April 2014; Published 7 May 2014

Academic Editor: Jyh-Hong Chou

We consider a multiserver queueing system with two input flows. Type-1 customers have preemptive priority and are lost during arrival only if all servers are occupied by type-1 customers. If all servers are occupied, but some provide service to type-2 customers, service of type-2 customer is terminated and type-1 customer occupies the server. If the number of busy servers is less than the threshold M during type-2 customer arrival epoch, this customer is accepted. Otherwise, it is lost or becomes a retrial customer. It will retry to obtain service. Type-2 customer whose service is terminated is lost or moves to the pool of retrial customers. The service time is exponentially distributed with the rate dependent on the customer's type. Such queueing system is suitable for modeling cognitive radio. Type-1 customers are interpreted as requests generated by primary users. Type-2 customers are generated by secondary or cognitive users. The problem of optimal choice of the threshold M is the subject of this paper. Behavior of the system is described by the multidimensional Markov chain. Its generator, ergodicity condition, and stationary distribution are given. The system performance measures are obtained. The numerical results show the effectiveness of considered admission control.

1. Introduction

Multiserver queueing system considered in this paper can be applied for modelling various real life systems. But the primary motivation of its consideration was potential applicability for modelling, performance evaluation, capacity planning, and optimizing cognitive radio systems. Recently, the technology of cognitive radio has attracted considerable attention of many researchers as a promising technology for optimization of the utilization of scare radio frequency spectrum. Dynamic spectrum access allows effective use of radio frequency and prevents its underutilization in many real world networks. It enables unlicensed users to temporarily "borrow" unused spectrum while ensuring that the rights of the incumbent license holders are respected [1]. Problems of optimization of joint access of the primary and secondary users can be effectively solved by means of queueing theory. So, although the term cognitive radio was first introduced in [2] only recently (in 1999), the literature devoted to application of queueing theory to cognitive radio is already extensive. A comprehensive survey [3] devoted to cognitive radio was published in 2006. A search in the database Scopus (using the key words "cognitive radio queue") gives today (the end of November, 2013) 143 references since 2007 including 31 references to papers published in 2013 and 44 published in 2012. Thus, in this paper we will not try to give more or less detailed survey of the existing results. The reader may get some knowledge about the state of the art in this field, for example, from the papers [1, 3–5].

As a rule, it is suggested in the considered models that the primary customers have preemptive priority over the secondary customers. A primary customer is lost during its arrival epoch only if all servers are occupied by primary customers. If all servers are occupied, but at least one of them provides service to the secondary customer, service of one secondary customer is terminated and the primary customer occupies the server. The forced termination of service of the secondary customers may imply at least two negative consequences: dissatisfaction of the secondary customers by the quality of service and wasting the throughput (bandwidth) due to the loss of some already done work. So, it is desirable to introduce some kind of control by admission of the secondary customers. For example, it sounds reasonable to stop admission of the secondary customers when the number of busy servers is large (more than some threshold) and, correspondingly, the risk of the forced termination of service of the secondary customers is high.

Such a kind of admission control was offered, for example, in [6]. It was shown in [6] that the appropriate choice of the threshold may lead to maximization of the throughput of the system. In the model considered in our paper, we apply essentially the same strategy of admission control as the one in [6]. If the total number of servers is equal to N, we fix the threshold M, $0 < M \leq N$. We assume that the secondary customer is accepted for service in the system only if the number of busy servers during its arrival epoch is less than M. The case $M = N$ corresponds to the system without restriction of access. As disadvantage of our model comparing to the one considered in [6] we may mention our assumption that both types of the customers need for their processing exactly one server, while it is assumed in [6] that a primary customer occupies a whole group of servers (channel) and a secondary customer occupies one server (subchannel). The advantages of our model comparing to the one considered in [6] and to the overwhelming majority of the papers devoted to analysis of cognitive radio by means of queueing theory are as follows.

(i) We assume that arrival flow of primary and secondary customers is described by the marked Markovian arrival process (MMAP). This process is the generalization of well-known Markovian arrival process (MAP) to the case of heterogeneous customers. The MAP arrival process was introduced as a versatile Markovian point process (VMPP) by M.F. Neuts in the 70th. The original development of the VMPP contained extensive notations; however these notations were greatly simplified in [7] and ever since this process bears the name Markovian arrival process. The class of MAPs 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 the modern telecommunication networks; see, for example, [8, 9]. In [6] and practically all other papers, it is assumed that the arrival flows of primary and secondary customers are stationary Poisson. The stationary Poisson arrival process is the simplest case of the MAP. If one tries to describe some real life flow based on its traces by means of the stationary Poisson arrival process, he or she is able to fit only the mean arrival rate, but not the variance or higher moments of interarrival times and possible correlation between these times. Analysis of the systems with the MAP is much more complicated comparing to analysis of the system with the stationary Poisson arrival process. It is not possible to get simple formulas for the performance measures of the system. Instead, numerical algorithms should be developed. However, careful account of the correlation and possible high variability in the arrival process is necessary to get satisfactory prediction of values of performance measures. So, the MAP is now popular in the literature.

(ii) We assume that the secondary customer, which is not granted immediate access to the system during its arrival epoch, has options to leave the system permanently or to go to some virtual place called in the literature as orbit and retry to get access to the system after a random amount of time. We do not know papers where effect of retrials is taken into account for the secondary customers, while the retrials are a typical feature of many telecommunication networks. It is worth to mention here good survey of research in retrial queues given by A. Gomez-Corral in [10].

(iii) We allow the secondary customers to be nonpersistent (to leave the system after some unsuccessful retrial) and (or) nonpatient (to leave the system after some random period of staying in orbit).

The mentioned above disadvantage of our model consisting of assumption that both types of the customers need for processing one server, while in some systems another number of servers can be required, can be eliminated by means of considering generalization of our model to the case of the batch marked Markovian arrival process (BMMAP). Technique of analysis will be essentially the same with larger blocks of generator of the underlying Markov chain.

The rest of the paper is organized as follows. In Section 2, the model under consideration is described in detail. In Section 3, the behavior of the system under study is described by the level dependent multidimensional continuous-time Markov chain and the generator of this Markov chain is written down. In Section 4, the ergodicity condition of this Markov chain is derived and the stationary distribution of the system states is calculated. The expressions for key performance measures of the system are presented in Section 5. The numerical results showing reasonability of restriction of access of the secondary customers are given in Section 6. The importance of account of the correlation in the arrival process is clarified. Finally, Section 7 concludes the paper.

2. Mathematical Model

We consider the queueing system having N identical servers without a waiting space (buffer). Arrival of two types of customers is defined by the MMAP—marked Markovian arrival process. This process is defined by the irreducible continuous-time Markov chain ν_t, $t \geq 0$, having a finite state space $\{0, \dots, W\}$. The sojourn time of the chain ν_t in the state ν is exponentially distributed with the parameter λ_ν. After this time expires, with probability $p^{(0)}_{\nu,\nu'}$ the chain ν_t jumps to the state ν' without generation of customers, $\nu, \nu' = \overline{0, W}$, $\nu \neq \nu'$, or with probability $p^{(r)}_{\nu,\nu'}$ it jumps to the state ν' with generation of type-r customer, $r = 1, 2, \nu$, $\nu' = \overline{0, W}$. Here notation $\nu = \overline{0, W}$ means that ν takes the values in the set $\{0, 1, \ldots, W\}$.

The MMAP is completely characterized by the square matrices D_0, $D^{(r)}_1$, $r = 1, 2$, defined as follows: $(D^{(r)}_1)_{\nu,\nu'} = \lambda_\nu p^{(r)}_{\nu,\nu'}, \nu, \nu' = \overline{0, W}$, $r = 1, 2, (D_0)_{\nu,\nu} = -\lambda_\nu, \nu = \overline{0, W}$, $(D_0)_{\nu,\nu'} = \lambda_\nu p^{(0)}_{\nu,\nu'}$, and $\nu, \nu' = \overline{0, W}$, $\nu \neq \nu'$.

The matrix $D(1) = D_0 + D^{(1)}_1 + D^{(2)}_1$ is the generator of the Markov chain ν_t, $t \geq 0$. The average intensity of customers arrival (fundamental rate) λ is defined by the formula $\lambda = \boldsymbol{\theta}(D^{(1)}_1 + D^{(2)}_1)\mathbf{e}$, where $\boldsymbol{\theta}$ is the row vector of the stationary probabilities of the Markov chain ν_t. This vector is the unique solution to the system $\boldsymbol{\theta} D(1) = \mathbf{0}$, $\boldsymbol{\theta}\mathbf{e} = 1$. Here and throughout this paper $\mathbf{e}$ is a column vector of appropriate size consisting of 1's, and $\mathbf{0}$ is a row vector of appropriate size consisting of zeroes. The average intensity of type-r customers arrival λ_r is defined by the formula $\lambda_r = \boldsymbol{\theta} D^{(r)}_1 \mathbf{e}$, $r = 1, 2$.

The squared coefficient of variation c_{var} of intervals between successive arrivals is defined by $c_{\text{var}} = 2\lambda\boldsymbol{\theta}(-D_0)^{-1}\mathbf{e} - 1$. The coefficient of correlation c_{cor} of two successive intervals between arrivals is defined by $c_{\text{cor}} = (\lambda\boldsymbol{\theta}(-D_0)^{-1}(D(1) - D_0)(-D_0)^{-1}\mathbf{e} - 1)/c_{\text{var}}$.

Methods of the estimation of MMAP parameters using a finite set of observed data, such as a set of customer arrival times recorded at a real world system, are presented, for example, in the paper [11].

The service time distribution of type-r customers is exponentially distributed with the rate μ_r, $r = 1, 2$.

We assume that type-1 customers have preemptive priority over type-2 customers. Type-1 customer is always accepted to the system except the situation when, during its arrival moment, all servers are occupied by type-1 customers. In this situation, type-1 customer leaves the system without service (is lost). If all servers are occupied, but at least one of them provides service to type-2 customer during type-1 customer arrival epoch, service of one type-2 customer is terminated and type-1 customer occupies the corresponding server.

Admission to the system of type-2 customers is restricted via the threshold mechanism. Some preassigned threshold M, $0 < M \leq N$, is fixed. The incoming secondary customer is accepted for service in the system only if the number of busy servers during its arrival moment is less than M. This is equivalent to reservation of $N - M$ servers exclusively for service of type-1 customers. The case $M = N$ corresponds to the system without reservation (without restriction of access of type-2 customers). If type-2 customer is admitted to the system, it occupies an arbitrary free server and starts service. If type-2 customer does not get permission to enter the system, with probability $1 - q$, $0 \leq q \leq 1$, it leaves the system permanently (is lost). With the complementary probability, type-2 customer decides to retry to get access later. We say that this customer goes to a virtual place called orbit. A customer in orbit repeats the attempts to get access, independently of other customers from orbit, after a time interval having an exponential distribution with the parameter α, $\alpha > 0$. An attempt will be successful if the number of busy servers during the moment of this attempt is less than M. If the attempt is successful, the customer immediately occupies a free server and starts processing. If the attempt is not successful; with probability $1 - q$ the customer leaves the system permanently. With the complementary probability, type-2 customer returns into orbit.

Service of any type-2 customer may be terminated by the arrival of type-1 customer. In this situation, type-2 customer leaves the system with probability $1 - p$, $0 \leq p \leq 1$, or moves into orbit. The customers staying in orbit may be impatient and leave the system after a random amount of time having an exponential distribution with the parameter γ, $\gamma > 0$. If the customers are patient, we set $\gamma = 0$.

As it was already noted above, the described queueing system is suitable, for example, for modeling, performance evaluation, capacity planning, and optimizing cognitive radio systems. Type-1 customers are interpreted as requests generated by the primary users, while type-2 customers are interpreted as requests generated by the secondary users. The imposed restriction of access of type-2 customers may not look very reasonable because type-1 customers have preemptive priority anyway. But, as it was shown in [6] and will be shown for more complicated models in our paper, under the suitable choice of the threshold, in some situations the restriction may decrease the probability of loss and forced termination of service of type-2 customers and increase the throughput of the system.

In the rest of this paper, we will analyze stationary distribution of the system states and performance measures of the system under various fixed values of the threshold M and solve optimization problem.

We assume that the quality of operation of the system is evaluated by cost criterion:

$$\begin{aligned} J(M) = {} & a\lambda^{(2)}_{\text{out}} \\ & - \lambda_2 \Big(c_1 P^{(\text{ent-loss})} + c_2 P^{(\text{ent-to-orbit})} \\ & \quad + c_3 P^{(\text{termination-loss})} + c_4 P^{(\text{termination-to-orbit})} \\ & \quad + c_5 P^{(\text{loss-from-orbit})} \Big), \end{aligned} \tag{1}$$

where $\lambda^{(2)}_{\text{out}}$ is the intensity of the flow of type-2 customers that receive successful service in the system, a is the profit, which earns the system by successful service of each type-2 customer, $P^{(\text{ent-loss})}$ is the loss probability of type-2 customer at the entrance to the system due to the imposed restriction on access, $P^{(\text{ent-to-orbit})}$ is the probability that, due to the restriction, type-2 customer goes into orbit, $P^{(\text{termination-loss})}$ is the probability that service of type-2 customer is terminated and it is lost, $P^{(\text{termination-to-orbit})}$ is the probability that service of type-2 customer is terminated and it moves into orbit, $P^{(\text{loss-from-orbit})}$ is the loss probability of a customer from orbit, and c_l, $l = \overline{1,5}$, are the charges which should be paid for the corresponding losses.

It is necessary to find the value M^* of the threshold M which maximizes cost criterion (1). To this end, we have to have an opportunity to compute the values of all performance measures of the system, which appear at the right-hand side of (1), for any fixed value of the threshold M. In the next two sections we assume that the threshold M, $0 < M \leq N$, is fixed.

3. Process of the System States

Let

(i) i_t, $i_t \geq 0$, be the number of customers in orbit,

(ii) n_t, $n_t = \overline{0,N}$, the number of busy servers,

(iii) l_t, $l_t = \overline{0, \min\{n_t, M\}}$, the number of type-2 customers in service,

(iv) ν_t, $\nu_t = \overline{0,W}$, the state of underlying process of the MMAP during the moment t, $t \geq 0$.

It is easy to see that the four-dimensional process

$$\xi_t = \{i_t, n_t, l_t, \nu_t\}, \quad t \geq 0, \tag{2}$$

is an irreducible continuous-time multidimensional Markov chain.

Let us enumerate the states of the chain ξ_t in lexicographic order of the components (i, n, l, ν). The set of the states having value (i, n) of two first components will be called as a macrostate (i, n).

Let Q be the generator of the Markov chain ξ_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 intensities of the transitions of the chain ξ_t from the macrostate (i, n) to the macrostate (j, n'), $n, n' = \overline{0, \min\{i, N\}}$. The diagonal entries of the matrices $Q_{i,i}$ are negative. The modulus of the diagonal entry defines intensity of departure from the corresponding state of the Markov chain.

To write down the expression for the generator Q, we need some notation.

Let

(i) I be the identity matrix, and let O be a zero matrix. If the dimension of a matrix is not clear from context, it is indicated by the suffix. For example, $I_{\overline{W}}$ is an identity matrix of size $\overline{W} = W + 1$;

(ii) $\otimes$ indicate the Kronecker product of matrices; see [12];

(iii) $C_l = \text{diag}\{0, 1, \ldots, l\}$, $\overline{C}_l = \text{diag}\{l, l-1, \ldots, 0\}$, $l = \overline{0,M}$,

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

(iv) $\text{diag}\{A_1, \ldots, A_l\}$ a block-diagonal matrix with the diagonal blocks $A_1, \ldots, A_l$;

(v) $E_l^+, \widehat{E}_l^+$, $l = \overline{0, M-1}$, the matrices of size $(l+1) \times (l+2)$, defined as

$$\begin{gathered}
E_l^+ = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & 1 \end{pmatrix}; \\
\widehat{E}_l^+ = \begin{pmatrix} 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{pmatrix};
\end{gathered} \tag{3}$$

(vi) $E_l^-, \widehat{E}_l^-$, $l = \overline{1,M}$, the matrices of size $(l+1) \times l$, defined as

$$\begin{gathered}
E_l^- = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \\ 0 & 0 & \cdots & 0 \end{pmatrix}, \\
\widehat{E}_l^- = \begin{pmatrix} 0 & 0 & 0 & \cdots & 0 \\ 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix};
\end{gathered} \tag{4}$$

(vii) $\widetilde{I}$ the diagonal matrix of size $(M+1)(N+1-M/2)$ with the diagonal entries defined as follows: $\{\underbrace{0, \ldots, 0}_{(M+1)M/2}, 1, \ldots, 1\}$;

(viii) $E^-, \widehat{I}$ the square matrices of size $M+1$ defined by formulas:

$$\begin{gathered}
E^- = \begin{pmatrix} 0 & 0 & 0 & \cdots & 0 & 0 \\ 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 1 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & 0 \end{pmatrix}, \\
\widehat{I} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{pmatrix}.
\end{gathered} \tag{5}$$

Lemma 1. *Generator Q of the Markov chain ξ_t has the following block-tridiagonal structure:*

$$Q = \begin{pmatrix} Q_{0,0} & Q_{0,1} & O & O & \cdots \\ Q_{1,0} & Q_{1,1} & Q_{1,2} & O & \cdots \\ O & Q_{2,1} & Q_{2,2} & Q_{2,3} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}, \tag{6}$$

where nonzero blocks $Q_{i,j}$, $i, j \geq 0$, *are defined as follows:*

$$
\begin{aligned}
Q_{i,i} &= \begin{pmatrix} A_i^{(0)} & B_i^{(0)} & O & \cdots & O & O \\ F^{(1)} & A_i^{(1)} & B_i^{(1)} & \cdots & O & O \\ \vdots & \vdots & \ddots & \ddots & \vdots & \vdots \\ O & O & O & \ddots & A_i^{(N-1)} & B_i^{(N-1)} \\ O & O & O & \cdots & F^{(N)} & A_i^{(N)} \end{pmatrix} \\
&\quad + (1-q)\, \widetilde{I}_{(M+1)(N-M/2+1)} \otimes D_2 \\
&\quad + I_{(M+1)(N-M/2+1)} \otimes D_0, \qquad i \geq 0, \\
Q_{i,i+1} &= Q^{+} = \operatorname{diag}\left\{H^{(0)}, \ldots, H^{(N)}\right\}, \qquad i \geq 0, \\
Q_{i,i-1} &= \begin{pmatrix} L_i^{(0)} & \widetilde{B}_i^{(0)} & O & \cdots & O & O \\ O & L_i^{(1)} & \widetilde{B}_i^{(1)} & \cdots & O & O \\ O & O & L_i^{(2)} & \cdots & O & O \\ \vdots & \vdots & \ddots & \ddots & \vdots & \vdots \\ O & O & O & \ddots & L_i^{(N-1)} & \widetilde{B}_i^{(N-1)} \\ O & O & O & \cdots & O & L_i^{(N)} \end{pmatrix}, \\
&\qquad\qquad i \geq 1,
\end{aligned}
\tag{7}
$$

where

$$
\begin{aligned}
&A_i^{(n)} \\
&= \begin{cases} -\left(\mu_2 C_n + \mu_1 \overline{C}_n \right. \\ \qquad \left. + i(\alpha+\gamma) I_{n+1}\right) \otimes I_{\overline{W}}, & n < M, i \geq 0, \\ -\left(\mu_2 C_M + \mu_1 \widetilde{C}_n \right. \\ \qquad \left. + i((1-q)\alpha+\gamma) I_{M+1}\right) \otimes I_{\overline{W}}, & M \leq n < N, i \geq 0, \\ -\left(\mu_2 C_M + \mu_1 \widetilde{C}_n \right. \\ \qquad \left. + i((1-q)\alpha+\gamma) I_{M+1}\right) \\ \qquad \otimes I_{\overline{W}} + (1-p) E^{-} \otimes D_1 \\ \qquad + \widehat{I} \otimes D_1, & n = N, i \geq 0; \end{cases} \\
&B^{(n)} = \begin{cases} E_n^{+} \otimes D_2 + \widehat{E}_n^{+} \otimes D_1, & n < M, \\ I_{M+1} \otimes D_1, & M \leq n < N; \end{cases} \\
&F^{(n)} = \begin{cases} \left(\mu_2 C_n \widehat{E}_n^{-} + \mu_1 \overline{C}_n E_n^{-}\right) \otimes I_{\overline{W}}, & n \leq M, \\ \left(\mu_2 C_M E^{-} + \mu_1 \widetilde{C}_n\right) \otimes I_{\overline{W}}, & M < n \leq N; \end{cases} \\
&H^{(n)} = \begin{cases} O, & n < M, \\ q I_{M+1} \otimes D_2, & M \leq n < N, \\ p E^{-} \otimes D_1 + q I_{M+1} \otimes D_2, & n = N; \end{cases} \\
&\widetilde{B}_i^{(n)} = \begin{cases} i\alpha E_n^{+} \otimes I_{\overline{W}}, & n < M, i \geq 0, \\ O, & n \geq M, i \geq 0; \end{cases} \\
&L_i^{(n)} = \begin{cases} i\gamma I_{n+1} \otimes I_{\overline{W}}, & n < M, \\ i(\gamma + (1-q)\alpha) I_{M+1} \otimes I_{\overline{W}}, & M \leq n \leq N. \end{cases}
\end{aligned}
\tag{8}
$$

Proof of the lemma is implemented by careful analysis of the intensities of the transitions of the Markov chain ξ_t during an interval of time having an infinitesimal length and is omitted here.

Remark 2. It can be verified that the following limits exist:

$$
\begin{gathered}
Y_0 = \lim_{i \to \infty} R_i^{-1} Q_{i,i-1}, \qquad Y_1 = \lim_{i \to \infty} R_i^{-1} Q_{i,i} + I, \\
Y_2 = \lim_{i \to \infty} R_i^{-1} Q_{i,i+1},
\end{gathered}
\tag{9}
$$

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

It is easy to check that here the matrix R_i is the block-diagonal matrix with the diagonal blocks $T_i^{(n)}$, $n \in \{0, \ldots, N\}$, $i \geq 0$, defined as follows:

$$
T_i^{(n)} = \begin{cases} -A_i^{(n)} + I_{n+1} \otimes \Lambda, & n = \overline{0, M-1}, \\ -A_i^{(n)} - (1-q) I_{M+1} \\ \quad \otimes \Sigma_2 + I_{M+1} \otimes \Lambda, & n = \overline{M, N-1}, \\ \left(\mu_2 C_M + \mu_1 \widetilde{C}_N + i((1-q)\alpha + \gamma)\right) \\ \quad \otimes I_{\overline{W}} - \widehat{I} \otimes \Sigma_1 - (1-q) I_{M+1} \\ \quad \otimes \Sigma_2 + I_{M+1} \otimes \Lambda, & n = N, \end{cases}
\tag{10}
$$

where Σ_1, Σ_2, and Λ are the diagonal matrices, the diagonal entries of which are defined as the corresponding diagonal entries of the matrices D_1, D_2, and $-D_0$, respectively.

Existence of the limits Y_k, $k = 0, 1, 2$, implies that the Markov chain ξ_t, $t \geq 0$, belongs to the class of continuous-time asymptotically quasi-Toeplitz Markov chains (AQTMC); see [13]. So, results from [13] can be used to derive the ergodicity condition for the Markov chain ξ_t and compute its stationary distribution.

4. Ergodicity Condition and Stationary Distribution of the Markov Chain

Theorem 3. *If* $q \neq 1$ *or* $\gamma \neq 0$, *then the Markov chain* ξ_t *is ergodic for any set of parameters of the queueing system under study.*

If $q = 1$ *and* $\gamma = 0$, *then the Markov chain* ξ_t *is ergodic if the following condition is fulfilled:*

$$
\mathbf{x}_M \left(\mu_2 C_M + \mu_1 \overline{C}_M\right) \mathbf{e} > \lambda_2 + p \lambda_1 \mathbf{x}_N \widehat{\mathbf{e}},
\tag{11}
$$

where $\widehat{\mathbf{e}}$ is the column vector of size $M+1$ having first zero component and other components equal to 1 and the vector $\mathbf{x} = (\mathbf{x}_M, \ldots, \mathbf{x}_N)$, $\mathbf{x}_n = (\mathbf{x}(n,0), \ldots, \mathbf{x}(n,M))$, $n = \overline{M,N}$, is the unique solution to the system

$$\mathbf{x}A = \mathbf{0}, \qquad \mathbf{x}\mathbf{e} = 1. \tag{12}$$

Here the matrix A is defined by formulas

$$A = \begin{pmatrix} \widetilde{A}^{(M)} & \lambda_1 I_{M+1} & O & \cdots & O & O & O \\ \widetilde{F}^{(M+1)} & \widetilde{A}^{(M+1)} & \lambda_1 I_{M+1} & \cdots & O & O & O \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ O & O & O & \cdots & \widetilde{F}^{(N-1)} & \widetilde{A}^{(N-1)} & \lambda_1 I_{M+1} \\ O & O & O & \cdots & O & \widetilde{F}^{(N)} & \widetilde{A}^{(N)} \end{pmatrix}, \tag{13}$$

$$\widetilde{A}^{(n)} = \begin{cases} -\mu_1 \widetilde{C}_M \left(I - E_M^- E_{M-1}^+\right) - \lambda_1 I_{M+1}, & n = M < N, \\ -\mu_1 \widetilde{C}_M \left(I - E_M^- E_{M-1}^+\right) & \\ \quad -\lambda_1 \left(I - E^- - \widehat{I}\right), & n = M = N, \\ -\mu_2 C_M - \mu_1 \widetilde{C}_n - \lambda_1 I_{M+1}, & M < n < N, \\ -\mu_2 C_M - \mu_1 \widetilde{C}_N - \lambda_1 \left(I - E^- - \widehat{I}\right), & n = N > M, \end{cases} \tag{14}$$

and $\widetilde{F}^{(n)} = (\mu_2 C_M E^- + \mu_1 \widetilde{C}_n)$, $n = \overline{M+1,N}$.

Proof. It follows from [13], that sufficient condition for ergodicity of the AQTMC ξ_t, $t \geq 0$, is the fulfillment of the inequality

$$\mathbf{y}Y_0\mathbf{e} > \mathbf{y}Y_2\mathbf{e}, \tag{15}$$

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

$$\mathbf{y}\left(Y_0 + Y_1 + Y_2\right) = \mathbf{y}, \qquad \mathbf{y}\mathbf{e} = 1. \tag{16}$$

Let us separately consider two cases. Let first $q \neq 1$ or $\gamma \neq 0$. In this case, it can be verified that the matrices Y_0, Y_1, and Y_2 are defined by expressions

$$Y_0 = \begin{pmatrix} \frac{\gamma}{\gamma+\alpha} I_{\overline{W}} & \frac{\alpha}{\gamma+\alpha} E_0^+ \otimes I_{\overline{W}} & \cdots & O & O & O & \cdots & O \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ O & O & \cdots & \frac{\gamma}{\gamma+\alpha} I_{M\overline{W}} & \frac{\alpha}{\gamma+\alpha} E_M^+ \otimes I_{\overline{W}} & O & \cdots & O \\ O & O & \cdots & O & I_{(M+1)\overline{W}} & O & \cdots & O \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ O & O & \cdots & O & O & O & \cdots & O \\ O & O & \cdots & O & O & O & \cdots & I_{(M+1)\overline{W}} \end{pmatrix}, \tag{17}$$

$$Y_1 = O, \qquad Y_2 = O.$$

It is evident that in this case inequality (15) and system (16) can be rewritten as $\mathbf{y}Y_0\mathbf{e} > 0$, $\mathbf{y}Y_0 = \mathbf{y}$, $\mathbf{y}\mathbf{e} = 1$. So inequality (15) is equivalent to the inequality $\mathbf{y}Y_0\mathbf{e} = \mathbf{y}\mathbf{e} = 1 > 0$ that trivially holds true for all values of the parameters of the system under study.

Let now $q = 1$ and $\gamma = 0$, then the matrices Y_0, Y_1, and Y_2 are defined by expressions

$$Y_0 = T^{-1} \begin{pmatrix} O & E_0^+ \otimes I_{\overline{W}} & O & \cdots & O & \cdots & O \\ \vdots & \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ O & O & O & \cdots & E_{M-1}^+ \otimes I_{\overline{W}} & \cdots & O \\ O & O & O & \cdots & O & \cdots & O \\ \vdots & \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ O & O & O & \cdots & O & \cdots & O \end{pmatrix};$$

$$Y_1 = \widetilde{I} \otimes I_{\overline{W}} + T^{-1} \begin{pmatrix} O & \cdots & O & O & O & \cdots & O & O \\ \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ O & \cdots & O & O & O & \cdots & O & O \\ O & \cdots & F^{(M)} & A^{(M)} & B^{(M)} & \cdots & O & O \\ O & \cdots & O & F^{(M+1)} & A^{(M+1)} & \cdots & O & O \\ \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ O & \cdots & O & O & O & \cdots & A^{(N-1)} & B^{(N-1)} \\ O & \cdots & O & O & O & \cdots & F^{(N)} & A^{(N)} \end{pmatrix},$$

$$Y_2 = T^{-1} \operatorname{diag}\left\{O_{\overline{W}}, O_{2\overline{W}}, \ldots, O_{M\overline{W}}, H^{(M)}, H^{(M+1)}, \ldots, H^{(N)}\right\}, \tag{18}$$

where

$$T = \operatorname{diag}\left\{I_{\overline{W}}, I_{2\overline{W}}, \ldots, I_{M\overline{W}}, T^{(M)}, T^{(M+1)}, \ldots, T^{(N)}\right\},$$

$$T^{(n)} = \begin{cases} \left(\mu_2 C_M + \mu_1 \widetilde{C}_n\right) \otimes I_{\overline{W}} + I_{M+1} \otimes \Lambda, & M \le n < N, \\ \left(\mu_2 C_M + \mu_1 \widetilde{C}_N\right) \otimes I_{\overline{W}} \\ \quad -\widehat{I} \otimes \Sigma_1 + I_{M+1} \otimes \Lambda, & n = N, \end{cases}$$

$$A^{(n)} = \begin{cases} -\left(\mu_2 C_M + \mu_1 \widetilde{C}_n\right) \otimes I_{\overline{W}} \\ \quad + I_{M+1} \otimes D_0, & M \le n < N, \\ -\left(\mu_2 C_M + \mu_1 \widetilde{C}_n\right) \otimes I_{\overline{W}} + \widehat{I} \otimes D_1 \\ \quad + (1-p) E^{-} \otimes D_1 + I_{M+1} \otimes D_0, & n = N. \end{cases} \tag{19}$$

Matrix $Y = Y_0 + Y_1 + Y_2$ has a form

$$Y = \widetilde{I} \otimes I_{\overline{W}} + T^{-1} G, \tag{20}$$

where the matrix G has a structure

$$G = \begin{pmatrix} G_{11} & G_{12} \\ G_{21} & G_{22} \end{pmatrix}, \tag{21}$$

where

$$G_{11} = \begin{pmatrix} O & E_0^{+} \otimes I_{\overline{W}} & O & \cdots & O \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ O & O & O & \cdots & E_{M-2}^{+} \otimes I_{\overline{W}} \\ O & O & O & \cdots & O \end{pmatrix},$$

$$G_{12} = \begin{pmatrix} O & O & \cdots & O \\ \vdots & \vdots & \ddots & \vdots \\ E_{M-1}^{+} \otimes I_{\overline{W}} & O & \cdots & O \end{pmatrix},$$

$$G_{21} = \begin{pmatrix} O & \cdots & O & F^{(M)} \\ O & \cdots & O & O \\ \vdots & \ddots & \vdots & \vdots \\ O & \cdots & O & O \end{pmatrix},$$

$$G_{22} = \begin{pmatrix} A^{(M)} + H^{(M)} & B^{(M)} & O & \cdots & O & O \\ F^{(M+1)} & A^{(M+1)} + H^{(M+1)} & B^{(M+1)} & \cdots & O & O \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ O & O & O & \cdots & A^{(N-1)} + H^{(N-1)} & B^{(N-1)} \\ O & O & O & \cdots & F^{(N)} & A^{(N)} + H^{(N)} \end{pmatrix}. \tag{22}$$

Let us represent the solution $\mathbf{y}$ of system (16) as $\mathbf{y} = (\mathbf{y}_0, \mathbf{y}_1, \ldots, \mathbf{y}_N)$. Taking into account the structure of the matrix G, it is easy to see that $\mathbf{y}_0 = \mathbf{y}_1 = \ldots = \mathbf{y}_{M-2} = \mathbf{0}$. So, the vector $\mathbf{y}$ is defined as

$$\mathbf{y} = \left(\mathbf{0}, \ldots, \mathbf{0}, \mathbf{y}_{M-1}, \mathbf{z}\right), \tag{23}$$

where $\mathbf{z} = (\mathbf{y}_M, \ldots, \mathbf{y}_N)$.

Having in mind this form of the vector $\mathbf{y}$, system (16) can be rewritten in the form

$$\begin{gathered} \mathbf{y}_{M-1} = \mathbf{z}\widetilde{T}^{-1}F, \\ \mathbf{z}Z = \mathbf{0}, \qquad \mathbf{z}\mathbf{e} + \mathbf{y}_{M-1}\mathbf{e} = 1, \end{gathered} \tag{24}$$

where

$$\begin{gathered} F = \Bigl(F^{(M)}, \underbrace{O_{(M+1)\overline{W}}, \ldots, O_{(M+1)\overline{W}}}_{N-M}\Bigr)^T, \\ Z = \widetilde{T}^{-1} \times \left(G_{22} + \hat{I} \otimes F^{(M)} E^{+}_{M-1}\right) \\ \widetilde{T} = \operatorname{diag}\left\{T^{(M)}, \ldots, T^{(N)}\right\}. \end{gathered} \tag{25}$$

Let us analyze equation $\mathbf{z}Z = \mathbf{0}$ taking into account more explicit forms of the matrix Z:

$$\begin{aligned} \mathbf{0} &= \mathbf{z}Z \\ &= \mathbf{z}\widetilde{T}^{-1} \times \left[\begin{pmatrix} -\mu_1\widetilde{C}_M\left(I - \widetilde{E}\right) & O & \cdots & O & O \\ \widetilde{F}^{(M+1)} & -\mu_2 C_M - \mu_1\widetilde{C}_{M+1} & \cdots & O & O \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ O & O & \cdots & -\mu_2 C_M - \mu_1\widetilde{C}_{N-1} & O \\ O & O & \cdots & \widetilde{F}^{(N)} & -\mu_2 C_M - \mu_1\widetilde{C}_N \end{pmatrix} \right. \\ &\quad \left. \otimes I_{\overline{W}} + \begin{pmatrix} I_{M+1} \otimes \left(D_0 + D_2\right) & I_{M+1} \otimes D_1 & \cdots & O & O \\ O & I \otimes \left(D_0 + D_2\right) & \cdots & O & O \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ O & O & \cdots & I \otimes \left(D_0 + D_2\right) & I \otimes D_1 \\ O & O & \cdots & O & I \otimes \left(D_0 + D_2\right) + \left(\hat{I} + E^{-}\right) \otimes D_1 \end{pmatrix} \right], \end{aligned} \tag{26}$$

where $\widetilde{E} = E^{-}_{M} E^{+}_{M-1}$.

By postmultiplying (26) by $\mathbf{e}_{(N-M+1)(M+1)} \otimes I_{\overline{W}}$ we obtain the following equation:

$$\mathbf{z}\widetilde{T}^{-1}\left(\mathbf{e}_{(N-M+1)(M+1)} \otimes \left(D_0 + D_1 + D_2\right)\right) = \mathbf{0}. \tag{27}$$

It follows from (27) that the vector $\mathbf{z}\widetilde{T}^{-1}$ can be represented in the form

$$\mathbf{z}\widetilde{T}^{-1} = \boldsymbol{\phi} \otimes \boldsymbol{\theta}, \tag{28}$$

where $\boldsymbol{\phi}$ is some row vector of size $(N - M + 1)(M + 1)$ and $\boldsymbol{\theta}$ is the invariant probability vector of the underlying Markov chain of the MMAP.

By substituting the vector $\mathbf{z}\widetilde{T}^{-1}$ in form (28) into (26), postmultiplying this equation by $I_{(N-M+1)(M+1)} \otimes \mathbf{e}_{\overline{W}}$ and taking into account relations $\boldsymbol{\theta}D_1\mathbf{e} = \lambda_1$, $\boldsymbol{\theta}(D_0 + D_2)\mathbf{e} = \boldsymbol{\theta}(-D_1)\mathbf{e} = -\lambda_1$, and $\boldsymbol{\theta}\mathbf{e} = 1$, we conclude that the unknown vector $\boldsymbol{\phi}$ is the solution of the system

$$\boldsymbol{\phi}A = \mathbf{0}, \tag{29}$$

where the matrix A is given in the statement of Theorem 3 under proof.

It can be verified that the matrix A is the generator of two-dimensional Markov chain $\{n_t, l_t\}$ 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, $l_t = \overline{0, M}$, in the situation when the system is overloaded; that is, the number of customers in orbit is huge. It follows from (29) that the vector $\boldsymbol{\phi}$ defines, up to the normalizing factor c, the joint stationary distribution of the Markov chain $\{n_t, l_t\}$. So, the vector $\boldsymbol{\phi}$ can be represented in the form

$$\boldsymbol{\phi} = c\mathbf{x}, \tag{30}$$

where the vector $\mathbf{x}$ is given in the statement of Theorem 3.

Thus, the vector $\mathbf{z}\widetilde{T}^{-1}$ is defined by

$$\mathbf{z}\widetilde{T}^{-1} = c\mathbf{x} \otimes \boldsymbol{\theta} \tag{31}$$

and, using the so called mixed product rule for Kronecker product of matrices, see [12]; the left-hand side of inequality (15) can be rewritten as

$$\begin{aligned} \mathbf{y}Y_0\mathbf{e} &= \mathbf{y}_{M-1}\left(E^{+}_{M} \otimes I_{\overline{W}}\right)\mathbf{e} \\ &= \mathbf{z}\widetilde{T}^{-1}F\left(E^{+}_{M} \otimes I_{\overline{W}}\right)\mathbf{e} \\ &= c\mathbf{x} \otimes \boldsymbol{\theta}F\mathbf{e} \\ &= c\left(\mathbf{x}_M \otimes \boldsymbol{\theta}\right)\left(\left(\mu_2 C_M \widehat{E}^{-}_{M} + \mu_1 \overline{C}_M E^{-}_{M}\right) \otimes I_{\overline{W}}\right)\mathbf{e} \end{aligned}$$

$$\begin{aligned}&= c\left(\mathbf{x}_M\left(\mu_2 C_M \widehat{E}_M^- + \mu_1 \overline{C}_M E_M^-\right)\right) \otimes \left(\boldsymbol{\theta} I_{\overline{W}}\right)\mathbf{e}\\ &= c\left(\mathbf{x}_M\left(\mu_2 C_M \widehat{E}_M^- + \mu_1 \overline{C}_M E_M^-\right)\right)\mathbf{e} \otimes 1\\ &= c\mathbf{x}_M\left(\mu_2 C_M \widehat{E}_M^- + \mu_1 \overline{C}_M E_M^-\right)\mathbf{e}\\ &= c\mathbf{x}_M\left(\mu_2 C_M + \mu_1 \overline{C}_M\right)\mathbf{e}.\end{aligned} \tag{32}$$

The right-hand side of inequality (15) can be rewritten as follows:

$$\begin{aligned}\mathbf{y}Y_2\mathbf{e} &= \mathbf{z}\widetilde{T}^{-1}\operatorname{diag}\left\{H_M, H_{M+1}, \dots, H_N\right\}\\ &= c\left(\sum_{n=M}^{N-1}\left(\mathbf{x}_n \otimes \boldsymbol{\theta}\right)\left(I_{M+1} \otimes D_2\right)\mathbf{e}\right.\\ &\qquad \left. + \left(\mathbf{x}_N \otimes \boldsymbol{\theta}\right)\left(pE^- \otimes D_1 + I_{M+1} \otimes D_2\right)\mathbf{e}\right)\\ &= c\left(\lambda_2 \sum_{n=M}^{N}\mathbf{x}_n\mathbf{e} + p\lambda_1\mathbf{x}_N\widehat{\mathbf{e}}\right)\\ &= c\left(\lambda_2 + p\lambda_1\mathbf{x}_N\widehat{\mathbf{e}}\right).\end{aligned} \tag{33}$$

Then, inequality (15) can be rewritten as follows:

$$\mathbf{x}_M\left(\mu_2 C_M + \mu_1 \overline{C}_M\right)\mathbf{e} > \lambda_2 + p\lambda_1\mathbf{x}_N\widehat{\mathbf{e}}. \tag{34}$$

Theorem is proved. □

Remark 4. Condition (11) is intuitively clear. In the overloaded system, a customer may leave orbit only in the situation when the number of busy servers becomes less than M. The components $\mathbf{x}(M, l)$, $l = \overline{0, M}$, of the vector $\mathbf{x}_M$ define the probability that, at an arbitrary time, M servers are busy and l of them provide service to type-2 customers. So, the left-hand side of inequality (11) defines intensity of the service completions when M servers are busy. This intensity coincides with the intensity of customers' departure from orbit. The right-hand side of inequality (11) defines the intensity of type-2 customers' arrival into orbit. It is equal to the sum of the intensity λ_2 of customers' arrival from outside and the intensity $p\lambda_1\mathbf{x}_N\widehat{\mathbf{e}}$ of customers' arrival due to force termination.

Thus, ergodicity condition (11) requires that, in the situation when the system is overloaded, the intensity of customers' arrival into orbit is less than the intensity of customers' departure from orbit.

Further, we assume that condition (11) is fulfilled. Then the following limits (stationary probabilities) exist:

$$\begin{aligned}&\pi(i, n, l, \nu) = \lim_{t \to \infty} P\left\{i_t = i,\ n_t = n,\ l_t = l,\ \nu_t = \nu\right\},\\ &i \geq 0, \quad n = \overline{0, N}, \quad l = \overline{0, \min\{n, M\}}, \quad \nu = \overline{0, W}.\end{aligned} \tag{35}$$

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

$$\begin{aligned}\boldsymbol{\pi}(i, n, l) &= \left(\pi(i, n, l, 0), \pi(i, n, l, 1), \dots, \pi(i, n, l, W)\right),\\ &\qquad l = \overline{0, \min\{n, M\}},\\ \boldsymbol{\pi}(i, n) &= \left(\boldsymbol{\pi}(i, n, 0), \boldsymbol{\pi}(i, n, 1), \dots, \boldsymbol{\pi}(i, n, \min\{n, M\})\right),\\ &\qquad n = \overline{0, N},\\ \boldsymbol{\pi}_i &= \left(\boldsymbol{\pi}(i, 0), \boldsymbol{\pi}(i, 1), \dots, \boldsymbol{\pi}(i, N)\right), \quad i \geq 0.\end{aligned} \tag{36}$$

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

$$\left(\boldsymbol{\pi}_0, \boldsymbol{\pi}_1, \dots\right)Q = \mathbf{0}, \qquad \left(\boldsymbol{\pi}_0, \boldsymbol{\pi}_1, \dots\right)\mathbf{e} = 1, \tag{37}$$

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

Note that in the case $q = 0$, $p = 0$ we get the model where retrials are not taken into account. In this case, the generator Q reduces to the finite block $Q_{0,0}$ and system (37) is finite. The probability vectors can be computed by the direct solution of system (37) on computer or by means of the numerically stable algorithms developed in [14, 15]. In general case, system (37) is infinite and cannot be directly solved on computer. It can be solved by means of the numerically stable algorithm developed in [13]. The algorithm presented in [13] is oriented to more general forms of the generator Q (blocks above the off-diagonal blocks can be not equal to zero). Variant of the algorithm exactly oriented to a block-tridiagonal form of the generator Q can be found, for example, in [16].

5. Performance Measures of the System

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

The distribution of the number of the customers in orbit is

$$\lim_{t \to \infty} P\left\{i_t = i\right\} = \boldsymbol{\pi}_i\mathbf{e}, \quad i \geq 0. \tag{38}$$

The average number of customers in orbit is

$$L_{\text{orbit}} = \sum_{i=1}^{\infty} i\boldsymbol{\pi}_i\mathbf{e}. \tag{39}$$

The average number of customers in the system is

$$L = \sum_{i=0}^{\infty}\sum_{n=0}^{N}(i + n)\boldsymbol{\pi}(i, n)\mathbf{e}. \tag{40}$$

The average number of busy servers is

$$N_{\text{server}} = \sum_{i=0}^{\infty}\sum_{n=1}^{N} n\boldsymbol{\pi}(i, n)\mathbf{e}. \tag{41}$$

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

$$N_{\text{server}}^{(1)} = \sum_{i=0}^{\infty}\sum_{n=1}^{N}\sum_{l=0}^{\min\{n,M\}} (n-l)\,\boldsymbol{\pi}(i,n,l)\,\mathbf{e}. \tag{42}$$

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

$$N_{\text{server}}^{(2)} = \sum_{i=0}^{\infty}\sum_{n=1}^{N}\sum_{l=1}^{\min\{n,M\}} l\boldsymbol{\pi}(i,n,l)\,\mathbf{e} = N_{\text{server}} - N_{\text{server}}^{(1)}. \tag{43}$$

The intensity of output of type-1 customers is

$$\lambda_{\text{out}}^{(1)} = \mu_1 N_{\text{server}}^{(1)}. \tag{44}$$

The intensity of output of type-2 customers is

$$\lambda_{\text{out}}^{(2)} = \mu_2 N_{\text{server}}^{(2)}. \tag{45}$$

The intensity of output of customers from the system is

$$\lambda_{\text{out}} = \lambda_{\text{out}}^{(1)} + \lambda_{\text{out}}^{(2)}. \tag{46}$$

The blocking (loss) probability of type-1 customers is

$$P_1^{(\text{loss})} = \lambda_1^{-1}\sum_{i=0}^{\infty}\boldsymbol{\pi}(i,N,0)\,D_1\mathbf{e} = 1 - \frac{\lambda_{\text{out}}^{(1)}}{\lambda_1}. \tag{47}$$

The blocking (loss) probability of type-2 customers is

$$P_2^{(\text{loss})} = 1 - \frac{\lambda_{\text{out}}^{(2)}}{\lambda_2}. \tag{48}$$

The blocking (loss) probability of an arbitrary customer is

$$P^{(\text{loss})} = 1 - \frac{\lambda_{\text{out}}}{\lambda}. \tag{49}$$

The probability of type-2 customer loss due to the imposed restriction (type-2 customer is not granted access to the system if the number of busy servers is greater or equal to M) is

$$P^{(\text{ent-loss})} = (1-q)\,\lambda_2^{-1}\sum_{i=0}^{\infty}\sum_{n=M}^{N}\boldsymbol{\pi}(i,n)\left(I_{M+1}\otimes D_2\right)\mathbf{e}. \tag{50}$$

The probability that type-2 customer will go into orbit due to the imposed restriction is

$$P^{(\text{ent-to-orbit})} = q\lambda_2^{-1}\sum_{i=0}^{\infty}\sum_{n=M}^{N}\boldsymbol{\pi}(i,n)\left(I_{M+1}\otimes D_2\right)\mathbf{e}. \tag{51}$$

The probability that an arbitrary type-2 customer will be forced to terminate service and go into orbit is

$$P^{(\text{termination-to-orbit})} = p\lambda_2^{-1}\sum_{i=0}^{\infty}\sum_{l=1}^{M}\boldsymbol{\pi}(i,N,l)\,D_1\mathbf{e}. \tag{52}$$

The probability that an arbitrary type-2 customer will be forced to terminate service and will be lost is

$$P^{(\text{termination-loss})} = (1-p)\,\lambda_2^{-1}\sum_{i=0}^{\infty}\sum_{l=1}^{M}\boldsymbol{\pi}(i,N,l)\otimes D_1\mathbf{e}. \tag{53}$$

The probability of an arbitrary type-2 customer loss from orbit is

$$P^{(\text{loss-from-orbit})} = P_2^{(\text{loss})} - P^{(\text{ent-loss})} - P^{(\text{termination-loss})}. \tag{54}$$

The probability that an arbitrary customer from orbit will make an attempt to receive service when the number of busy servers is greater or equal to M and return to orbit is

$$P^{(\text{return-to-orbit})} = q\tilde{\alpha}^{-1}\sum_{i=1}^{\infty}\sum_{n=M}^{N} i\alpha\boldsymbol{\pi}(i,n)\,\mathbf{e}, \tag{55}$$

where $\tilde{\alpha} = \alpha L_{\text{orbit}}$.

The probability that an arbitrary customer from orbit will make an attempt to receive service when the number of busy servers is greater or equal to M and leave the system without service is

$$P_1^{(\text{loss-from-orbit})} = (1-q)\,\tilde{\alpha}^{-1}\sum_{i=1}^{\infty}\sum_{n=M}^{N} i\alpha\boldsymbol{\pi}(i,n)\,\mathbf{e}. \tag{56}$$

6. Optimization Problem and Numerical Examples

As it was mentioned in the description of the mathematical model, our goal is to find the value M^* of the threshold M, $1 \le M \le N$, which provides the maximal value of cost criterion (1). Analytical results presented in the previous sections allow us to compute the performance measures involved in cost criterion (1) under any fixed value of M, $1 \le M \le N$. So, the problem of finding the optimal value of M in the finite set $1 \le M \le N$ can be solved.

As one of the advantages of our model comparing to other models existing in the literature, we mentioned in Section 1 that we use the MMAP instead of the stationary Poisson processes of customers. This allows taking into account possible correlation in the arrival process. To illustrate the importance of account of impact of correlation, let us consider three different arrival processes having the same intensity of arrival of each type of customers, but different coefficients of correlation of successive interarrival times in the arrival process.

For this purpose, let us introduce three MMAPs defined by the matrices D_0, $D_1^{(1)}$, and $D_1^{(2)}$. All these MMAPs have the same average total arrival rate $\lambda = 4$, the average intensity of priority customers $\lambda_1 = 4/3$, and the average intensity of nonpriority customers $\lambda_2 = 8/3$, but different coefficients of correlation. MMAP^a denotes the MMAP arrival process with coefficient of correlation $c_{\text{cor}} = a$.

The first process coded as MMAP^0 is defined by the matrices $D_0 = -4$, $D_1^{(1)} = 4/3$, and $D_1^{(2)} = 8/3$. It has the coefficient of correlation $c_{\text{cor}} = 0$ and the coefficient of variation $c_{\text{var}} = 1$. In this case, the arrival processes of priority and nonpriority customers are defined as the stationary Poisson processes.

The second process MMAP$^{0.2}$ is defined by the matrices

$$D_0 = \begin{pmatrix} -5.408 & 0 \\ 0 & -0.1755 \end{pmatrix},$$
$$D_1^{(1)} = \begin{pmatrix} 1.7906 & 0.012 \\ 0.03257 & 0.02593 \end{pmatrix}, \quad (57)$$
$$D_1^{(2)} = \begin{pmatrix} 3.5814 & 0.024 \\ 0.06515 & 0.05185 \end{pmatrix}$$

and has the coefficient of correlation $c_{\text{cor}} = 0.2$, and the coefficient of variation $c_{\text{var}} = 12.35$.

The third process MMAP$^{0.4}$ is defined by the matrices

$$D_0 = \begin{pmatrix} -13.775 & 0.081 \\ 0.004 & -0.444 \end{pmatrix},$$
$$D_1^{(1)} = \begin{pmatrix} 4.544 & 0.021 \\ 0.016 & 0.1305 \end{pmatrix}, \quad (58)$$
$$D_1^{(2)} = \begin{pmatrix} 9.088 & 0.041 \\ 0.0325 & 0.261 \end{pmatrix}.$$

It has the coefficient of correlation $c_{\text{cor}} = 0.4$, and the coefficient of variation $c_{\text{var}} = 12.35$.

The rest of the parameters of the queueing model are assumed to be as follows:

(i) the number of servers is $N = 30$;

(ii) the service intensity of type-1 customers is $\mu_1 = 0.08$;

(iii) the service intensity of type-2 customers is $\mu_2 = 0.2$;

(iv) the intensity of impatience of customers from orbit is $\gamma = 0.005$;

(v) the intensity of retrials is $\alpha = 0.15$;

(vi) the probabilities q and p are equal to 0.9 and 0.1, correspondingly.

Let us vary the threshold M in the interval $[1, N]$, compute, and analyze dynamics of key performance measures of the system.

First of all, it is worth to note the evident fact that, because type-1 customers have preemptive priority, the probability $P_1^{(\text{loss})}$ of type-1 customer loss and the average number $N_{\text{server}}^{(1)}$ of busy servers providing service to type-1 customers do not change when the threshold M varies. The values of these performance measures for arrival flows with the same mean arrival rate but different coefficient of correlation are given in Table 1.

Table 1: Probability $P_1^{(\text{loss})}$ of type-1 customer loss and the average number $N_{\text{server}}^{(1)}$ of busy servers providing service to type-1 customers for various correlations in the arrival process.

c_{cor}	$P_1^{(\text{loss})}$	$N_{\text{server}}^{(1)}$
0	0.000986	16.65
0.2	0.011974	16.4337
0.4	0.205498	13.2389

It is evidently seen from this table that the correlation in the arrival process may drastically change the performance measures of the queueing model. If an arrival process in some real life system is correlated while one will try to model the arrival process by the stationary Poisson process, he or she will get too optimistic forecasting of the system performance. In our example, for the stationary Poisson process the predicted value of the probability $P_1^{(\text{loss})}$ of type-1 customer loss is less than 10^{-3}. But if the correlation in the real arrival process is equal to 0.2, the probability $P_1^{(\text{loss})}$ is more than 10 times higher and, if correlation in the arrival process is equal to 0.4, the probability $P_1^{(\text{loss})}$ is more than 200 times higher.

Note that this result does not sound very surprising because service of type-1 customers may be described by Erlang loss model $MAP/M/N/N$ and analogous effect for this model was previously reported in [14]. Intuitive explanation of this effect is the following. Arrivals of customers in the stationary Poisson process are more or less uniformly distributed in time and the servers of corresponding queueing system are loaded more or less uniformly. Positive correlation in the arrival process implies that customers arrive rarely during some time intervals, while a lot of customers arrive in other intervals. It is said that such a process is "bursty." This nonuniform arrival of customers implies that during some time intervals the system starves, many servers are idle, while during some other intervals a lot of customers is lost due to the system overflow.

Concerning the behavior of the majority of characteristics of processing type-2 customers when the threshold M varies, it is quite predictable. When the number M increases, the probabilities $P^{(\text{ent-loss})}$ and $P^{(\text{ent-to-orbit})}$ decrease, while the probabilities $P^{(\text{termination-loss})}$ and $P^{(\text{termination-to-orbit})}$ increase. Because the probability $P_2^{(\text{loss})}$ of type-2 customer loss is the sum of the probabilities $P^{(\text{ent-loss})}$, $P^{(\text{termination-loss})}$, and $P^{(\text{loss-from-orbit})}$, some of them having an opposite dynamics, behavior of the probability $P_2^{(\text{loss})}$ is more complicated. This probability decreases with increase of M until the threshold reaches some critical value. Then the probability $P_2^{(\text{loss})}$ starts sharply increasing. Correspondingly, the intensity $\lambda_{\text{out}}^{(2)}$ of the flow of type-2 customers that received successful service in the system, which is the most important performance measure of the system, has maximum at some point inside the region $[1, N]$.

The cost criterion (1), which is the weighted sum of $\lambda_{\text{out}}^{(2)}$ and loss probabilities, also reaches the maximum at some point. To demonstrate this, let us fix the following values of the cost coefficients $a = 20$, $c_1 = 3$, $c_2 = 1$, $c_3 = 20$, $c_4 = 10$, and $c_5 = 3$. Dynamics of cost criterion (1) for three MMAPs, defined above, having the same mean arrival rate but different correlation is presented in Figure 1.

In Table 2, we present the optimal values M^* of the threshold and cost criterion $J(M^*)$, the value of cost criterion for the system without admission control $J(N)$, the absolute value of the profit gained by control $J(M^*) - J(N)$, and the relative value of the gain $(J(M^*) - J(N))/J(N) \times 100\%$ for various correlations in the arrival process.

Table 2: Information about the optimal values of the threshold, cost criterion, and profit in comparison to the system without admission control for various correlation in the arrival process.

c_{cor}	M^*	$J(M^*)$	$J(N)$	$J(M^*) - J(N)$	$\dfrac{J(M^*) - J(N)}{J(N)} \times 100\%$
0	29	38.5723	35.1728	3.3995	9.66
0.2	28	27.9364	19.9398	7.9986	40.10
0.4	27	20.8424	12.2544	8.5880	70.08

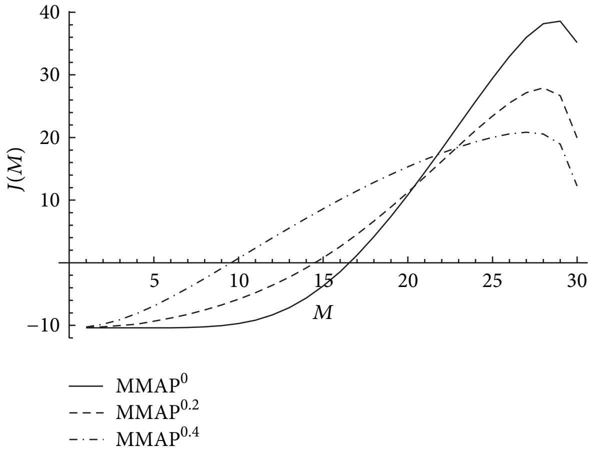


Figure 1: Dynamics of the cost criterion $J(M)$ for three MMAPs arrival processes with different coefficients of correlation.

It follows from this table that higher correlation in arrival process implies necessity of more strict restriction of access of secondary customers and higher profit obtained via the optimal control by admission of secondary customers.

7. Conclusion

We analyzed the multiserver queueing model of the MMAP/$M_2/N/N$ type suitable for modeling systems of cognitive radio. Primary customers have preemptive priority. Access of the secondary customers is restricted via threshold mechanism aiming to provide maximally effective processing of secondary customers. The secondary customers have the option to retry for service later in the case of access deny. Under the fixed value of the threshold, behavior of the queueing system is described by the level dependent multidimensional Markov chain. Sufficient condition for ergodicity of this chain in simple analytically tractable form is derived. The expressions for the main performance measures of the system are derived. Optimization problem is considered. Provided results of the numerical experiments illustrate high effectiveness of the used strategy of the restriction of access of the secondary customers and necessity of careful account correlation in arrival process.

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. 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.

[2] 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.

[3] I. F. Akyildiz, W.-Y. 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.

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

[5] 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.

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

[7] 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.

[8] D. P. Heyman and D. Lucantoni, “Modeling multiple IP traffic streams with rate limits,” *IEEE/ACM Transactions on Networking*, vol. 11, no. 6, pp. 948–958, 2003.

[9] 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.

[10] A. Gómez-Corral, “A bibliographical guide to the analysis of retrial queues through matrix analytic techniques,” *Annals of Operations Research*, vol. 141, pp. 163–191, 2006.

[11] 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.

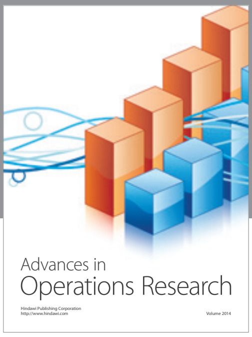
[12] A. Graham, *Kronecker Products and Matrix Calculus with Applications*, Ellis Horwood, Chichester, UK, 1981.

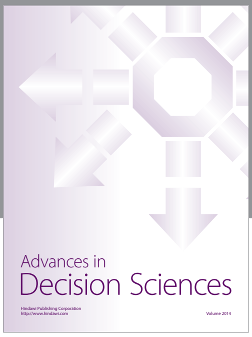
[13] V. Klimenok and A. Dudin, “Multi-dimensional asymptotically quasi-Toeplitz Markov chains and their application in queueing theory,” *Queueing Systems*, vol. 54, no. 4, pp. 245–259, 2006.

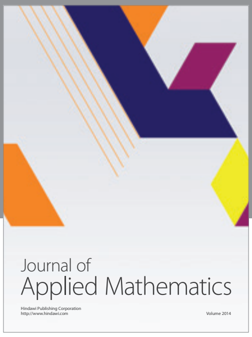
[14] V. Klimenok, C. S. Kim, D. Orlovsky, and A. Dudin, “Lack of invariant property of the Erlang loss model in case of *MAP* input,” *Queueing Systems*, vol. 49, no. 2, pp. 187–213, 2005.

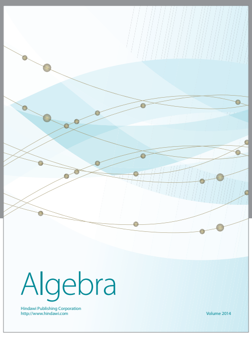
[15] C. Kim, S. Dudin, O. Taramin, and J. Baek, “Queueing system $MAP|PH|N|N+R$ with impatient heterogeneous customers as a model of call center,” *Applied Mathematical Modelling*, vol. 37, no. 3, pp. 958–976, 2013.

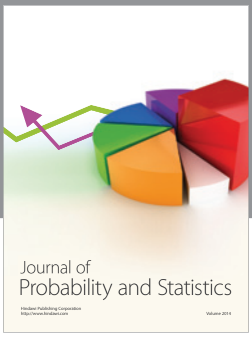
[16] S. A. Dudin, “The $MAP/M/N$ retrial queueing system with time-phased batch arrivals,” *Problems of Information Transmission*, vol. 45, no. 3, pp. 270–281, 2009.


Advances in
Operations Research

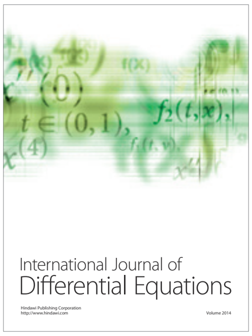

Advances in
Decision Sciences


Journal of
Applied Mathematics

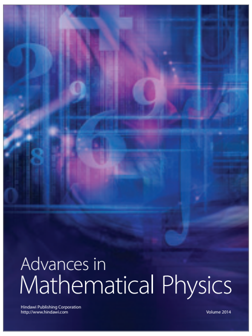

Algebra

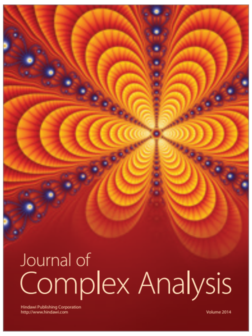

Journal of
Probability and Statistics

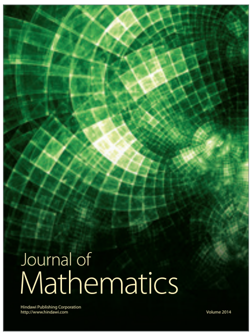
The Scientific
World Journal

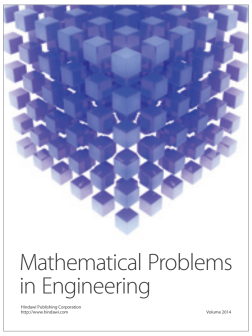

International Journal of
Differential Equations

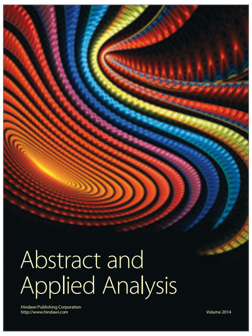
International Journal of
Combinatorics

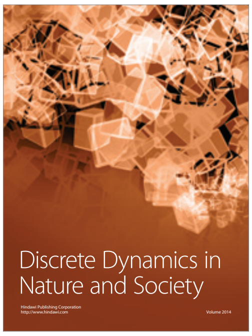

Advances in
Mathematical Physics

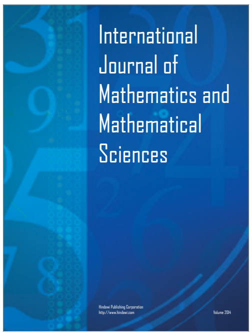

Journal of
Complex Analysis


Journal of
Mathematics


Mathematical Problems
in Engineering

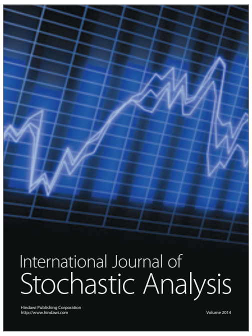

Abstract and
Applied Analysis


Discrete Dynamics in
Nature and Society


International
Journal of
Mathematics and
Mathematical
Sciences

Journal of
Discrete Mathematics

Journal of
Function Spaces


International Journal of
Stochastic Analysis

Journal of
Optimization