



Queueing System with Two Types of Customers and Limited Processor Sharing

Valentina Klimenok^(✉) , Alexander Dudin^{}, and Viktor Boksha

Department of Applied Mathematics and Computer Science,
Belarusian State University, 220030 Minsk, Belarus
{klimenok,dudin,fpm.boksha}@bsu.by

Abstract. In this paper a processor-sharing queueing system is investigated. Two types of customers enter the system according a marked Markovian arrival process. It is assumed that the number of customers of each type simultaneously being serviced is limited. The service times of customers have a phase type distribution the parameters of which depend both on the type of a customer and on the number of customers of this type in the system. The operation of the system is described in terms of a multi-dimensional Markov chain. We calculate the stationary probabilities, the main performance characteristics of the system and derive the Laplace–Stieltjes transform of the sojourn time distribution. We also present illustrative numerical examples to show the behavior of the performance measures of the system and to solve numerically an optimization problem.

Keywords: Processor sharing · Marked Markovian arrival process · Phase type distribution · Stationary distribution · Performance measures · Sojourn time

1 Introduction

Processor sharing technology is very popular in computer systems and telecommunications networks. It can be found a number of examples of real processor sharing systems and their mathematical models in the literature, see, e.g. the papers [1–8]. Most often, it is assumed that the processor can be used by an unlimited number of users, the input flow is stationary Poisson, and the service times are distributed exponentially. More general systems have been considered in the papers [9, 10] where it was assumed that customers arrive into the system according to Markovian arrival process (*MAP*) and service times have a phase type distribution. In these papers, homogeneous traffic is assumed, which is not always suitable for describing next-generation wireless communication networks, implying, in particular, the use of the Internet of Things and the presence of interaction between H2H users and M2M devices, see, e.g. [6–8]. The presence of heterogeneous requests gives rise to the need to develop new mechanisms

to maintain the specified quality of service parameters for both H2H users and M2M devices. At the same time, with an increase in the intensity of the proposed load, the planners at the base station of the LTE network must determine the optimal strategy for the allocation of radio resources based on the established restrictions, for example, the probability of loss of requests from H2H users and the average transmission time of data blocks from M2M devices.

The queueing system considered in this paper significantly expands the capabilities of modeling real systems with processor sharing. We believe that there are restrictions on the number of users of different types simultaneously in service, and we do not introduce restrictive assumptions such as the homogeneity and uncorrelated nature of the customers flow, as well as the exponential distribution of service times for customers of different types. We assume that the input flow to the system is correlated and described by the marked Markov arrival process (*MMAP*) introduced in the paper [11]. For a more adequate description of the service process, we use a phase type distribution (*PH*) which is successfully used to approximate an arbitrary distribution.

Thus, in this paper we consider a queueing system with processor sharing which receives two types of customers arriving according to a *MMAP*. If at the moment of a customer arrival the number of customers of this type on the server is greater than a predetermined threshold, then the customer leaves the system un-handled, it is considered lost. Otherwise, the customer takes up part of the throughput of the channel and is serviced for a period of time having a *PH* distribution, the parameters of which differ for customers of different types.

2 Mathematical Model

We consider a queueing system with two type of customers and processor sharing. Customers of different types arrive into the system according to the *MMAP* under control of the irreducible Markov chain ν_t , $t \geq 0$, which takes values in the set $\{0, 1, 2, \dots, W\}$ and is called as an underlying process of the *MMAP*. The transitions of the underlying process accompanied by an arrival of a customer of type k are stored as entries of the matrix D_k , $k = 1, 2$, of order $\bar{W} \times \bar{W}$ where $\bar{W} = W + 1$ and idle transitions of this process are described by the matrix D_0 .

The arrival rate of customers of type k in the *MMAP* is given by $\lambda_k = \boldsymbol{\theta} D_k \mathbf{e}$, $k = 1, 2$, where the vector $\boldsymbol{\theta}$, is defined as the unique solution of the system $\boldsymbol{\theta} D(1) = \mathbf{0}$, $\boldsymbol{\theta} \mathbf{e} = 1$. The total arrival rate is $\lambda = \lambda_1 + \lambda_2$.

The variance of inter-arrival times of customers of type k is calculated by the formula

$$v_k = \frac{2\boldsymbol{\theta}(-D_0 - \sum_{l=1, l \neq k}^2 D_l)^{-1} \mathbf{e}}{\lambda_k} - \left(\frac{1}{\lambda_k}\right)^2, \quad k = 1, 2.$$

The coefficient of correlation of the lengths of two adjacent intervals between the arrivals of customers of type k is calculated by

$$c_{cor}^{(k)} = \left[\frac{\theta(D_0 + \sum_{l=1, l \neq k}^2 D_l)^{-1}}{\lambda_k} D_k (D_0 + \sum_{l=1, l \neq k}^2 D_l)^{-1} \mathbf{e} - \left(\frac{1}{\lambda_k} \right)^2 \right] v_k^{-1}, \quad k = 1, 2.$$

A detailed description of the *MMAP* can be found, for example, in [11].

In this paper we assume that the server can simultaneously serve up to N customers of type 1 and up to R customers of type 2. If only one customer of the k th type is serviced on the server, then its service time has the *PH* distribution given by the irreducible representation (β_k, S_k) and the underlying process $m_t^{(k)}, t \geq 0$, with the state space $\{1, \dots, M_k, M_k + 1\}$, where the state $M_k + 1$ is absorbing. The intensities of transitions to the absorbing state are determined by the column vector $S_0^{(k)} = -S_k \mathbf{e}$. The service rate of a customer of type k is calculated as $\mu_k = (\beta_k (-S_k)^{-1} \mathbf{e})^{-1}$.

The customers of each type divide the throughput of the server allocated to them equally. If the server simultaneously serves n_k customers of the k th type, then the service time of any of these customers has the *PH* distribution given by the irreducible representation $(\beta_k, \frac{1}{n_k} S_k)$ and the underlying process $m_t^{(k)}, t \geq 0$, with the state space $\{1, \dots, M_k, M_k + 1\}$, where the state $M_k + 1$ is absorbing. The intensities of transitions to the absorbing state are determined by the column vector $\frac{1}{n_k} S_0^{(k)}$.

If an incoming customer of type 1 finds $n < N$ customers on the server, then it is sent for service. In this case, the throughput of the server allocated to customers of the 1st type is divided equally between $n + 1$ customers. Otherwise, the customer leaves the system un-handled, it is considered lost. Similarly, if a customer of the 2nd type finds $r < R$ customers on the server, then it is sent for service. The throughput of the server allocated to customers of type 2 is divided equally between $r + 1$ customers. Otherwise, the customer is lost.

3 Process of the System States

The operation of the system is described by the regular irreducible Markov chain

$$\xi_t = \{n_t, r_t, \eta_t^{(1)}, \eta_t^{(2)}, \dots, \eta_t^{(M_1)}, \tau_t^{(1)}, \tau_t^{(2)}, \dots, \tau_t^{(M_2)}, \nu_t\},$$

where at the moment t

- n_t is the number of customers of type 1 on the server, $n_t = \overline{0, N}$;
- r_t is the number of customers of type 2 on the server, $r_t = \overline{0, R}$;
- $\eta_t^{(m^{(1)})}$ is the number of customers of type 1 that are served in the phase $m^{(1)}$, $\eta_t^{(m^{(1)})} = \overline{0, n_t}, m^{(1)} = \overline{1, M_1}$;
- $\tau_t^{(m^{(2)})}$ is the number of customers of type 2 that are served in the phase $m^{(2)}$, $\tau_t^{(m^{(2)})} = \overline{0, r_t}, m^{(2)} = \overline{1, M_2}$;
- ν_t is the state of underlying process of the *MMAP*, $\nu_t = \overline{0, W}$,

In the following we will also use the processes

$$\mathbf{u}_t^{(1)} = \{\eta_t^{(1)}, \eta_t^{(2)}, \dots, \eta_t^{(M_1)}\}; \quad \mathbf{u}_t^{(2)} = \{\tau_t^{(1)}, \tau_t^{(2)}, \dots, \tau_t^{(M_2)}\}.$$

Let us arrange the states of the considered Markov chain $\xi_t, t \geq 0$, as follows. We enumerate the components n_t, r_t in the direct lexicographic order and, for fixed values of these components, we renumber the states of the processes $\mathbf{u}_t^{(1)}$ and $\mathbf{u}_t^{(2)}$ in the reverse lexicographic order.

To further describe the transition rates of the chain, we need the matrices $P_i(\cdot), A_i(\cdot, \cdot)$, and $L_i(\cdot, \cdot)$, which have the following probabilistic sense: the matrix $L_l(n, \tilde{S}_k)$ contains the transition rates of the process $\mathbf{u}_t^{(k)}$, leading to the end of servicing of one of $n - l$ customers of the k th type; the matrix $P_n(\beta_k)$ contains the transition probabilities of the process $\mathbf{u}_t^{(k)}$ leading to an increase in the number of customers of the k th type on the server from n to $n + 1$; the matrix $A_n(l, \tilde{S}_k)$ contains the transition rates of the process $\mathbf{u}_t^{(k)}$ in its state space without increasing or decreasing the number of customers of the k th type. Here $\tilde{S}_k = \begin{pmatrix} 0 & O \\ \mathbf{S}_0^{(k)} & S_l \end{pmatrix}, k = 1, 2$. Algorithm for calculating matrices $P_i(\cdot), A_i(\cdot, \cdot)$, and $L_i(\cdot, \cdot)$ follows from the results of V. Ramaswami and D. Lucantoni published in the papers [12, 13].

Let us introduce the notation $Q_{n,n'}$ for the matrices of transition rates of the chain from the states corresponding to the value n of the first component to the states corresponding to the value n' of this component, $n, n' = \overline{0, N}$. We also introduce the following notation:

- $C_n^m = \binom{n}{m} = \frac{n!}{m!(n-m)!}$;
- $diag\{a_1, a_2, \dots, a_n\}$ is a block diagonal matrix in which the diagonal blocks are equal to the elements listed in brackets, and the other blocks are zero;
- $diag^+\{a_1, a_2, \dots, a_n\}$ ($diag^-\{a_1, a_2, \dots, a_n\}$) is a square block matrix in which the off-diagonal (below-diagonal) blocks are equal to the elements listed in brackets, and the other blocks are zero.

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

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

where

$$\begin{aligned}
 Q_{0,0} &= \text{diag}^- \left\{ \frac{1}{r} L_{R-r}(R, \tilde{S}_2), r = \overline{1, R} \right\} \otimes I_{\bar{W}} \\
 &+ \text{diag} \left\{ 0, \frac{1}{r} A_r(R, S_2), r = \overline{1, R} \right\} \oplus D_0 + \text{diag} \left\{ O_{\sum_{r=0}^{R-1} C_{r+M_2-1}^{M_2-1}}, I_{C_{R+M_2-1}^{M_2-1}} \right\} \otimes D_2 \\
 &+ \text{diag}^+ \left\{ P_r(\beta_2), r = \overline{0, R-1} \right\} \otimes D_2 + \Delta_0; \\
 Q_{n,n+1} &= P_n(\beta_1) \otimes I_{\sum_{r=0}^R C_{r+M_2-1}^{M_2-1}} \otimes D_1, \quad 0 \leq n \leq N-1; \\
 Q_{n,n-1} &= \frac{1}{n} L_{N-n}(N, \tilde{S}_1) \otimes I_{\sum_{r=0}^R C_{r+M_2-1}^{M_2-1}} \otimes I_{\bar{W}}, \quad 1 \leq n \leq N; \\
 Q_{n,n} &= I_{C_{n+M_1-1}^{M_1-1}} \otimes \text{diag}^- \left\{ \frac{1}{r} L_{R-r}(R, \tilde{S}_2), r = \overline{1, R} \right\} \otimes I_{\bar{W}} \\
 &+ \frac{1}{n} A_n(N, \tilde{S}_1) \oplus \text{diag} \left\{ 0, \frac{1}{r} A_r(R, S_2), r = \overline{1, R} \right\} \oplus (D_0 + \delta_{n,N} D_1) \\
 &+ I_{C_{n+M_1-1}^{M_1-1}} \otimes \text{diag} \left\{ O_{\sum_{r=0}^{R-1} C_{r+M_2-1}^{M_2-1}}, I_{C_{R+M_2-1}^{M_2-1}} \right\} \otimes D_2 \\
 &+ I_{C_{n+M_1-1}^{M_1-1}} \otimes \text{diag}^+ \left\{ P_r(\beta_2), r = \overline{0, R-1} \right\} \otimes D_2 + \Delta_n, \quad 1 \leq n \leq N,
 \end{aligned}$$

where $\otimes(\oplus)$ denotes the Kronecker product (sum) of matrices, $\delta_{n,N}$ is the Kronecker symbol, $\Delta_n, n = \overline{0, N}$, are diagonal matrices, which are constructed so that the equality $Qe = \mathbf{0}^T$ holds.

Proof. The generator block $Q_{0,0}$ contains the transition rates in the set of states corresponding to the absence of customers of type 1. The corresponding transitions occur when

- a) one of the customers of type 2 finishes the service. The corresponding rates are given by the matrix $\text{diag}^- \left\{ \frac{1}{r} L_{R-r}(R, \tilde{S}_2), r = \overline{1, R} \right\} \otimes I_{\bar{W}}$;
- b) the number of customers of type 2 that are in a certain phase of servicing is changed or the *MMAP* underlying process makes an idle transition. The corresponding rates are given by the matrix $\text{diag} \left\{ 0, \frac{1}{r} A_r(R, S_2), r = \overline{1, R} \right\} \oplus D_0$;
- c) a customer of type 2 arrives and take place on the server (the matrix $\text{diag}^+ \left\{ P_r(\beta_2), r = \overline{0, R-1} \right\} \otimes D_2$) or, if all places for customers of this type are occupied, the customer leaves the system un-handled (the matrix $\text{diag} \left\{ O_{\sum_{r=0}^{R-1} C_{r+M_2-1}^{M_2-1}}, I_{C_{R+M_2-1}^{M_2-1}} \right\} \otimes D_2$).

Block $Q_{n,n}, n = \overline{1, N}$, contains the transition rates in the set of states corresponding to the presence of n customers of type 1 in the system. The expression for this block differs from the expression for the block $Q_{0,0}$ only in the second term, which in this case specifies the transition rates of the processes of servicing customers of types 1 and 2 in their sets of states without changing their numbers or the *MMAP* idle transition, or the loss of customer of type 1.

Block $Q_{n,n+1}, n = \overline{0, N-1}$, contains the rates of transitions accompanied by the arrival of a customer of type 1 which takes up place on the server.

Block $Q_{n,n-1}, n = \overline{1, N}$, contains the rates of transitions accompanied by the departure of the serviced customer of type 1 from the system.

All other blocks of the generator are zero matrices, since they consist of the rates of two or more transitions of the considered Markov chain on an infinitely small time interval.

4 Stationary Distribution. Performance Measures

In accordance with the described ordering of the states of the Markov chain ξ_t , we form the row vectors $\mathbf{p}_n, n = \overline{0, N}$, of the stationary probabilities of the states of the chain corresponding to the value n of the first component n_t . Let $\mathbf{p} = (\mathbf{p}_0, \mathbf{p}_1, \dots, \mathbf{p}_N)$ be the vector of steady state probabilities of the chain. This vector is the unique solution to the system of linear algebraic equation $\mathbf{p}Q = \mathbf{0}, \mathbf{p}\mathbf{e} = 1$. If the dimension of this system is large, the solution can be calculated using the algorithm developed in [14].

Based on the stationary distribution, we can obtain formulas for calculating a number of stationary performance characteristics of the system. Below we present some of them.

- Joint distribution of the number of type 1 customers on the server, the number of type 1 customers in different service phases, and the states of the *MMAP*

$$\mathbf{p}_n^* = \mathbf{p}_n(I_{C_{n+M_1-1}^{M_1-1}} \otimes \mathbf{e}_{\sum_{r=0}^R C_{r+M_2-1}^{M_2-1}} \otimes I_W), n = \overline{0, N}.$$

- Distribution of the number of customers of type 1 in the system $p_n = \mathbf{p}_n^* \mathbf{e}, n = \overline{0, N}$.
- Joint distribution of the number of type 2 customers on the server, the number of type 2 customers in different service phases, and the states of the *MMAP*

$$\mathbf{q}_r^* = \sum_{n=0}^N \mathbf{p}_n(I^{(n,r)} \otimes I_W), r = \overline{0, R},$$

where

$$I^{(n,r)} = \begin{pmatrix} O_{C_{n+M_1-1}^{M_1-1}} & \sum_{m=0}^{r-1} C_{m+M_2-1}^{M_2-1} \times C_{r+M_2-1}^{M_2-1} \\ \mathbf{e}_{C_{n+M_1-1}^{M_1-1}} & \otimes I_{C_{r+M_2-1}^{M_2-1}} \\ O_{C_{n+M_1-1}^{M_1-1}} & \sum_{m=r+1}^R C_{m+M_2-1}^{M_2-1} \times C_{r+M_2-1}^{M_2-1} \end{pmatrix}.$$

- Distribution of the number of customers of type 2 in the system $q_r = \mathbf{q}_r^* \mathbf{e}, r = \overline{0, R}$.
- The probability of losing a customer of the k th type

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

where λ_k is the arrival rate of customers of k th type, φ_k is the output rate of customers of k th type. The value of φ_k is calculated as

$$\begin{aligned}\varphi_1 &= \sum_{n=1}^N \mathbf{p}_n^* (I_{C_{n+M_1-1}^{M_1-1}} \otimes \mathbf{e}_{\bar{W}}) \frac{1}{n} L_{N-n}(N, \tilde{S}_1) \mathbf{e}, \\ \varphi_2 &= \sum_{r=1}^R \mathbf{q}_r^* (I_{C_{r+M_2-1}^{M_2-1}} \otimes \mathbf{e}_{\bar{W}}) \frac{1}{r} L_{R-r}(R, \tilde{S}_2) \mathbf{e}.\end{aligned}$$

5 Sojourn Time Distribution

Denote by $v_n^{(1)}(\eta^{(1)}, \eta^{(2)}, \dots, \eta^{(M_1)}, \tilde{\eta}, \nu, s)$ the Laplace-Stieltjes transform (*LST*) of the virtual sojourn time distribution of a customer of type 1 for which service began with the phase $\tilde{\eta}$, and which found n customers of the first type in the system, the number of customers in phase $m^{(1)}$ equal to $\eta^{(m^{(1)})}$, and the underlying process of the *MMAP* in the state ν , $n = \overline{0, N-1}$, $\eta^{(m^{(1)})} = \overline{0, n}$, $m^{(1)} = \overline{1, M_1}$, $\nu = \overline{0, \bar{W}}$.

Similarly, let $v_r^{(2)}(\tau^{(2)}, \dots, \tau^{(M_2)}, \tilde{\tau}, \nu, s)$ be the Laplace-Stieltjes transform of the virtual sojourn time distribution of a customer of type 2 for which service began with the phase $\tilde{\tau}$, and which found in the system r customers of the second type, the number of customers in the phase $m^{(2)}$ equal to $\tau^{(m^{(2)})}$, and the underlying process of the *MMAP* in the state ν , $r = \overline{0, R-1}$, $\tau^{(m^{(2)})} = \overline{0, r}$, $m^{(2)} = \overline{1, M_2}$, $\nu = \overline{0, \bar{W}}$. First we derive formulas of the conditional *LSTs* $v_n^{(1)}(\eta^{(1)}, \eta^{(2)}, \dots, \eta^{(M_1)}, \tilde{\eta}, \nu, s)$. Let us arrange these *LSTs* in the reverse lexicographic order of arguments $\eta^{(1)}, \eta^{(2)}, \dots, \eta^{(M_1)}$, in the direct lexicographic order of arguments $\tilde{\eta}, \nu$ and form the column vectors

$$\mathbf{v}_n^{(1)}(s), n = \overline{0, N-1}, \quad \mathbf{v}^{(1)}(s) = ((\mathbf{v}_0^{(1)}(s))^T, (\mathbf{v}_1^{(1)}(s))^T, \dots, (\mathbf{v}_{N-1}^{(1)}(s))^T)^T.$$

Similarly, for customers of type 2, we form the column vectors

$$\mathbf{v}_r^{(2)}(s), r = \overline{0, R-1}, \quad \mathbf{v}^{(2)}(s) = ((\mathbf{v}_0^{(2)}(s))^T, (\mathbf{v}_1^{(2)}(s))^T, \dots, (\mathbf{v}_{R-1}^{(2)}(s))^T)^T.$$

Theorem 1. *The Laplace-Stieltjes transform vector $\mathbf{v}^{(1)}(s)$ is calculated as follows:*

$$\mathbf{v}^{(1)}(s) = (sI - A^{(1)})^{-1} \mathbf{b}^{(1)}, \quad (1)$$

where

$$\begin{aligned}A^{(1)} &= \text{diag}\left\{\left[\frac{1}{n+1} A_n(N, S_1) + \Delta_n\right] \oplus S_1 \oplus (D_0 + D_2), n = \overline{0, N-1}\right\} \\ &\quad + \text{diag}^- \left\{ \frac{1}{n+1} L_{N-n}(N, \tilde{S}_1) \otimes I_{M_1 \bar{W}}, n = \overline{1, N-1} \right\} \\ &\quad + \text{diag}^+ \left\{ P_n(\beta_1) \otimes I_{M_1} \otimes D_1, n = \overline{0, N-2} \right\} \\ &\quad + \text{diag}\left\{ O_{\bar{W} \sum_{n=0}^{N-2} C_{n+M_1-1}^{M_1-1}}, I_{C_{N+M_1-2}^{M_1-1}} \right\} \otimes I_{M_1} \otimes D_1\}, \\ \mathbf{b}^{(1)} &= \text{diag}\left\{ I_{C_{n+M_1-1}^{M_1-1}} \otimes \frac{1}{n+1} \mathbf{S}_0^{(1)} \otimes I_{\bar{W}}, n = \overline{0, N-1} \right\} \mathbf{e}.\end{aligned}$$

Proof. Using the probabilistic interpretation of the Laplace-Stieltjes transform, we obtain the following equations for the vectors $\mathbf{v}_n^{(1)}(s), n = \overline{0, N-1}$:

$$\begin{aligned}
 \mathbf{v}_n^{(1)}(s) = & \int_0^\infty e^{-st} e^{[\frac{1}{n+1} A_n(N, S_1) + \Delta_n] \oplus S_1 t} (I_{C_{n+M_1-1}^{M_1-1}} \otimes \frac{1}{n+1} \mathbf{S}_0^{(1)}) \otimes e^{(D_0+D_2)t} dt \mathbf{e} \\
 & + \int_0^\infty e^{-st} e^{[\frac{1}{n+1} A_n(N, S_1) + \Delta_n] \oplus S_1 t} (L_{N-n}(N, \tilde{S}_1) \otimes I_{M_1}) \otimes e^{(D_0+D_2)t} dt \mathbf{ev}_{n-1}^{(1)}(s) \\
 & + \int_0^\infty (e^{-st} e^{[\frac{1}{n+1} A_n(N, S_1) + \Delta_n] \oplus S_1 t} \otimes e^{(D_0+D_2)t}) (P_n(\boldsymbol{\beta}_1) \otimes I_{M_1} \otimes D_1) dt \\
 & \times \mathbf{v}_{\min\{n+1, N-1\}}^{(1)}(s) \mathbf{e}. \tag{2}
 \end{aligned}$$

Let us explain the meaning of the terms on the right-hand side of (2):

- the first integral (first term) is the probability that the incoming virtual customer will be serviced before any of the n customers of type 1 that are already on the server at the time of the virtual customer arriving, and during the time of servicing the virtual customer there will be no catastrophe.
- the integral in the second term is the vector of probabilities that after the arrival of the virtual customer one of the n customers of type 1 that are already on the server at the time of the arrival of the virtual customer will be served first, and no catastrophe will occur during the service of this first customer. After the first of the mentioned n customers is served, the server resource is redistributed between the remaining i customers, including the virtual one, and the further scenario of servicing the virtual customer up to the distribution of the *MMAP* states and service phases will be the same as at the moment of the arrival of a virtual customer that found $n-1$ customers in the system. By definition, the corresponding vector of *LSTs* is $\mathbf{v}_{n-1}^{(1)}(s)$. The product of the integral and $\mathbf{v}_{n-1}^{(1)}(s)$ will give the required vector of *LSTs* of the service time distribution of the virtual customer.
- when describing the third term, we will distinguish between the cases $n < N-1$ and $n = N-1$. In both cases, the integral in the third term is a vector of probabilities that after the arrival of the virtual customer, the first event that entails a change in the number of customers on the server will be the arrival of a customer of type 1 and no catastrophe will occur in the time before it arrives. In the case $n < N-1$, after this customer arrives, the server resource will be redistributed between $n+2$ customers, including the virtual one, and the further scenario of servicing the virtual customer up to the distribution of the *MMAP* states and servicing phases will be the same as at the moment of arrival of the virtual customer that found $n+1$ customers in the system. By definition, the corresponding vector of *LSTs* is $\mathbf{v}_{n+1}^{(1)}(s)$. The product of

the integral and $\mathbf{v}_{n+1}^{(1)}(s)$ will give the required vector of *LSTs* of service time of the virtual customer. In the case $n = N - 1$ the received customer will be rejected, since the server already contains N customers, including the virtual one. Then the further scenario of servicing the virtual customer, up to the distribution of the states of the *MMAP*, will be the same as at the moment of the arrival of the virtual customer that found $N - 1$ customers in the system. By definition, the corresponding vector of *LSTs* is $\mathbf{v}_{N-1}^{(1)}(s)$. The product of the integral and $\mathbf{v}_{N-1}^{(1)}(s)$ will give the required vector of *LSTs* of service time of the virtual customer.

After calculating the integrals in (2) and a number of algebraic transformations, we obtain the required formula (1).

Corollary 1. *The Laplace-Stieltjes transform vector $\mathbf{v}^{(2)}(s)$ is calculated as follows:*

$$\mathbf{v}^{(2)}(s) = (sI_{\overline{W}} - A^{(2)})^{-1}\mathbf{b}^{(2)},$$

where the matrix $A^{(2)}$ and the vector $\mathbf{b}^{(2)}$ are obtained from the matrix $A^{(1)}$ and the vector $\mathbf{b}^{(1)}$, respectively, by replacing N by R and permutation of indices 1 and 2.

Having known the Laplace-Stieltjes transforms defined in Theorem 1 and Corollary 1, we can find all the moments of the sojourn time, in particular, the mean and the variance of this time.

The corresponding mean (variance) for customers of type 1 we denote as $\bar{v}_n^{(1)}(\eta^{(1)}, \eta^{(2)}, \dots, \eta^{(M_1)}, \tilde{\eta}, \nu)$ ($d_n^{(1)}(\eta^{(1)}, \eta^{(2)}, \dots, \eta^{(M_1)}, \tilde{\eta}, \nu)$) and for customers of type 2 as $\bar{v}_r^{(2)}(\tau^{(1)}, \dots, \tau^{(M_2)}, \tilde{\tau}, \nu)$ ($d_n^{(2)}(\tau^{(1)}, \tau^{(2)}, \dots, \tau^{(M_2)}, \tilde{\tau}, \nu)$).

We renumber the values $\bar{v}_n^{(1)}(\eta^{(1)}, \eta^{(2)}, \dots, \eta^{(M_1)}, \tilde{\eta}, \nu)$, $(\bar{v}_n^{(1)}(\eta^{(1)}, \eta^{(2)}, \dots, \eta^{(M_1)}, \tilde{\eta}, \nu))^2$ and $d_n^{(1)}(\eta^{(1)}, \eta^{(2)}, \dots, \eta^{(M_1)}, \tilde{\eta}, \nu)$ in the lexicographic order described above and form the corresponding column vectors

$$\bar{\mathbf{v}}_n^{(1)}, \bar{\bar{\mathbf{v}}}_n^{(1)}, \bar{\mathbf{d}}_n^{(1)}, n = \overline{0, N - 1}.$$

In turn, from these vectors we form the column vectors

$$\bar{\mathbf{v}}^{(1)} = ((\bar{\mathbf{v}}_0^{(1)})^T, (\bar{\mathbf{v}}_1^{(1)})^T, \dots, (\bar{\mathbf{v}}_{N-1}^{(1)})^T)^T, \quad \bar{\bar{\mathbf{v}}}^{(1)} = ((\bar{\bar{\mathbf{v}}}_0^{(1)})^T, (\bar{\bar{\mathbf{v}}}_1^{(1)})^T, \dots, (\bar{\bar{\mathbf{v}}}_{N-1}^{(1)})^T)^T, \\ \mathbf{d}^{(1)} = ((\mathbf{d}_0^{(1)})^T, (\mathbf{d}_1^{(1)})^T, \dots, (\mathbf{d}_{N-1}^{(1)})^T)^T.$$

By analogy we introduce the column vectors $\bar{\mathbf{v}}^{(2)}, \bar{\bar{\mathbf{v}}}^{(2)}, \mathbf{d}^{(2)}$.

Corollary 2. *The vector of conditional means, $\bar{\mathbf{v}}^{(k)}$, and the vector of conditional variances, $\mathbf{d}^{(k)}$, of the sojourn times of a customer of type k are calculated by the following formulas:*

$$\bar{\mathbf{v}}^{(k)} = -(A^{(k)})^{-1}\mathbf{e}, \quad \mathbf{d}^{(k)} = 2(A^{(k)})^{-2}\mathbf{e} - \bar{\mathbf{v}}^{(k)}, \quad k = 1, 2.$$

To calculate the Laplace-Stieltjes transforms of the sojourn time distributions of the customers of type 1 and 2 admitted into the system, we introduce into consideration the vector \mathbf{p}^+ (\mathbf{q}^+), the components of which define the joint distribution of the number of type 1 (type 2) customers that are in different service phases and the states of the *MMAP* immediately after the moment the customer of type 1 (type 2) has been admitted into the system. It is easy to see that these vectors are calculated as follows:

$$\mathbf{p}^+ = \lambda_1^{-1}(\mathbf{p}_0^*, \mathbf{p}_1^*, \dots, \mathbf{p}_{N-1}^*)[\text{diag}\{P_n(\boldsymbol{\beta}_1), n = \overline{0, N-1}\} \otimes D_1],$$

$$\mathbf{q}^+ = \lambda_2^{-1}(\mathbf{q}_0^*, \mathbf{q}_1^*, \dots, \mathbf{q}_{R-1}^*)[\text{diag}\{P_r(\boldsymbol{\beta}_2), r = \overline{0, R-1}\} \otimes D_2].$$

Theorem 2. *The Laplace-Stieltjes transformations of the sojourn time distributions of the customers of type 1 and type 2 accepted to the system are calculated as*

$$v^{(1)}(s) = \mathbf{p}^+ \mathbf{v}^{(1)}(s), \quad v^{(2)}(s) = \mathbf{q}^+ \mathbf{v}^{(2)}(s).$$

Corollary 3. *The means and variances of the sojourn times of customers of type 1 and type 2 accepted to the system are calculated using the following formulas:*

$$\bar{v}^{(1)} = \mathbf{p}^+ \bar{\mathbf{v}}^{(1)}, \quad d^{(1)} = \mathbf{p}^+ \mathbf{d}^{(1)}; \quad \bar{v}^{(2)} = \mathbf{q}^+ \bar{\mathbf{v}}^{(2)}, \quad d^{(2)} = \mathbf{q}^+ \mathbf{d}^{(2)}.$$

6 Numerical Results

In this section we conduct a number of numerical experiments aimed at studying the behavior of the performance characteristics of the system depending on its parameters and at solving optimization problems. To carry out the experiments, a computer program was written in Python using built-in packages for processing matrices, calculating complex mathematical formulas and executed in the PyCharm 2019.3.4 (Professional Edition) program.

In **Experiment 1** we analyse the dependence of the loss probabilities, $P_{loss,k}, k = 1, 2$, and the mean sojourn times, $\bar{v}^{(k)}, k = 1, 2$, on the maximum number of channels allocated for customers of type k . In this experiment we used the following input data.

The *MMAP* is specified by the matrices D_0, D_1, D_2 , where

$$D_0 = \begin{pmatrix} -86 & 0.01 \\ 0.02 & -2.76 \end{pmatrix}, \quad D_1 = \begin{pmatrix} 59.5 & 0.693 \\ 0.14 & 1.778 \end{pmatrix}, \quad D_2 = \begin{pmatrix} 25.5 & 0.297 \\ 0.06 & 0.762 \end{pmatrix}.$$

With such matrices $\lambda = 12.43, \lambda_1 = 0.7\lambda$ and $\lambda_2 = 0.3\lambda, c_{cor}^{(1)} = 0.39, c_{cor}^{(2)} = 0.33$.

The *PH* distribution of the service time of a single customer of type 1 is given by the vector $\boldsymbol{\beta}^{(1)} = (1, 0)$ and the matrix $S^{(1)} = \begin{pmatrix} -80 & 80 \\ 0 & -80 \end{pmatrix}$. This means that the service time has Erlang distribution of order 2 with parameter 80 and the service rate $\mu_1 = 40$.

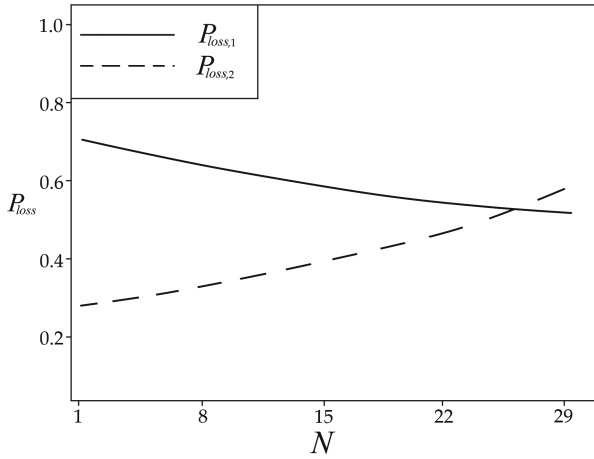


Fig. 1. $P_{loss,1}$ and $P_{loss,2}$ vs N under fixed number of channels $N + R = 30$

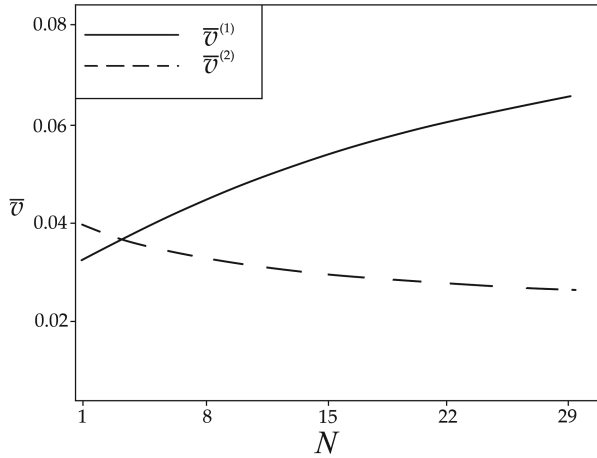


Fig. 2. $\bar{v}^{(1)}$ and $\bar{v}^{(2)}$ vs N under fixed number of channels $N + R = 30$

The PH distribution of the service time of a single customer of type 2 is given by the vector $\beta^{(2)} = (1, 0)$ and the matrix $S^{(2)} = \begin{pmatrix} -20 & 20 \\ 0 & -20 \end{pmatrix}$. This means that the service time has Erlang distribution of order 2 with parameter 20 and the service rate $\mu_2 = 10$.

The total number of channels, into which the throughput of the servers is divided, is $N + R = 30$.

It is seen from Fig. 1 that $P_{loss,1}$ decreases and $P_{loss,2}$ increases. This is due to the fact that with an increase in N the possible number of type 1 customers in the system increases and the smaller part of the customers will be lost. Taking

into account the equality $N + R = 30$, with an increase in N the value of R decreases and more and more customers are lost.

Figure 2 shows that $\bar{v}^{(1)}$ is an increasing function of N . This is due to the fact that with an increase in N the throughput allocated for a customer of type 1 decreases and hence the sojourn time increases. Due to the relation $N + R = 30$ when N increases then R decreases. That entails an increase in the throughput available for a customer of type 2 and a decrease in the time for servicing the customer.

Experiment 2. In this experiment, we solve numerically the optimization problem which consists in the optimal sharing of the throughput $\mu = \mu_1 + \mu_2$ of the server between customers of types 1 and 2 and the optimal choice of the maximum numbers of simultaneously served customers of types 1 and 2 under the given restrictions on the minimum throughput allocated for each customer.

As a criterion for the quality of the operation of the system, we use the economic functional, which is the average penalty per unit of time

$$J = a\bar{N} + c_1\lambda_1P_{loss,1} + c_2\lambda_2P_{loss,2}, \tag{3}$$

where a is the penalty charged per unit of time spent by one customer of type 1 in the system, c_k is the penalty charged for the lost customer of type k , $k = 1, 2$.

The problem is to choose the parameters μ_1 , N and R which provide the minimum to criterion (3) under the following conditions:

$$\mu_1 + \mu_2 = \mu = const, \quad \gamma_1 = \frac{\mu_1}{N} = const, \quad \gamma_2 = \frac{\mu_2}{R} = const.$$

Here γ_k is the minimum throughput of the server that can be used to provide service to a customer of type k .

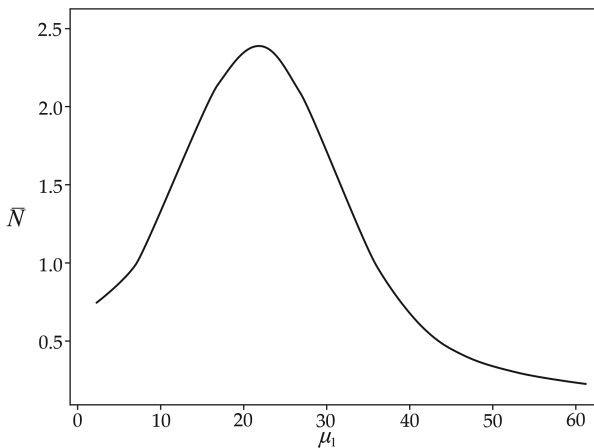


Fig. 3. \bar{N} vs μ_1 under restrictions $\mu = 70, \gamma_1 = 2, \gamma_2 = 7$

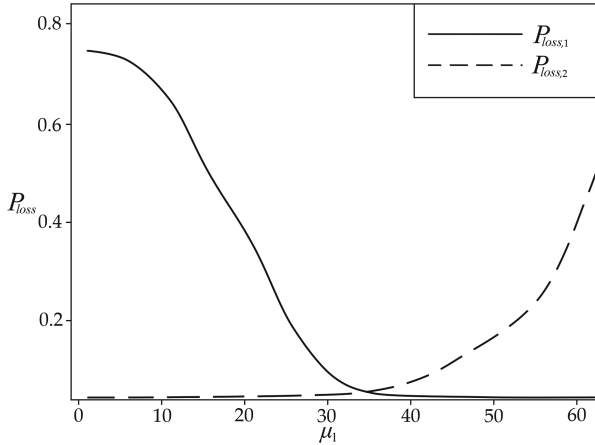


Fig. 4. $P_{loss,1}, P_{loss,2}$ vs μ_1 under restrictions $\mu = 70, \gamma_1 = 2, \gamma_2 = 7$

In the experiment, we will use the *MMAP* specified in Experiment 1. The shape of service time distributions is the same as in Experiment 1. In the course of the current experiment, we will only change the service rates μ_1 and μ_2 multiplying the matrices S_1, S_2 by the corresponding constants. We fix the values of μ, γ_1, γ_2 as $\mu = 70, \gamma_1 = 2, \gamma_2 = 7$.

For these initial data, let us look at the graphs of the dependence of the mean number of customers of the type 1, \bar{N} , and the probabilities of losses of customers of different types, $P_{loss,1}, P_{loss,2}$, which are shown in Fig. 3 and 4.

Having calculated the dependence of $\bar{N}, P_{loss,1}, P_{loss,2}$ on μ_1 we can calculate the dependence of the cost criteria J on μ_1 under different cost coefficients. Let

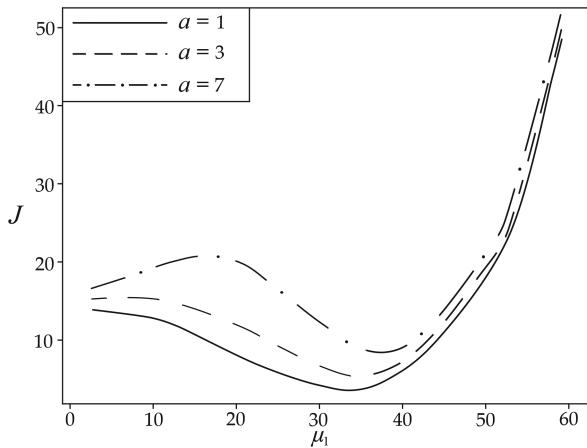


Fig. 5. J vs μ_1 for $c_1 = 1, c_2 = 20, a = 1, 3, 7$ under restrictions $\mu = 70, \gamma_1 = 2, \gamma_2 = 7$

us consider the following cost coefficients: $a = 1, 3, 7, c_1 = 1, c_2 = 20$. The results of calculation are presented in Fig. 5 and in Table 1.

Table 1. Values of N, R, J as functions of μ_1 for $c_1 = 1, c_2 = 20, a = 1, 3, 7$ under restrictions $\mu = 70, \gamma_1 = 2, \gamma_2 = 7$

μ_1	N	R	$J, a = 1$	$J, a = 3$	$J, a = 7$
2	1	9	13.09	14.60	17.64
5	2	9	12.81	14.76	18.67
10	5	8	11.69	14.91	21.34
15	7	7	9.24	13.58	22.24
20	10	7	6.54	11.40	21.11
25	12	6	3.96	8.17	16.57
30	15	5	2.45	5.58	11.83
35	17	5	2.15	4.04	7.83
40	20	4	4.13	5.32	7.72
45	22	3	8.82	9.65	11.32
50	25	2	16.99	17.62	18.89
55	27	2	22.84	23.35	24.37
60	30	1	44.79	45.22	46.07
63	31	1	50.32	50.70	51.48

It is seen from Fig. 5 and Table 1 that for input data under consideration the server throughput is divided approximately in half between customers of types 1 and 2. In the case $a = 1, 3$, it is optimal to divide the throughput allocated to customers of types 1 and 2 as 17:5. When $a = 7$, this proportion changes as 20:4.

7 Conclusion

We analysed a queuing system with the *MMAP* of customers of two types, processor sharing and a limited number of places for customers of different types. We described the system operation by the multi-dimensional Markov chain, calculated its stationary distribution and the main performance characteristics. The Laplace-Stieltjes transform of the sojourn time distribution is found. Formulas for means and variances of the sojourn time are obtained. We carried out numerical experiments to study the behavior of the system performance characteristics and to find the optimal strategy for sharing the processor throughput between users of different types. The results obtained can be used in the study and planning of telecommunication networks for various purposes, in particular, the Internet of Things.

References

1. Ghosh, A., Banik, A.D.: An algorithmic analysis of the *BMAP/MSP/1* generalized processor-sharing queue. *Comput. Oper. Res.* **79**, 1–11 (2017)
2. Telek, M., van Houdt, B.: Response time distribution of a class of limited processor sharing queues. *ACM SIGMETRICS Perform. Eval. Rev.* **45**, 143–155 (2018). <https://doi.org/10.1145/3199524.3199548>
3. Yashkov, S., Yashkova, A.: Processor sharing: a survey of the mathematical theory. *Autom. Remote. Control.* **68**, 662–731 (2007)
4. Zhen, Q., Knessl, C.: On sojourn times in the finite capacity *M/M/1* queue with processor sharing. *Oper. Res. Lett.* **37**, 447–450 (2009)
5. Masuyama, H., Takine, T.: Sojourn time distribution in a *MAP/M/1* processor-sharing queue. *Oper. Res. Lett.* **31**, 406–412 (2003)
6. Kennedy, E., Bulega, T.: Resource sharing between M2M and H2H traffic under time-controlled scheduling scheme in LTE networks. In: Proceedings of the 8th International Conference on Telecommunication Systems Services and Applications (TSSA) (2014). <https://doi.org/10.1109/TSSA.2014.7065909>
7. Fawal, A., Najem, M., Mansour, A., Roy, F., Jeune, D.: CTMC modelling for H2H/M2M coexistence in LTE-A/LTE-M networks. *J. Eng.* **12**, 1954–1962 (2018)
8. Ahmadi, M., Golkarifard, M., Movaghar, A., Yousefi, H.: Processor sharing queues with impatient customers and state-dependent rates. *IEEE/ACM Trans. Netw.* **29**(6), 2467–2477 (2021). <https://doi.org/10.1109/TNET.2021.3091189>
9. Dudin, S., Dudin, A., Dudina, O., Samouylov, K.: Analysis of a retrial queue with limited processor sharing operating in the random environment. In: Koucheryavy, Y., Mamatas, L., Matta, I., Ometov, A., Papadimitriou, P. (eds.) *WWIC 2017*. LNCS, vol. 10372, pp. 38–49. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-61382-6_4
10. Dudin, A., Dudin, S., Dudina, O., Samouylov, K.: Analysis of queueing model with limited processor sharing discipline and customers impatience. *Oper. Res. Perspect.* **5**, 245–255 (2018)
11. He, Q.M.: Queues with marked customers. *Adv. Appl. Probab.* **28**, 567–587 (1996)
12. Ramaswami, V.: Independent Markov processes in parallel. *Commun. Statist.-Stochastic Models* **1**, 419–432 (1985)
13. Ramaswami, V., Lucantoni, D.M.: Algorithms for the multi-server queue with phase-type service. *Commun. Statist.-Stochastic Models* **1**, 393–417 (1985)
14. Klimenok, V.I., Kim, C.S., Orlovsky, D.S., Dudin, A.N.: Lack of invariant property of Erlang *BMAP/PH/N/0* model. *Queueing Syst.* **49**, 187–213 (2005)