Analysis of a Retrial Queue with Limited Processor Sharing Operating in the Random Environment

Sergey Dudin, Alexander Dudin, Olga Dudina, Konstantin Samouylov

Analysis of a retrial queue with limited processor sharing operating in the random environment

Sergey Dudin[1,2], Alexander Dudin[1⋆], Olga Dudina [1,2], and Konstantin Samouylov[2]

[1] Belarusian State University, 4 Nezavisimosti Ave.,220030, Minsk, Belarus
[2] RUDN University, 6 Miklukho-Maklaya st., 117198, Moscow, Russia
dudins@bsu.by,dudin@bsu.by,dudina@bsu.by,ksam@sci.pfu.edu.ru

Abstract. Queueing system with limited processor sharing, which operates in the Markovian random environment, is considered. Parameters of the system (pattern of the arrival rate, capacity of the server, i.e., the number of customers than can share the server simultaneously, the service intensity, the impatience rate, etc.) depend on the state of the random environment. Customers arriving when the server capacity is exhausted join orbit and retry for service later. The stationary distribution of the system states (including the number of customers in orbit and in service) is computed and expressions for the key performance measures of the system are derived. Numerical example illustrates possibility of optimal adjustment of the server capacity to the state of the random environment.

Keywords: Processor sharing, Markovian arrival process, Random environment

1 Introduction

Processor sharing discipline is widely applied for modelling and analysis of communication systems and networks. For references and examples of real world applications, the recent papers [1, 2] can be recommended along with the known surveys [3, 4]. Generally speaking, a processor can be shared by infinitely many users. However, in many applications, especially to wireless communication networks, too small share of the bandwidth of the channel assigned to a customer may lead to poor service and its termination due to too long service. Therefore, the **limited** processor sharing is often considered. This kind of processor sharing suggests that the maximal number of users who obtain service simultaneously is fixed. This number is called as the server capacity. Customers arriving when capacity of the server is not exhausted immediately start service with the rate inversely proportional to the number of customers in service.

The model considered in this paper has the following features previously not addressed or only partially addressed in the relevant literature.

⋆ Corresponding author.

1) It is usually assumed that an arriving customer can be lost or queued to the finite or infinite buffer if the capacity of the server is exhausted. In our paper, we assume a more realistic in application to wireless networks scenario that such the arriving customer virtually moves to so-called orbit from which he/she makes the repeated attempts (retrials) to obtain service after the random time intervals. It is well-known that the phenomenon of retrials is typical in wireless communication networks and that analysis of retrial queueing models is more difficult comparing to the queues with buffers, see, e.g., [5].

2) The customers in service can be impatient. They may leave the server before service completion after an exponentially distributed amount of time the parameter of which depends on the number of customers in service.

3) We assume that the system operates in the random environment (RE). This means that the parameters of the system (pattern of the arrival rate, capacity of the server, intensities of service and impatience rate, etc.) depend on the state of the RE. They instantaneously change their values at the moment of a jump of the RE to another state. As special case, our model includes the systems with processor sharing and unreliable servers, see, e.g., [1, 2, 6]. Consideration of the system operating in the RE is important for potential applications in wireless networks because the server capacity and other parameters can be changed due to redistribution of the system resources among the existing servers due to many reasons, including the users mobility, noise in radio-channel, etc. We assume that the behavior of the RE does not depend on the state of the system while such a dependence is suggested in [7]. However, in that paper capacity of the system does not depend on the state of the RE while we allow such a dependence.

4) As well as in [2], we assume quite general model of arrival process, namely, Markovian arrival process, while the overwhelming majority of existing papers deal with the stationary Poisson arrival process. Model considered in [8] assumes the batch Markovian arrival process. However, the number of customers, which can get service simultaneously, is not limited in that paper.

2 Mathematical model

We consider a retrial single-server queueing system with limited server (processor) sharing discipline.

All system parameters depend on the state of the RE. The RE is defined by the stochastic process $r_t,\ t \geq 0$. This process is an irreducible continuous-time Markov chain with finite state space $\{1, 2, \ldots, R\}$ and the infinitesimal generator H.

The structure of the system under study is presented in Figure 1.

Arrival of customers to the system is described by the process which is a slight generalization of the well-known Markovian Arrival Process (MAP) introduced in [9]. Arrival are governed by the underlying process $\{r_t, \nu_t\},\ t \geq 0$, where r_t is the state of the RE and the process ν_t with finite state space $\{0, 1, \ldots, W\}$ is defined as follows. Under the fixed state r of the RE the process ν_t behaves as an

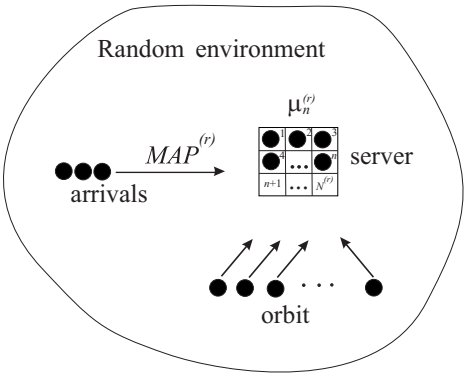

Fig. 1. Queueing system under study

irreducible continuous-time Markov chain. The sojourn time of this chain in the state ν is exponentially distributed with the positive finite parameter $\lambda_\nu^{(r)}$. When the sojourn time in the state ν expires, with probability $p_0^{(r)}(\nu,\nu')$ the process ν_t jumps to the state ν' without generation of a customer, $\nu,\nu' = \overline{0,W}$, $\nu \neq \nu'$, $r = \overline{1,R}$. With probability $p_1^{(r)}(\nu,\nu')$, the process ν_t jumps to the state ν' with generation of a customer, $\nu,\nu' = \overline{0,W}$, $r = \overline{1,R}$.

The behavior of the arrival process under the fixed state r of the RE is completely characterized by the matrices $D_0^{(r)}$ and $D_1^{(r)}$ defined by the entries

$$(D_0^{(r)})_{\nu,\nu} = -\lambda_\nu^{(r)}, \nu = \overline{0,W},\ (D_0^{(r)})_{\nu,\nu'} = \lambda_\nu^{(r)} p_0^{(r)}(\nu,\nu'),\ \nu,\nu' = \overline{0,W}, \nu \neq \nu',$$

$$(D_1^{(r)})_{\nu,\nu'} = \lambda_\nu^{(r)} p_1^{(r)}(\nu,\nu'),\ \nu,\nu' = \overline{0,W},\ r = \overline{1,R}.$$

The average arrival rate $\lambda^{(r)}$ under the fixed state r of the RE is given as $\lambda^{(r)} = \boldsymbol{\theta}^{(r)} D_1^{(r)} \mathbf{e}$ where $\boldsymbol{\theta}^{(r)}$ is the invariant vector of the stationary distribution of the Markov chain ν_t, $t \geq 0$, under the fixed state r. The vector $\boldsymbol{\theta}^{(r)}$ is the unique solution to the system $\boldsymbol{\theta}^{(r)} D^{(r)}(1) = \mathbf{0}$, $\boldsymbol{\theta}^{(r)}\mathbf{e} = 1$. Here and in the sequel $\mathbf{0}$ is the zero row vector and $\mathbf{e}$ is the column vector of appropriate size consisting of ones.

Let us introduce the following matrices: $\tilde{D}_1 = \mathrm{diag}\{D_1^{(r)},\ r = \overline{1,R}\}$, $\tilde{D}_0 = H \otimes I_{W+1} + \mathrm{diag}\{D_0^{(r)},\ r = \overline{1,R}\}$, where $\mathrm{diag}\{\dots\}$ denotes the diagonal matrix with the diagonal entries listed in the brackets.

The averaged (over all states of the RE) intensity of input flow of customers λ is defined as $\lambda = \boldsymbol{\theta}\tilde{D}_1\mathbf{e}$ where the vector $\boldsymbol{\theta}$ is the unique solution of the system

$$\boldsymbol{\theta}(\tilde{D}_0 + \tilde{D}_1) = \mathbf{0},\ \boldsymbol{\theta}\mathbf{e} = 1.$$

The squared coefficient of variation c_{var} of intervals between successive arrivals is given as $c_{var} = 2\lambda\boldsymbol{\theta}(-\tilde{D}_0)^{-1}\mathbf{e} - 1$. The coefficient of correlation c_{cor} of two successive intervals between arrivals is given as $c_{cor} = (\lambda\boldsymbol{\theta}(-\tilde{D}_0)^{-1}\tilde{D}_1(-\tilde{D}_0)^{-1}\mathbf{e} - 1)/c_{var}$.

Under the fixed state r of the RE, up to $N^{(r)}$ customers can obtain service simultaneously. We call the number $N^{(r)}$ as server capacity under the state r of the RE, $r = \overline{1, R}$. Without the loss of generality, let us assume that the states of the RE are enumerated in ascending order of the server capacity, i.e.,

$$0 \le N^{(1)} \le N^{(2)} \le \cdots \le N^{(R)}.$$

We permit that the server capacity can be equal to 0 under some states of the RE. This allows us to consider the model with server breakdowns as the partial case of the model under study and use the presented below results for analysis of the model with server breakdowns.

If during an arbitrary customer arrival epoch the number of customers in service is less than $N^{(r)}$, the customer is admitted and starts obtaining service immediately. Otherwise, with probability $q^{(r)}$, $0 \le q^{(r)} \le 1$, the arriving customer joins orbit and retries later or leaves the system permanently with the complimentary probability. Each customer from orbit makes the repeated attempts (retrials) to obtain service after an exponentially distributed with the parameter $\gamma^{(r)}$, $0 \le \gamma^{(r)} < \infty$, time. If the attempt will be successful, i.e. if the number of customers in service is less than $N^{(r)}$, the retrial customer is accepted for service. Otherwise, the retrial customer returns to the orbit with probability $q^{(r)}$ or leaves the system permanently with the complimentary probability.

The service rate of each customer depends on the number of customers that obtain service. Under the fixed state r of the RE, the service rate of each customer is $\mu_n^{(r)}$ where $0 \le \mu_n^{(r)} < \infty$ if n customers are obtaining service simultaneously, $n = \overline{1, N^{(r)}}$. For the sake of mathematical generality, in our analysis we do not impose any special restrictions on dependence of values $\mu_n^{(r)}$ on n. However, it looks realistic to assume that the increase of the number n of simultaneously serviced customers implies the decrease of the individual service rate, i.e., for each state r of the RE, the following inequalities are satisfied: $\mu_1^{(r)} \ge \mu_2^{(r)} \ge \cdots \ge \mu_{N^{(r)}}^{(r)}$. The most popular in the literature form of dependence of the intensity $\mu_n^{(r)}$ on n is: $\mu_n^{(r)} = \frac{\mu^{(r)}}{n}$ where $\mu^{(r)}$ is the fixed constant characterizing the total service rate under the state r of the RE. This popular dependence satisfies our assumption as a very particular case.

The customers obtaining service can be impatient, i.e., a customer can leave server without completing service. We assume that the individual customer's intensity of impatience also depends on the number of customers in service and if there are n customers in service, each customer leaves the server due to impatience after an exponentially distributed with the parameter $\beta_n^{(r)}$ time, $\beta_n^{(r)} \ge 0$. After leaving the server due to impatience, a customer joins the orbit with probability $a^{(r)}$ or leaves the system permanently with the complimentary probability.

Because the server capacity depends on the state of the RE, the transition of the RE from one state to another one may lead to decreasing the server capacity. We assume that in the case where n customer obtain service during the epoch of the transition of the RE from the state r to the state r', $r' < r$, and $n > N^{(r')}$, i.e. the "new" server capacity is less that the number of customers in service, then

$n - N^{(r')}$ customers are forced to terminate service. If the customer is forced to terminate service, he/she leaves the system permanently with probability $1 - p^{(r)}$, $r = \overline{2, R}$, or joins orbit with the complimentary probability. If the transition of the underlying process of the RE leads to increasing the server capacity, the number of customers that obtain service at the epoch of transition does not change.

Our goal is to analyse the stationary behavior of the described queueing model under the fixed parameters of the RE and system having in mind a possibility of further use of the results of analysis for various managerial purposes, e.g., adjusting the values of capacities $N^{(r)}$ to the corresponding arrival rates and requirements to the maximal admissible value of a customer loss probability.

3 Process of system states and its stationary distribution

Let i_t, $i_t \geq 0$, be the number of customers in orbit, r_t, $r_t = \overline{1, R}$, be the state of the RE, n_t, $n_t = \overline{0, N^{(r_t)}}$, be the number of customers in service, ν_t, $\nu_t = \overline{0, W}$, be the state of the second component of the underlying process of customers arrivals at the moment t, $t \geq 0$.

It is easy to see that the process $\xi_t = \{i_t, r_t, n_t, \nu_t\}$, $t \geq 0$, is the four dimensional irreducible Markov chain.

We enumerate all states of the Markov chain ξ_t, $t \geq 0$, in the lexicographic order of the components (i, r, n, ν). Let us call the set of the states having value (i, r) of the two first components of the Markov chain as the macro-state (i, r).

Let $\mathbf{A}$ be the generator of the Markov chain ξ_t, $t \geq 0$. It is formed by the blocks $\mathbf{A}_{i,j}$, consisting of the matrices $(A_{i,j})_{r,r'}$ that define (except the diagonal entries of the matrix $\mathbf{A}_{i,i}$) the intensities of the transitions of the Markov chain ξ_t, $t \geq 0$, from the macro-state (i, r) to the macro-state (j, r'), r, $r = \overline{1, R}$. The diagonal entries of the matrix $\mathbf{A}_{i,i}$ are negative. The modulus of each element defines the intensity of departing from the corresponding state of the Markov chain ξ_t, $t \geq 0$.

Let us introduce the following notation.

- I is an identity matrix, O is a zero matrix of appropriate dimension;
- $\otimes$ is the symbol of Kronecker's product of matrices;
- $\bar{W} = W + 1$;
- $\tilde{N} = \max\{N^{(R)} - N^{(1)}, 1\}$;
- E_r^-, $r = \overline{1, R}$, is the square matrix of size $N^{(r)}+1$ with all zero entries except the subdiagonal entries $(E_r^-)_{n,n-1}$, $n = \overline{1, N^{(r)}}$, which are equal to 1;
- $\hat{I}_r$, $r = \overline{1, R}$, is the diagonal matrix of size $N^{(r)} + 1$ having the form $\hat{I}_r = \text{diag}\{0, \ldots, 0, 1\}$;
- E_r^+, $r = \overline{1, R}$, is the square matrix of size $N^{(r)}+1$ with all zero entries except the overdiagonal entries $(E_r^+)_{n,n+1}$, $n = \overline{0, N^{(r)} - 1}$, which are equal to 1;
- M_r, $r = \overline{1, R}$, is the diagonal matrix of size $N^{(r)} + 1$ having the form $M_r = \text{diag}\{n\mu_n^{(r)}$, $n = \overline{0, N^{(r)}}\}$;

- B_r, $r = \overline{1,R}$, is the diagonal matrix of size $N^{(r)}+1$ having the form $B_r = \text{diag}\{n\beta_n^{(r)},\ n = \overline{0,N^{(r)}}\}$;
- $P_{r,r'}$, $r = \overline{1,R}$, $r' = \overline{r+1,R}$, is the matrix of size $(N^{(r)}+1)\times(N^{(r')}+1)$ that has the form $P_{r,r'} = (I_{N^{(r)}+1}|O)$, i.e., $P_{r,r'}$ is obtained from the identity matrix $I_{N^{(r)}+1}$ by supplementing it from the right by zero matrix of the corresponding size;
- $p^{(r)}(n,k)$, $0 \le n \le k$, is the probability that n customers join the orbit and $k-n$ customers leave the system permanently when the state of the RE was r and k customers are forced to terminate service. This probability is defined as $p^{(r)}(n,k) = C_k^n(1-p^{(r)})^{k-n}(p^{(r)})^n,\ k = \overline{1,\tilde{N}}$;
- $Z_{r,r'}^{(n)}$, $r = \overline{1,R}$, $r' = \overline{1,r-1}$, $n = \overline{0,\tilde{N}}$, is the matrix of size $(N^{(r)}+1)\times(N^{(r')}+1)$ that has the following non-zero entries:

$$(Z_{r,r'}^{(0)})_{l,l} = 1,\ l = \overline{0,N^{(r')}},\ (Z_{r,r'}^{(n)})_{l,N^{(r')}} = p^{(r)}(n,l-N^{(r')}),\ l = \overline{N^{(r')}+1,N^{(r)}}.$$

Lemma 1. *The generator* $\mathbf{A}$ *has the following block structure:*

$$\mathbf{A} = \begin{pmatrix} \mathbf{A}_{0,0} & \mathbf{A}_{0,1} & \mathbf{A}_{0,2} & \dots & \mathbf{A}_{0,\tilde{N}} & O & O & \dots \\ \mathbf{A}_{1,0} & \mathbf{A}_{1,1} & \mathbf{A}_{1,2} & \dots & \mathbf{A}_{1,\tilde{N}} & \mathbf{A}_{1,\tilde{N}+1} & O & \dots \\ O & \mathbf{A}_{2,1} & \mathbf{A}_{2,2} & \dots & \mathbf{A}_{2,\tilde{N}} & \mathbf{A}_{2,\tilde{N}+1} & \mathbf{A}_{2,\tilde{N}+2} & \dots \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}. \tag{1}$$

The non-zero blocks $\mathbf{A}_{i,j}$, $i,j \ge 0$, *are defined as follows:*

- $\mathbf{A}_{i,i} = (\mathbf{A}_{i,i})_{r,r'}, r, r' = \overline{1,R}$, $i \ge 0$, *where*

$$(\mathbf{A}_{i,i})_{r,r} = I_{N^{(r)}+1}\otimes D_0^{(r)} + ((1-q^{(r)})\hat{I}_r + E_r^+)\otimes D_1^{(r)} - (M_r(I-E_r^-)+$$

$$+B_r(I-(1-a^{(r)})E_r^-) + i\gamma^{(r)}(I_{N^{(r)}+1} - q^{(r)}\hat{I}_r))\otimes I_{\bar{W}} + (H)_{r,r}I_{(N^{(r)}+1)\bar{W}},$$

$$(\mathbf{A}_{i,i})_{r,r'} = (H)_{r,r'}Z_{r,r'}^{(0)}\otimes I_{\bar{W}},\ r' < r,\ (\mathbf{A}_{i,i})_{r,r'} = (H)_{r,r'}P_{r,r'}\otimes I_{\bar{W}},\ r' > r.$$

- $\mathbf{A}_{i,i+n} = \mathbf{A}_n^+ = (\mathbf{A}_n^+)_{r,r'}, r, r' = \overline{1,R}$, $i \ge 0$, *where*

$$(\mathbf{A}_n^+)_{r,r} = \delta_{n,1}(q^{(r)}\hat{I}_r\otimes D_1^{(r)} + a^{(r)}B_rE_r^-\otimes I_{\bar{W}}),\ \ n = \overline{1,\tilde{N}},$$

$\delta_{i,j}$ indicates the Kronecker delta,

$$(\mathbf{A}_n^+)_{r,r'} = (H)_{r,r'}Z_{r,r'}^{(n)}\otimes I_{\bar{W}},\ r = \overline{1,R},\ r' = \overline{1,r-1},$$

$$(\mathbf{A}_n^+)_{r,r'} = O,\ r = \overline{1,R},\ r' = \overline{r+1,R},\ n = \overline{1,\tilde{N}}.$$

- $\mathbf{A}_{i,i-1} = \text{diag}\{i\gamma^{(r)}((1-q^{(r)})\hat{I}_R + E_r^+)\otimes I_{\bar{W}},\ r = \overline{1,R}\}$, $i \ge 1$.

Proof of the lemma is performed by means of analysis of the intensities of all possible transitions of the Markov chain ξ_t during the time interval having infinitesimal length. Existence of $\tilde{N}+1$ non-zero block diagonals in the matrix $\mathbf{A}$ is explained by the fact that the number $\tilde{N}$ (the maximal difference of the system capacities under various states of the RE, if this difference is not equal to zero, or 1, otherwise) defines the maximal number of customers that can join orbit simultaneously due to a customer arrival when the number of customers in service is equal to the server's capacity or due to the service forced termination caused by the reduction of the server capacity.

Analysis of the Markov chain having the generator $\mathbf{A}$ defined by Lemma 1 is non-trivial due to two essential reasons. The first reason is that the matrix $\mathbf{A}$ is not the block-tridiagonal. The second reason is that this matrix does not have Toeplitz-like structure, i.e. the form of the blocks $\mathbf{A}_{i,j}$ depends not only on the difference $j-i$ but depends on i and j separately. Fortunately, the Markov chains, which have the generator in form (1), are known in the literature.

Lemma 2. *The Markov chain ξ_t, $t \geq 0$, belongs to the class of continuous-time asymptotically quasi-Toeplitz Markov chains (AQTMC), see [10].*

To prove this lemma, it is required to verify that the following limits exist:

$$Y^{(0)} = \lim_{i\to\infty} R_i^{-1}\mathbf{A}_{i,i-1},\ Y^{(1)} = \lim_{i\to\infty} R_i^{-1}\mathbf{A}_{i,i} + I,$$

$$Y^{(n)} = \lim_{i\to\infty} R_i^{-1}\mathbf{A}_{i,i+n-1},\ n = \overline{2,\tilde{N}+1},$$

where R_i is a diagonal matrix with the diagonal entries which are defined as the moduli of the corresponding diagonal entries of the matrix $\mathbf{A}_{i,i}$, $i \geq 0$.

By the direct calculation of these limits, it is possible to show that they exist and the matrices $Y^{(n)}$, $n = \overline{0,\tilde{N}+1}$, have the following form:

- $Y^{(0)} = \mathrm{diag}\{Y_1^{(0)},\ldots,Y_R^{(0)}\}$, where

$$Y_r^{(0)} = \begin{cases} O, & \text{if } \gamma^{(r)} = 0, \\ E_r^+ \otimes I_{\bar{W}}, & \text{if } \gamma^{(r)} > 0,\ q^{(r)} = 1, \\ (E_r^+ + \hat{I}_r) \otimes I_{\bar{W}}, & \text{if } \gamma^{(r)} > 0,\ q^{(r)} \neq 1; \end{cases}$$

- $Y^{(1)} = (Y^{(1)})_{r,r'}, r, r' = \overline{1,R}$, where
  $(Y^{(1)})_{r,r'} = R_1^{(r)}(\mathbf{A}_{0,0})_{r,r'} + \delta_{r-r',0}\hat{I}_r \otimes I_{\bar{W}}$, if $q^{(r)} = 1, \gamma^{(r)} > 0$,
  $(Y^{(1)})_{r,r'} = R_2^{(r)}(\mathbf{A}_{0,0})_{r,r'} + \delta_{r-r',0} I_{N^{(r)}+1} \otimes I_{\bar{W}}$, if $\gamma^{(r)} = 0$,

$$(Y^{(1)})_{r,r'} = O, \text{ if } q^{(r)} \neq 1,\ \gamma^{(r)} > 0,\ r,\ r' = \overline{1,R},$$

$$R_1^{(r)} = \hat{I}_r \otimes (((\mu_{N^{(r)}}^{(r)} + \beta_{N^{(r)}}^{(r)})N^{(r)} - (H)_{r,r})I_{\bar{W}} + \Sigma_0^{(r)} - (1-q^{(r)})\Sigma_1^{(r)})^{-1},\ r = \overline{1,R},$$

$$R_2^{(r)} = ((M_r + B^{(r)}) \otimes I_{\bar{W}} + I_{N^{(r)}} \otimes (\Sigma_0^{(r)} -$$

$$-(H)_{r,r}I_{\bar{W}}) - (1-q^{(r)})\hat{I}_r \otimes \Sigma_1^{(r)})^{-1},\ r = \overline{1,R},$$

$$\Sigma_0^{(r)} = \mathrm{diag}\{(-D_0^{(r)})_{l,l},\ l = \overline{0,W}\},\ \Sigma_1^{(r)} = \mathrm{diag}\{(D_1^{(r)})_{l,l},\ l = \overline{0,W}\}.$$

- $Y^{(n)} = (Y^{(n)})_{r,r'}, r, r' = \overline{1, R}$, $n = \overline{2, \tilde{N}+1}$, where

$$(Y^{(n)})_{r,r'} = \begin{cases} R_1^{(r)} \mathbf{A}_{n-1}^+, & \text{if } q^{(r)} = 1,\ \gamma^{(r)} > 0, \\ R_2^{(r)} \mathbf{A}_{n-1}^+, & \text{if } \gamma^{(r)} = 0, \\ O, & \text{if } q^{(r)} \neq 1,\ \gamma^{(r)} > 0. \end{cases}$$

It is possible to verify that the matrices $Y^{(n)}$, $n = \overline{0, \tilde{N}+1}$, are sub-stochastic while their sum is the stochastic. According to definition of $AQTMC$, this means that the Markov chain ξ_t belongs to the class of $AQTMC$. Lemma 2 is proven.

Therefore, it is possible to apply the theory of $AQTMC$ from [10] for analysis of the Markov chain ξ_t. First of all, it is necessary to obtain the conditions on the system parameters which guarantee existence of the steady state distribution (the ergodicity) of the Markov chain ξ_t. According to [10], the sufficient condition of the ergodicity of the chain is the fulfillment of the following inequality:

$$\mathbf{y} Y^{(0)} \mathbf{e} > \mathbf{y} \sum_{n=2}^{\tilde{N}+1} (n-1) Y^{(n)} \mathbf{e} \tag{2}$$

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

$$\mathbf{y} \sum_{n=0}^{\tilde{N}+1} Y^{(n)} = \mathbf{y},\ \mathbf{y}\mathbf{e} = 1. \tag{3}$$

If the customers from orbit are persistent for all states of the RE, i.e. $q^{(r)} = 1$, $r = \overline{1, R}$, the procedure for verifying the existence of the steady state distribution is the following. Because system (3) is the finite system of the linear algebraic equations, its solving on computer is not the difficult task. By substituting the obtained solution to inequality (2), one can easily check whether or not the steady state distribution exist under these values of the system parameters.

If customers from orbit non-persistent ($q^{(r)} \neq 1$) at least for one state r of the RE such as $\gamma^{(r)} \neq 0$, the following statement is true.

Lemma 3. *If customers from orbit not absolutely persistent ($q^{(r)} \neq 1$) at least for one state r of the RE such as $\gamma^{(r)} \neq 0$, then the Markov chain ξ_t is ergodic for any set of the system parameters.*

Proof is implemented by analogy with Theorem 2 from [11].

In the sequel, we assume that the ergodicity condition is fulfilled. Then, the following stationary probabilities exist:

$$\pi(i, r, n, \nu) = \lim_{t \to \infty} P\{i_t = i,\ r_t = r, n_t = n,\ \nu_t = \nu\},$$

$$i \geq 0,\ r = \overline{1, R},\ n = \overline{0, N^{(r)}}, \nu = \overline{0, W}.$$

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

$$\boldsymbol{\pi}(i, r, n) = (\pi(i, r, n, 0), \pi(i, r, n, 1), \ldots, \pi(i, r, n, W)),$$

$$\boldsymbol{\pi}(i,r) = (\boldsymbol{\pi}(i,r,0), \boldsymbol{\pi}(i,r,1), \dots, \boldsymbol{\pi}(i,r,N^{(r)})),\ r = \overline{1,R},$$
$$\boldsymbol{\pi}_i = (\boldsymbol{\pi}(i,1), \boldsymbol{\pi}(i,2), \dots, \boldsymbol{\pi}(i,R)),\ i \geq 0.$$

It is well known that the vectors $\boldsymbol{\pi}_i$, $i \geq 0$, satisfy the system

$$(\boldsymbol{\pi}_0, \boldsymbol{\pi}_1, \dots)\mathbf{A} = \mathbf{0},\ \ (\boldsymbol{\pi}_0, \boldsymbol{\pi}_1, \dots)\mathbf{e} = 1. \tag{4}$$

System (4) is infinite. Therefore, its solution is a difficult problem. However, this problem can be successfully solved using the numerically stable algorithm developed in [10].

4 Performance measures

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

The average number of customers in the service area is

$$N = \sum_{i=0}^{\infty} \sum_{r=1}^{R} \sum_{n=0}^{N^{(r)}} n\boldsymbol{\pi}(i,r,n)\mathbf{e}.$$

The average number of customers in orbit is $L = \sum_{i=1}^{\infty} i\boldsymbol{\pi}_i\mathbf{e}$.

The intensity of output of successfully serviced customers is

$$\lambda_{out} = \sum_{i=0}^{\infty} \sum_{r=1}^{R} \sum_{n=1}^{N^{(r)}} n\mu_n^{(r)}\boldsymbol{\pi}(i,r,n)\mathbf{e}.$$

The intensity of flow of customers who leaves the server due to impatience is

$$\lambda_{imp} = \sum_{i=0}^{\infty} \sum_{r=1}^{R} \sum_{n=1}^{N^{(r)}} n\beta_n^{(r)}\boldsymbol{\pi}(i,r,n)\mathbf{e}.$$

The probability that a customer arrives to the system when the server already reached its capacity and leaves the system is $P_{ent} = \lambda^{-1} \sum_{i=0}^{\infty} \sum_{r=1}^{R} (1 - q^{(r)})\boldsymbol{\pi}(i,r,N^{(r)})D_1^{(r)}\mathbf{e}$.

The probability that a customer leaves the system forever due to impatience is $P_{imp} = \lambda^{-1} \sum_{i=0}^{\infty} \sum_{r=1}^{R} (1 - a^{(r)}) \sum_{n=1}^{N^{(r)}} n\beta_n^{(r)}\boldsymbol{\pi}(i,r,n)\mathbf{e}$.

The probability of customers loss due to the decrease of the number of servers caused by change of the state of the RE is $P_{RE} = \frac{1}{\lambda} \sum_{i=0}^{\infty} \sum_{r=2}^{R} \sum_{r'=1}^{r-1} (1 - p^{(r)})(H)_{r,r'} \sum_{n=N^{(r')}+1}^{N^{(r)}} (n - N^{(r')})\boldsymbol{\pi}(i,r,n)\mathbf{e}$.

The probability that an arbitrary customer from orbit makes an attempt to receive service when the server capacity is exhausted and permanently leaves the system is $P_{retry} = \lambda^{-1} \sum_{i=1}^{\infty} \sum_{r=1}^{R} i\gamma^{(r)}(1 - q^{(r)})\boldsymbol{\pi}(i,r,N^{(r)})\mathbf{e}$.

The loss probability of an arbitrary customer is

$$P_{loss} = 1 - \frac{\lambda_{out}}{\lambda} = P_{retry} + P_{imp} + P_{ent} + P_{RE}. \tag{5}$$

5 Numerical example

Let us consider the queueing system that operates in the RE having three states ($R = 3$) with the generator $H = \begin{pmatrix} -0.06 & 0.04 & 0.02 \\ 0.0002 & -0.0005 & 0.0003 \\ 0.0004 & 0.0005 & -0.0009 \end{pmatrix}$ and stationary distribution given by the vector $\boldsymbol{\psi} = (0.00439, 0.6735, 0.32211)$.

We assume that the server doesn't work (is broken, takes vacation, etc.) during the RE's stay in state 1.

Under state 2 of the RE, the server can serve up to 10 customers simultaneously. When there are n customers on service, the service intensity of one customer is determined as $\mu_n^{(2)} = \frac{10.0-0.3n}{n}$, $n = \overline{1, 10}$.

Under state 3 of the RE, the server can serve up to 15 customers simultaneously. When there are n customers on service, the service intensity of one customer is determined as $\mu_n^{(3)} = \frac{20.0-0.3n}{n}$, $n = \overline{1, 15}$.

The individual intensities of impatience are given by $\beta_n^{(2)} = 0.03n$, $\beta_n^{(3)} = 0.04n$.

To define the arrival arrival flow under various states of the RE, let us consider the MAP arrival flow that defined by the matrices

$$D_0 = \begin{pmatrix} -1.35164 & 0 \\ 0 & -0.04387 \end{pmatrix}, D_1 = \begin{pmatrix} 1.34265 & 0.00899 \\ 0.024435 & 0.019435 \end{pmatrix}.$$

This arrival flow has the coefficient of correlation $c_{cor} = 0.2$ and the coefficient of variation $c_{var} = 12.34$.

We assume that under state 1 of the RE the arrival flow is defined by the matrices $D_0^{(1)} = 2D_0$ and $D_1^{(1)} = 2D_1$, under state 2 of the RE the arrival flow is defined by the matrices $D_0^{(2)} = 5D_0$ and $D_1^{(2)} = 5D_1$, under state 3 of the RE the arrival flow is defined by the matrices $D_0^{(3)} = 8D_0$ and $D_1^{(3)} = 8D_1$. The intensities of arrivals are $\lambda^{(1)} = 2$, $\lambda^{(2)} = 5$, and $\lambda^{(3)} = 8$, correspondingly.

The rest of the system parameters are as follows: $q^{(1)} = 0.95$, $a^{(1)} = 0.5$, $\gamma^{(1)} = 0.2$; $q^{(2)} = 0.9$, $a^{(2)} = 0.4$, $p^{(2)} = 0.6$, $\gamma^{(2)} = 0.2$; $q^{(3)} = 0.9$, $a^{(3)} = 0.4$, $p^{(3)} = 0.7$, $\gamma^{(3)} = 0.2$.

As the main performance measure of the system, we will consider the loss probability of an arbitrary customer P_{loss}. The goal of the experiment is to find the values of the server capacities $N^{(2)}$ and $N^{(3)}$ which provide the minimal value of this probability. To this end, we will compute the values of this probability for various combinations of $N^{(2)}$ and $N^{(3)}$ from the set $N^{(2)} = \overline{1, \min\{N^{(3)}, 10\}}$ and $N^{(3)} = \overline{1, 15}$. Note, that because service is not provided when the RE stays in state 1 we have $N^{(1)} = 0$.

It is clear that the problem of choosing the optimal combination of $N^{(2)}$ and $N^{(3)}$ is not trivial. If these values are chosen be small, the probabilities of customers loss upon arrival P_{ent} or retrial P_{retry} may be high. This is confirmed by Figure 2.

If the values of $N^{(2)}$ and $N^{(3)}$ are chosen be large, the probabilities of customers loss upon arrival or retrial essentially decrease. However the probabilities

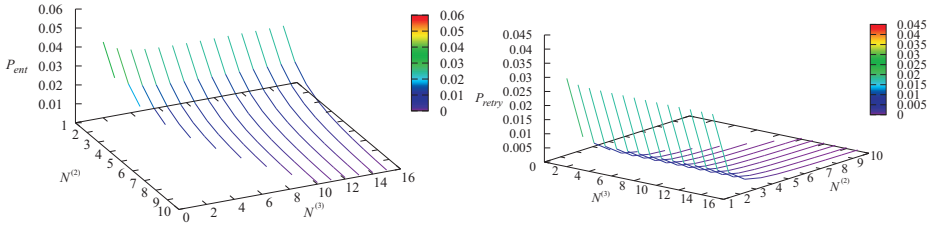

Fig. 2. Dependence of P_{ent} and P_{retry} on $N^{(2)}$ and $N^{(3)}$

of customers loss due impatience P_{imp} and due to the decrease of the server capacity P_{RE} grow. This is confirmed by Figure 3. The loss probability P_{loss} of an arbitrary customer is the sum of the probabilities P_{ent}, P_{retry}, P_{imp}, and P_{RE}. The surface giving dependence of P_{loss} on $N^{(2)}$ and $N^{(3)}$ is presented in Figure 4.

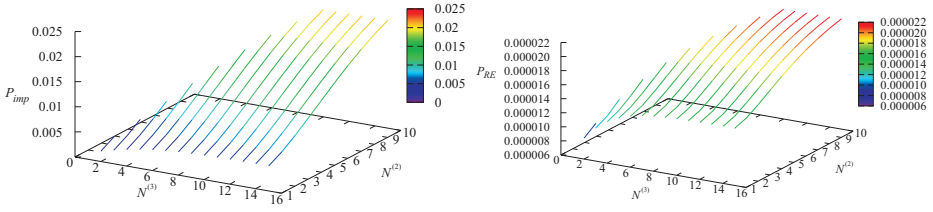

Fig. 3. Dependence of P_{imp} and P_{RE} on $N^{(2)}$ and $N^{(3)}$

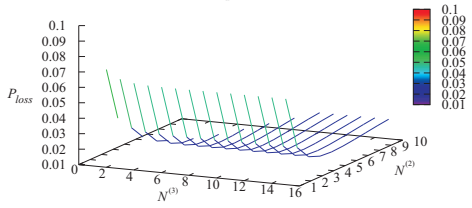

Fig. 4. Dependence of P_{loss} on $N^{(2)}$ and $N^{(3)}$

It is evidently seen from Figure 4, that there exit the points $(N^{(2)}, N^{(3)})$ providing some trade-off in situation when the summands in expression (5) for probability P_{loss} demonstrate the opposite behavior when $N^{(2)}$ and $N^{(3)}$ increase and decrease. The minimal value of the loss probability $P_{loss} = 0.0177847$ is reached for $N^{(2)} = 5$ and $N^{(3)} = 7$. If we do not control capacity of the server and accept all customers for each state of the RE (up to 10 when the RE is in state 2 and up to 15 when the RE is in state 3), then $P_{loss} = 0.022658$. Therefore, admission for simultaneous service of less customers than the maximally possible allows essentially decrease the customer loss probability.

6 Conclusion

A retrial queueing system with limited processor sharing discipline and impatient customers, which operates in the RE, is analysed. An arbitrary dependence of the individual service and impatience rates on the number of customers in service is allowed. The behavior of the system is described by the multi-dimensional asymptotically quasi-Toeplitz Markov chain. Expressions for key performance measures of the system are presented. Feasibility of the described algorithmic results is numerically illustrated. It is shown that the results can be used for the optimal adjustment of capacity of the server at each state of the RE.

7 Acknowledgments

The publication was financially supported by the Ministry of Education and Science of the Russian Federation (the Agreement number 02.a03.21.0008) and by the Belarusian Republican Foundation for Fundamental Research (grant F16MV-003).

References

1. Samouylov, K., Naumov, V., Sopin, E., Gudkova, I., Shorgin S.: Sojourn time analysis for processor sharing loss system with unreliable server. Lecture Notes in Computer Science. 9247, 284-297 (2016)
2. Samouylov, K.E., Sopin, E.S., and Gudkova, I.A. Sojourn Time Analysis for Processor Sharing Loss Queuing System with Service Interruptions and MAP Arrivals. Communications in Computer and Informational Science. 678, 406-417, (2017)
3. Yashkov, S.: Processor-sharing queues: some progress in analysis. Queuing Systems. 2, 1-17 (1987)
4. Yashkov, S., Yashkova, A.: Processor sharing: a survey of the mathematical theory. Automation and Remote Control. 68, 1662-1731 (2007)
5. Artalejo J. R., Gomez-Corral, A.: Retrial queueing systems: A computational approach. Berlin-Heidelberg: Springer. (2008)
6. Nunez-Queija, R.: Sojourn times in a processor sharing queue with service interruptions. Queueing Systems. 34, 351-386 (2000)
7. Nunez-Queija, R.: Sojourn times in non-homogeneous QBD processes with processor sharing. Stochastic Models. 17. 61-92 (2001)
8. Ghosh, A, Banik, A.D.: An algorithmic analysis of the $BMAP/MSP/1$ generalized processor-sharing queue. Computers and Operations Research. 79, 1-11 (2017)
9. Lucantoni, D.: New results on the single server queue with a batch Markovian arrival process. Communication in Statistics-Stochastic Models. 7, 1-46 (1991)
10. Klimenok, V. Dudin, A.: Multi-dimensional asymptotically quasi-Toeplitz Markov chains and their application in queueing theory, Queueing Systems. 54, 245–259 (2006)
11. Dudin, A., Kim, C., Dudin, S., Dudina, O. Priority retrial queueing model operating in random environment with varying number and reservation of servers. Applied Mathematics and Computation. 269, 674-690 (2015)