Retrial Tandem Queue with $BMAP$-Input and Semi-Markovian Service Process

Valentina Klimenok[1(✉)], Olga Dudina[1], Vladimir Vishnevsky[2], and Konstantin Samouylov[3]

[1] Department of Applied Mathematics and Computer Science,
Belarusian State University, 220030 Minsk, Belarus
klimenok@bsu.by, taramin@mail.ru
[2] Institute of Control Sciences of Russian Academy of Sciences, Moscow, Russia
vishn@inbox.ru
[3] Peoples' Friendship University of Russia, Moscow, Russia
ksam@sci.pfu.edu.ru

Abstract. We consider a tandem queueing system consisting of two stations. The input flow at the single-server first station is described by a $BMAP$ (batch Markovian arrival process). If a customer from this flow meets the busy server, it goes to the orbit of infinite size and tries its luck later on in exponentially distributed random time. The service time distribution at the first station is assumed to be semi-Markovian. After service at the first station a customer proceeds to the second station which is described by a multi-server queue without a buffer. The service time by the server of the second station is exponentially distributed. We derive the condition for the stable operation of the system and determine the stationary distribution of the system states. Some key performance measures are calculated and illustrative numerical results are presented.

Keywords: Tandem retrial queue · Batch Markovian arrival process · Semi-Markovian service process · Asymptotically quasi-Toeplitz Markov chain

1 Introduction

Retrial queues are good mathematical models for many telecommunication networks such as telephone switching systems, cellular mobile networks, local area networks under the protocol of random multiple access, etc. So, despite their complexity, retrial queueing systems are popular object for investigations. They have been extensively studied under a variety of scenarios, for references see, e.g., survey [1] and books [2,3].

The analysis of current situation makes clear the great importance of tandem retrial queueing system with a batch Markovian arrival process ($BMAP$). Tandem queues can be used for modeling real-life networks of linear topology as well as for the validation of general decomposition algorithms in networks, see, e.g., [4–6]. Thus, tandem queueing systems have found much interest in the

literature. An extensive survey of early papers on tandem queues can be seen in [7]. Most of these papers are devoted to exponential queueing models. Over the last three decades or so, the efforts of many investigators in tandem queues were in weakening the distribution assumptions on the service times as well as on the arrivals. In particular, the arrival process should be able to capture any correlation and burstiness that are commonly seen in the traffic of modern communication networks [6]. Such an arrival process was introduced in [8] and ever since this process is referred to as a batch Markovian arrival process ($BMAP$). $BMAP$ includes many input flows considered previously, such as stationary Poisson, Erlangian, Hyper-Markovian, Phase-Type (PH) renewal process, Markov Modulated Poisson Process ($MMPP$) and their superpositions. Tandem queues with the $BMAP$ input or its ordinary counterparts were considered in [9–16].

At the same time, to the best of our knowledge, only the papers [13,16,17] are devoted to investigation of tandem queues with $BMAP$ input where the effect of retrials is taken into account. The paper [13] considers the $MAP/PH/1 \rightarrow \cdot/PH/1/K+1$ tandem retrial queue. The paper [16] deals with a tandem retrial queue with two Markovian flows and reservation of servers on the second station is studied. The paper [17] is devoted to the tandem queue $BMAP/G/1 \rightarrow \cdot/M/N/0$ with retrials and group occupation of servers of the second station. Here the customers receive service individually, and upon completion of a service the customer's type is determined. This type identification is necessary to determine the nature of service, if any, offered at station 2. The customer's type is classified based on the number of servers (resources) required to process the request of the customer. The simultaneous initiation or occupation of several servers to a customer's request is typical for the so called non-elastic traffic in communication networks.

In this paper, we consider a generalization of the model [17] to the case of correlated semi-Markovian service process at the first station. Unlike the system [17], where service times at the first station are independent identically distributed random variables, in this paper we assume that these times can be significantly dependent and distributed according to different laws. This assumption makes the system under consideration more adequate model of real-life systems and processes but more difficult for analytical investigation.

The rest of the paper is organized as follows. In Sect. 2, the mathematical model is described. In Sect. 3, the embedded Markov chain is investigated, the condition for existence is derived, and the stationary distribution of the system states at the service completion epochs at the first station is calculated. The stationary distribution at an arbitrary time is derived in Sect. 4. The system performance measures are given in Sect. 5. In Sect. 6, the numerical examples illustrating the behavior of the system characteristics depending on its parameters are presented. Finally, the conclusions are given in Sect. 7.

2   Mathematical Model

We consider tandem queue consisting of two stations.

The first station has a single server. Customers arrive at the first station in batches. The input flow of batches is described by the $BMAP$. The $BMAP$ is defined by the underlying process ν_t, $t \geq 0$, which is an irreducible continuous time Markov chain with state space $\{0, ..., W\}$, and with the matrix generating function $D(z) = \sum\limits_{k=0}^{\infty} D_k z^k$, $|z| \leq 1$. Arrivals occur only at epochs of the jumps in the underlying process ν_t, $t \geq 0$. The intensities of the transitions of the process ν_t accompanied by the arrival of a batch of size k are defined by the matrices D_k, $k \geq 0$. The matrix $D(1)$ is the infinitesimal generator of the process ν_t. The stationary distribution vector $\boldsymbol{\theta}$ of this process satisfies the equations $\boldsymbol{\theta} D(1) = \mathbf{0}, \boldsymbol{\theta}\mathbf{e} = 1$, where $\mathbf{e}$ is a column vector consisting of 1's, and $\mathbf{0}$ is a row vector of 0's. The average intensity λ (fundamental rate) of the $BMAP$ is given by $\lambda = \boldsymbol{\theta} D'(z)|_{z=1}\mathbf{e}$. The average intensity λ_b of group arrivals is defined by $\lambda_b = \boldsymbol{\theta}(-D_0)\mathbf{e}$. The coefficient of variation c_{var} of intervals between successive group arrivals is defined by $c_{var}^2 = 2\lambda_b\boldsymbol{\theta}(-D_0)^{-1}\mathbf{e} - 1$. The coefficient of correlation c_{cor} of the successive intervals between group arrivals is given by $c_{cor} = (\lambda_b\boldsymbol{\theta}(-D_0)^{-1}(D(1) - D_0)(-D_0)^{-1}\mathbf{e} - 1)/c_{var}^2$. For more information about the $BMAP$, its history and properties see, e.g., [8].

If the arriving batch of primary customers meets a free first station server upon arrival, one customer automatically starts a service and the rest of the batch go to so called orbit. If the server is busy at an arrival epoch, all customers of the batch go to the orbit. From the orbit, they try their luck later on after a random amount of time. We assume that the total flow of retrials is such as the probability of generating a retrial attempt in the interval $(t, t + \Delta t)$ is equal to $\alpha_i \Delta t + o(\Delta t)$ when the number of customers in the orbit is equal to i, $i > 0$, $\alpha_0 = 0$. The orbit capacity is supposed to be unlimited. We do not fix the explicit dependence of the intensity α_i on i assuming only that $\lim\limits_{i \to \infty} \alpha_i = \infty$. Note that such dependence describes the classic retrial strategy ($\alpha_i = i\alpha$) and the linear strategy ($\alpha_i = i\alpha + \gamma$, $\alpha > 0$) as special cases.

The service time of primary and repeated customers at the first station is governed by the semi-Markovian process m_t, $t \geq 0$. It is characterized by the state space $\{1, ..., M\}$ and the semi-Markovian kernel $B(t) = (B_{m,m'}(t))_{m,m'=\overline{1,M}}$. The successive service times of customers are defined as the sojourn times of the process m_t, $t \geq 0$, in its states. The average service time is calculated as $b_1 = \boldsymbol{\delta} \int\limits_0^{\infty} t dB(t)\mathbf{e}$ where $\boldsymbol{\delta}$ is the unique solution to the system $\boldsymbol{\delta} B(\infty) = \boldsymbol{\delta}, \boldsymbol{\delta}\mathbf{e} = 1$.

After receiving service at the first station a customer proceeds to the second station which is represented by N independent identical servers. The service time by a server is exponentially distributed with the parameter $\mu > 0$. The service of an arbitrary customer at the second station requires a random number η of servers. Here η is an integer-valued random variable with the distribution

$q_n = P\{\eta = n\}$, $q_n \geq 0$, $n = \overline{0, N}$, $\sum_{n=0}^{N} q_n = 1$. No queue is allowed between the first and the second station.

In case a customer completes the service at the first station and does not see required number of free servers at the second station, it leaves the system forever with the probability p, $0 \leq p \leq 1$. With the probability $1 - p$ the customer waits until the required number of servers of the second station becomes free and then occupies these servers immediately. The waiting period is accompanied by blocking the first station server operating.

For further use in the sequel, we introduce the following notation:

- I is an identity matrix of appropriate dimension. When needed the dimension of the matrix will be identified with a suffix;

- $\otimes$ and $\oplus$ are symbols of the Kronecker product and sum of matrices, see, e.g., [18];

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

- $P(n, t)$, $n \geq 0$, are coefficients of the matrix expansion $e^{D(z)t} = \sum_{n=0}^{\infty} P(n, t) z^n$, $|z| \leq 1$. The (ν, ν')th entry of the matrix $P(n, t)$ defines the probability that n customers arrive in the $BMAP$ during the interval $(0, t]$ and the state of the underlying process of the $BMAP$ at the epoch t is ν' given $\nu_0 = \nu$, $\nu, \nu' = \overline{0, W}$;

- $\tilde{D}_k = D_k \otimes I_M$, $\hat{D}_k = I_{N+1} \otimes \tilde{D}_k$, $k \geq 0$, $\hat{D}(z) = \sum_{k=0}^{\infty} \hat{D}_k z^k$, $|z| \leq 1$;

- $H(t) = (H_{r,r'}(t))_{r,r'=\overline{0,N}}$, where $H_{r,r'}(t) = 0$ for $r \leq r'$ and, for $r > r'$, $H_{r,r'}(t)$ is the distribution function with the Laplace-Stieltjes transform $h_{r,r'}(s) = \prod_{l=r'+1}^{r} l\mu(l\mu + s)^{-1}$;

- Q_m, $m = \overline{1,3}$, are square matrices:

$$Q_1 = \begin{pmatrix} q_0 & q_1 & \cdots & q_N \\ 0 & q_0 & \cdots & q_{N-1} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & q_0 \end{pmatrix}, \quad Q_3 = \begin{pmatrix} 0 & \cdots & 0 & q_N \\ 0 & \cdots & 0 & q_{N-1} \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & 0 & q_0 \end{pmatrix},$$

$$Q_2 = diag\{\sum_{n=N-r+1}^{N} q_n, r = \overline{0, N}\};$$

- $\tilde{Q}_m = Q_m \otimes I_{\bar{W}}$, $\hat{Q}_m = \tilde{Q}_m \otimes I_M$, $m = \overline{1,3}$; $\quad \bar{Q} = \hat{Q}_1 + p\hat{Q}_2$;
- $Q = \tilde{Q}_1 + p\tilde{Q}_2 + (1-p)E\tilde{Q}_3$,
- $E = \bar{I}_{N+1} \otimes I_M$, where $\bar{I}_{\bar{W}}$ is a square matrix of size $N + 1$ whose below-diagonal entries are equal to 1 and the rest entries are zeroes.

3   The Stationary Distribution of the Embedded Markov Chain

Let t_n denote the time of the nth service completion at the first station. Consider the process

$$\xi_n = \{i_n, r_n, \nu_n, m_n\}, n \geq 1,$$

where

- i_n is the number of customers in the orbit at the epoch t_n, $i_n \geq 0$;
- r_n is the number of busy servers at the second station at the epoch $t_n - 0$, $r_n = \overline{0, N}$;
- ν_n is the state of the $BMAP$ underlying process at the epoch t_n, $\nu_n = \overline{0, W}$;
- m_n is the state of the service directing process m_t at the epoch $t_n + 0$, $m_n = \overline{1, M}$.

It is easy to see that the process ξ_n, $n \geq 1$, is a four-dimensional Markov chain which describes the process of the system operation at the service completion epochs.

Enumerate the states of the chain ξ_n, $n \geq 1$, in the lexicographic order and form the square matrices $P_{i,l}$, $i, l \geq 0$, of size $(N+1)\bar{W}M$ of transition probabilities of the chain from the states having the value i of the first component to the states having the value l of this component.

Lemma 1. The transition probability matrices $P_{i,l}$ are defined as follows:

$$P_{i,l} = 0, l < i - 1, i > 1,$$

$$P_{i,l} = \bar{Q}A_i[\alpha_i\Omega_{l-i+1} + \sum_{k=1}^{l-i+1}\hat{D}_k\Omega_{l-i-k+1}] + (1-p)[\sum_{n=0}^{l-i+1}\mathcal{H}_n\hat{Q}_3A_{i+n}\alpha_{i+n}\Omega_{l-i-n+1}$$

$$+ \sum_{n=0}^{l-i}\mathcal{H}_n\hat{Q}_3A_{i+n}\sum_{k=1}^{l-i-n+1}\hat{D}_k\Omega_{l-i-n-k+1}], \quad l \geq max\{0, i-1\}, i \geq 0, \tag{1}$$

$$A_i = \int_0^\infty e^{-\alpha_i t}e^{(\Delta \oplus \tilde{D}_0)t}dt \otimes I_M = (\alpha_i I - \Delta \oplus \tilde{D}_0)^{-1} \otimes I_M, i \geq 0,$$

$$\Omega_n = \int_0^\infty e^{\Delta t} \otimes P(n,t) \otimes dB(t), \quad \mathcal{H}_n = \int_0^\infty dH(t) \otimes P(n,t) \otimes I_M, n \geq 0,$$

$$\Delta = \begin{pmatrix} 0 & 0 & 0 & \cdots & 0 & 0 \\ \mu & -\mu & 0 & \cdots & 0 & 0 \\ 0 & 2\mu & -2\mu & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & N\mu & -N\mu \end{pmatrix}.$$

Proof. Formula (1) becomes clear if we take into account the probabilistic interpretation of the matrices which appear in the right hand side of (1).

The matrix $\bar{Q}$ defines probabilities that a customer served at the first station finds the required number of idle servers at the second station and occupies them or does not find the required number of idle servers and leaves the system.

The matrix Δ is an infinitesimal generator of the death process which describes the evolution of the number of occupied servers at the second station between two consecutive service completion epochs at the first station.

The matrix $A_i \alpha_i$ defines probabilities that, given i customers stay in the orbit after the service completion epoch at the first station, the next service at this station will be initiated by a customer from the orbit. The matrix $A_i \hat{D}_k$ has the analogous probabilistic sense with the only difference that the next service at the first station is initiated by a primary customer arriving in k-size batch.

The matrices Ω_n and $\mathcal{H}_n$ defines probabilities that during the service time at the first station and during the blocking time, respectively, n customers arrive into the system.

The matrix $(1-p)\mathcal{H}_n \tilde{Q}_3$ defines probabilities that a customer served at the first station does not find the required number of idle servers at the second station, causes the blocking of the first station server and during the blocking time n customers arrive into the system.

Using the above probabilistic interpretations and the total probability formula, we get expression (1) for the transition probability matrices.

Lemma 1 is proved.

It is seen from (1) that transition probability matrices $P_{i,l}$ depend on i and l and this dependence can not be reduced to the dependence on the difference $l-i$ only. It means that the Markov chain ξ_n, $n \geq 1$, is a level dependent one. At the same time the dependence of i vanishes when i tends to ∞ and the matrices $P_{i,l}$ approach to matrices that depend on the values i and l only via the difference $l-i$. It implies that the chain under consideration belongs to the class of asymptotically quasi-Toeplitz Markov chains (AQTMC), see [19]. So, the further investigation of the ξ_n, $n \geq 1$, will be based on the results given in [19].

Let us denote

$$Y_k = \lim_{i \to \infty} P_{i,i+k-1}, \; k \geq 0, \tag{2}$$

and let $Y(z)$ be the generating function of the matrices Y_k, $k \geq 0$. The matrices $\tilde{Y}_k$, $k \geq 0$, can be considered as transition probability matrices of a quasi-Toeplitz Markov chain $\tilde{\xi}_n$, $n \geq 1$, with the same state space as the chain ξ_n, $n \geq 1$. The chain $\tilde{\xi}_n$, $n \geq 1$, is called as limiting chain relative to the chain ξ_n, $n \geq 1$.

Corollary 1. The Markov chain ξ_n, $n \geq 1$, belongs to the class of asymptotically quasi-Toeplitz Markov chain. The generating function of its limiting chain transition probability matrices has form

$$Y(z) = [\bar{Q} + (1-p)\mathcal{H}(z)\hat{Q}_3]\Omega(z), \tag{3}$$

where

$$\Omega(z) = \sum_{n=0}^{\infty} \Omega_n z^n = \int_0^{\infty} e^{\Delta t} \otimes e^{D(z)t} \otimes dB(t),$$

$$\mathcal{H}(z) = \sum_{n=0}^{\infty} \mathcal{H}_n z^n = \int_0^{\infty} dH(t) \otimes e^{D(z)t} \otimes I_M, \; |z| \leq 1.$$

Proof. It is seen from (1) that transition probability matrices $P_{i,l}$ depend on i and l and this dependence can not be reduced to the dependence on the difference $l - i$ only. It means that the Markov chain ξ_n, $n \geq 1$, is a level dependent one. At the same time the dependence of i vanishes when i tends to ∞ and the matrices $P_{i,l}$ approach to matrices that depend on the values i and l only via the difference $l - i$. It implies that the chain under consideration belongs to the class of asymptotically quasi-Toeplitz Markov chains (AQTMC), see [19].

Using (2) and Lemma 1 we get the expression (3) for the generating function $Y(z)$.

In what follows we will use the results for asymptotically quasi-Toeplitz Markov chains given in [19] to derive the ergodicity condition and calculate the stationary distribution.

Theorem 1. The sufficient condition for ergodicity of the Markov chain ξ_n, $n \geq 1$, is the fulfillment of the inequality

$$\rho = \lambda[b_1 + (1-p)\boldsymbol{y} \int_0^{\infty} tdH(t)Q_3\mathbf{e}] < \mathbf{1}, \tag{4}$$

where the vector $\boldsymbol{y}$ is the unique solution of the system

$$\boldsymbol{y}Q^- \int_0^{\infty} e^{\Delta t} \otimes d\tilde{B}(t) = \boldsymbol{y}, \;\; \boldsymbol{y}\mathbf{e} = 1. \tag{5}$$

$$Q^- = Q_1 + pQ_2 + (1-p)EQ_3, \;\; \tilde{B}(t) = \boldsymbol{\delta}B(t)\mathbf{e}.$$

Proof. The matrix $Y(1)$ is an irreducible one. So, as follows from [19], the sufficient condition for ergodicity of the chain ξ_n, $n \geq 1$, is the fulfillment of the inequality

$$\boldsymbol{x}Y'(1)\mathbf{e} < 1, \tag{6}$$

where $\boldsymbol{x}$ is the unique solution of the system

$$\boldsymbol{x}Y(1) = \boldsymbol{x}, \;\; \boldsymbol{x}\mathbf{e} = 1. \tag{7}$$

Let the vector $\boldsymbol{x}$ be of the form

$$\boldsymbol{x} = \boldsymbol{y} \otimes \boldsymbol{\theta} \otimes \boldsymbol{v} \tag{8}$$

where $\boldsymbol{\theta}$ is the vector of the stationary distribution of the $BMAP$, $\boldsymbol{y}$ and $\boldsymbol{v}$ some stochastic vectors of size $N + 1$ and M, respectively, and the vector $\boldsymbol{y}$ is the unique solution of system (5).

It is verified by the direct substitution that the vector $\boldsymbol{x}$ of form (8) is the unique solution of system (7) and $\boldsymbol{v} = \boldsymbol{\delta}$. The last equation means that the vector $\boldsymbol{v}$ coincides with the stationary distribution vector of the service process $m_t, t \geq 0$.

We differentiate (3) at the point $z = 1$ and substitute the obtained expression for $Y'(1)$ and the vector $\boldsymbol{x}$ of form (8) into inequality (6). Then using that the vector $\boldsymbol{v} = \boldsymbol{\delta}$, we derive inequality (4).

Theorem 1 is proved.

Remark 1. Inequality (7) is intuitively clear on noting that the vector $\boldsymbol{y}$ gives the stationary distribution of the number of busy servers at the second station given the server of the first station works under overload conditions. Then $\boldsymbol{y} \int\limits_0^\infty t\, dH(t) Q_3 \mathbf{e}$ defines the average blocking time of the server of the first station under overload condition and ρ is the system load.

In what follows we suppose that inequality (4) is fulfilled.

Denote the stationary state probabilities of the Markov chain $\xi_n = \{i_n, r_n, \nu_n, m_n\}, n \geq 1$, by

$$\pi(i, r, \nu, m), \; i \geq 0, \; r = \overline{0, N}, \; \nu = \overline{0, W}, \; m = \overline{1, M}.$$

Form the row vectors of these probabilities

$$\boldsymbol{\pi}(i, r, \nu) = (\pi(i, r, \nu, 1), \dots, \pi(i, r, \nu, M)),$$

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

$$\boldsymbol{\pi}_i = (\boldsymbol{\pi}(i, 0), \boldsymbol{\pi}(i, 1), \dots, \boldsymbol{\pi}(i, N)), i \geq 0.$$

To calculate the vectors $\boldsymbol{\pi}_i, \; i \geq 0$, we use the numerically stable algorithm (see [19]) which has been elaborated for calculating the stationary distribution of the multi-dimensional asymptotically quasi-Toeplitz Markov chain.

The algorithm is based on censoring technique and asymptotic properties of the chain under consideration. It consists of the next principal steps:

1. Calculate the matrix G as the minimal nonnegative solution of the matrix equation $G = Y(G)$;
2. For preassigned sufficiently large integer i_0 calculate the matrices G_{i_0-1}, $G_{i_0-2}, \dots, G_0$ using the equation of the backward recursion

$$G_i = (I - \sum_{l=i+1}^{\infty} P_{i+1,l} G_{l-1} G_{l-2} \dots G_{i+1})^{-1} P_{i+1,i}, \; i = i_0 - 1, i_0 - 2, \dots, 0,$$

with the boundary condition $G_i = G, \; i \geq i_0$.

3. Calculate the matrices $\Phi_l, l \geq 1$, using recurrent formulas

$$\Phi_l = (\bar{P}_{0,l} + \sum_{i=1}^{l-1} \Phi_i \bar{P}_{i,l})(I - \bar{P}_{l,l})^{-1}, l \geq 1;$$

4. Calculate the vector $\boldsymbol{\pi}_0$ as the unique solution to the system

$$\boldsymbol{\pi}_0(I - \bar{P}_{0,0}) = \mathbf{0}, \ \boldsymbol{\pi}_0(I + \sum_{l=1}^{\infty} \Phi_l)\mathbf{e} = 1.$$

5. Calculate the vectors $\boldsymbol{\pi}_l$ by $\boldsymbol{\pi}_l = \boldsymbol{\pi}_0 \Phi_l, l \geq 1$.

4   The Stationary Distribution at an Arbitrary Time

Consider now the process of the system states at an arbitrary time

$$\zeta_t = \{i_t, r_t, \nu_t, m_t, k_t\}, \ t \geq 0,$$

where

- i_t is the number of customers in the orbit;
- r_t is the number of busy servers at the second station;
- ν_t is the state of the arrival directing process;
- m_t is the state of the service directing process;
- k_t is a random which takes values 0, 1, 2 depending on whether the server of the first station is idle, serves a customer, or it is blocked at time t, $t \geq 0$.

   The process ζ_t, $t \geq 0$, is non-Markovian. It can be classified as semi-regenerative processes, for definition see [20]. The stationary distribution of this process can be related to the stationary distribution of the embedded Markov chain ξ_n, $n \geq 1$, using the results [20] for Markov renewal and semi-regenerative processes.
   Let

$$p(i, r, \nu, m, k) == \lim_{t \to \infty} P\{i_t = i, \ r_t = r, \ \nu_t = \nu, \ m_t = m, \ k_t = k\},$$

$$i \geq 0, \ r = \overline{0, N}, \ \nu = \overline{0, W}, m = \overline{1, M}, \ k = \overline{0, 2},$$

be the steady-state probabilities of the process ζ_t, $t \geq 0$.
   Define the vectors of these probabilities

$$\boldsymbol{p}(i, r, \nu, k) = (p(i, r, \nu, 1, k), ..., p(i, r, \nu, M, k)),$$

$$\boldsymbol{p}(i, r, k) = (\boldsymbol{p}(i, r, 0, k), ..., \boldsymbol{p}(i, r, W, k)), \ \ \boldsymbol{p}_i(k) = (\boldsymbol{p}(i, 0, k), ..., \boldsymbol{p}(i, N, k)).$$

Theorem 2. The non-zero stationary probability vectors $\boldsymbol{p}_i(k)$, $i \geq 0$, $k = \overline{0,2}$, are related to the stationary probability vectors $\boldsymbol{\pi}_i$, $i \geq 0$, of the embedded Markov chain ξ_n, $n \geq 1$, as follows:

$$\boldsymbol{p}_0(0) = \lambda^{-1}\boldsymbol{\pi}_0[\bar{Q} + (1-p)\mathcal{H}_0\hat{Q}_3]A_0,$$

$$\boldsymbol{p}_i(0) = \lambda^{-1}[\boldsymbol{\pi}_i\bar{Q}A_i + (1-p)\sum_{l=0}^{i}\boldsymbol{\pi}_l\mathcal{H}_{i-l}\hat{Q}_3A_i], \; i \geq 1,$$

$$\boldsymbol{p}_i(1) = \lambda^{-1}\{\sum_{l=0}^{i}\boldsymbol{\pi}_l[\bar{Q}A_l\sum_{k=1}^{i-l+1}\hat{D}_k\tilde{\Omega}_{i-l-k+1} + (1-p)\sum_{k=0}^{i-l}\mathcal{H}_k\hat{Q}_3A_{l+k}$$

$$\times \sum_{m=1}^{i-l-k+1}\hat{D}_m\tilde{\Omega}_{i-l-k-m+1}] + \sum_{l=0}^{i+1}\boldsymbol{\pi}_l[\bar{Q}A_l\alpha_l\tilde{\Omega}_{i-l+1}$$

$$+ (1-p)\sum_{k=0}^{i-l+1}\mathcal{H}_k\hat{Q}_3A_{l+k}\alpha_{l+k}\tilde{\Omega}_{i-l-k+1}]\}, \; i \geq 0,$$

$$\boldsymbol{p}_i(2) = \lambda^{-1}(1-p)\sum_{l=0}^{i}\boldsymbol{\pi}_l\sum_{k=0}^{i-l}(\mathcal{H}_k + \delta_{0,k}I)\hat{Q}_2\int_0^{\infty}e^{-\mu Rt}\otimes P(i-l-k,t)\otimes I_M dt,$$

where $\delta_{i,j}$ is Kronecker's symbol, $\tilde{\Omega}_n = \int_0^{\infty}e^{\Delta t}\otimes P(n,t)\otimes(I_M - \nabla_B(t))dt, n \geq 0$, $\nabla_B(t)$ is the diagonal matrix with diagonal elements $(B(t)\boldsymbol{e})_j, j = \overline{1,M}$.

Corollary 2. The stationary probability vectors $\boldsymbol{p}_i$, $i \geq 0$, can be calculated by the following formula:

$$\boldsymbol{p}_0 = \boldsymbol{p}_0(0), \; \boldsymbol{p}_i = \boldsymbol{p}_i(0) + \sum_{m=1}^{2}\boldsymbol{p}_{i-1}(m), \; i \geq 1.$$

5   Performance Measures

Basing on the stationary distribution we can calculate different performance characteristics of the system. The most important performance measures are calculated as follows:

- Mean number of customers at the first station at the service completion epoch at this station and at an arbitrary time

$$\tilde{L} = \boldsymbol{\Pi}'(1)\mathbf{e}, \qquad L = \boldsymbol{P}'(1)\mathbf{e},$$

where $\boldsymbol{\Pi}(z) = \sum_{i=0}^{\infty}\boldsymbol{\pi}_i z^i$, $\boldsymbol{P}(z) = \sum_{i=0}^{\infty}\boldsymbol{p}_i z^i$, $|z| \leq 1$.

- Variance of the mean number of customers at the first station at the service completion epoch at this station and at an arbitrary time

$$\tilde{D} = \boldsymbol{\Pi}''(1)\mathbf{e} + \tilde{L} - \tilde{L}^2, \qquad D = \boldsymbol{P}''(1)\mathbf{e} + L - L^2.$$

- The vector of the stationary distribution of the number of busy servers at the second station at the service completion epoch at the first station and at an arbitrary time

$$\tilde{\boldsymbol{r}} = \boldsymbol{\Pi}(1)(I_{N+1} \otimes \mathbf{e}_{\bar{W}M}), \qquad \boldsymbol{r} = \boldsymbol{P}(1)(I_{N+1} \otimes \mathbf{e}_{\bar{W}M}).$$

- Mean number of busy servers at the second station at the service completion epoch at the first station and at an arbitrary time

$$\tilde{N}_{busy} = \tilde{\boldsymbol{r}}diag\{r, r = \overline{0, N}\}\mathbf{e}, \qquad N_{busy} = \boldsymbol{r}diag\{r, r = \overline{0, N}\}\mathbf{e}.$$

- The probability that an arbitrary customer leaves the system or causes the blocking of the server at the first station

$$P_{loss} = p\boldsymbol{\Pi}(1)\hat{Q}_2\mathbf{e}, \qquad P_{block} = (1 - p)\boldsymbol{\Pi}(1)\hat{Q}_2\mathbf{e}.$$

- The probability of immediate access to the first station server

$$P_{imm} = -\lambda^{-1} \sum_{i=0}^{\infty} \boldsymbol{p}_i(0)(\mathbf{e}_{(N+1)M} \otimes D_0\mathbf{e}).$$

6   Numerical Examples

The proposed algorithms for calculating the stationary distributions were realized as computer programm using tools of software "Sirius++" [21]. In this section we present numerical examples demonstrating the behavior of the performance measures as function of the system load and intensity of retrials.

We consider $BMAP$ having the fundamental rate $\lambda = 10$, the correlation coefficient $c_{cor} = 0.1$ and characterized by the matrices

$$D_0 = \begin{pmatrix} -5.39233 & 4.33008 \times 10^{-5} \\ 1.74403 \times 10^{-5} & -0.70174 \end{pmatrix}, \ D_1 = D_3 = \begin{pmatrix} 1.60698 & 0.01072 \\ 0.11744 & 0.09307 \end{pmatrix},$$

$$D_2 = \begin{pmatrix} 2.14264 & 0.01429 \\ 0.15659 & 0.12410 \end{pmatrix}.$$

Semi-Markovian kernel $B(t)$ is defined as $B(t) = diag\{B_1(t), B_2(t)\}P$, where the transition matrix P has the form $P = \begin{pmatrix} 0.6 & 0.4 \\ 0.35 & 0.65 \end{pmatrix}$,

$B_1(t)$ and $B_2(t)$ define the Erlangian distributions of order 3 with the parameters 20 and 50, respectively. The average service time b_1 is equal to 0.102.
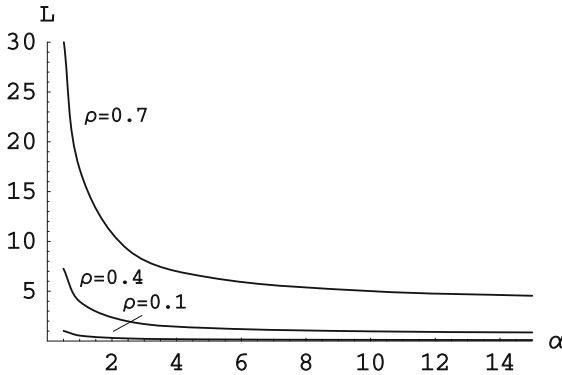
Fig. 1. The mean number of customers at the first station as a function of retrial rate α for different values of system load ρ.
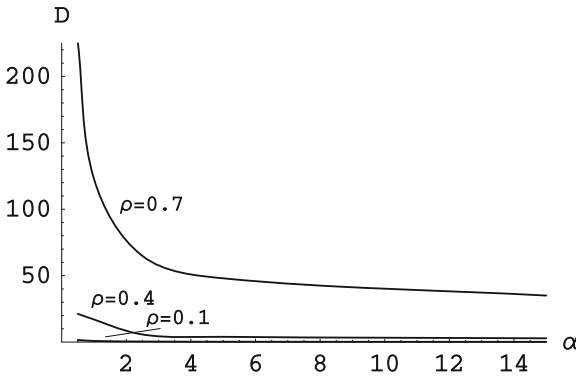

Fig. 2. The variance of the mean number of customers at the first station as a function of retrial rate α for different values of system load ρ.

We assume that the number of servers at the second station $N = 3$, the probability p is equal to 0.6. The parameters of service at the second station are as follows: $q_0 = 0$, $q_1 = 0.9$, $q_2 = q_3 = 0.05$, and the service rate $\mu = 3$.

We consider the classical retrial strategy: $\alpha_i = i\alpha$, $i \geq 0$, and vary the retrial rate α in the interval $[0.5, 15]$.

Figures 1, 2 illustrate the dependence of mean number L of customers and the variance D of the number of customers at the first station on the value α for three different values of system load ρ.

The dependence of the loss probability P_{loss} and the probability P_{imm} of immediate access to the first station server on the value α under the different values of system load is presented in Figs. 3 and 4.

The system load ($\rho = 0.1, 0.4, 0.7$) in this experiment varies by means of scaling the fundamental rate ($\lambda = 0.8, 3.12, 5.45$, respectively). In turn, the
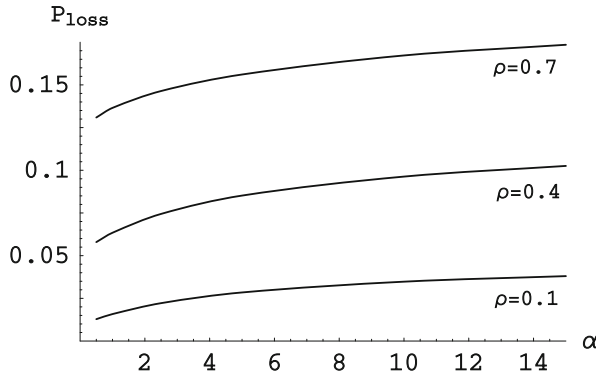
Fig. 3. The loss probability as a function of retrial rate α for different values of system load ρ.
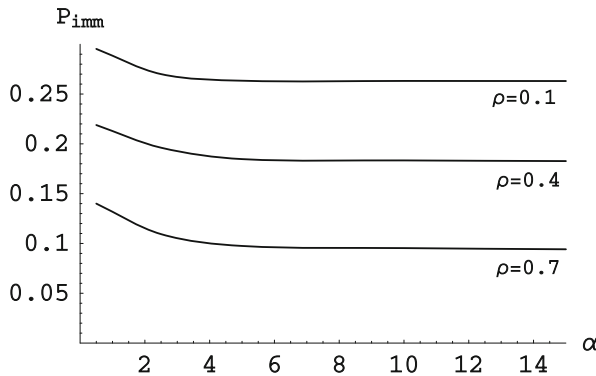

Fig. 4. The probability of immediate access to the first station server as a function of retrial rate α for different values of system load ρ.

fundamental rate λ of the $BMAP$ varies by multiplying the matrices D_k, $k = \overline{0,3}$, by some positive constant.

Figures show that the mean number of customers, the variance of the number of customers at the first station and the loss probability increase, while the probability of immediate access to the first station server without visiting the orbit decrease when the input intensity λ (and the system load ρ) increases. It confirms the fact that the system load has a great impact on the performance measures: the increase of the load makes the quality of service in the system worse.

It is also seen from the figures that all characteristics are sensitive with respect to the value of retrial rate α. Under the fixed value of system load ρ, the probability P_{imm} becomes worse, while characteristics L and D become better when α increases, but when α becomes greater than 4 the change of these characteristics becomes unessential. The increase of retrial rate α has a

negative influence on the loss probability P_{loss}. This probability increases when α increases.

The obtained results illustrate the importance of computational investigation of the tandem queue under consideration and can be used for correct prediction of system operation.

7   Conclusion

In this paper, the tandem queue with repeated attempts and semi-Markovian service process is investigated. The processes of the system states at embedded epochs and at arbitrary time are studied. The condition for the stationary distribution existence is derived and the algorithms for calculating the stationary distribution are presented. The key performance measures are calculated. The numerical examples demonstrating the dependence of the performance measures on the system load and intensity of retrials are given. The results of this paper can be exploited for capacity planning, performance evaluations, and optimization of real-life tandem queues and two-node networks with the random multiple access to the first station in the case of correlated bursty traffic as well as for validation of general networks decomposition algorithms.

References

1. Gomez-Corral, A.: A bibliographical guide to the analysis of retrial queues through matrix analytic techniques. Ann. Oper. Res. **141**, 163–191 (2006)
2. Artalejo, J.R., Gomez-Corral, A.: Retrial Queueing Systems: A Computational Approach. Springer, Berlin (2008)
3. Falin, G., Templeton, J.: Retrial Queues. Chapman and Hall, London (1997)
4. Balsamo, S., Persone, V.D.N., Inverardi, P.: A review on queueing network models with finite capacity queues for software architectures performance prediction. Perform. Eval. **51**, 269–288 (2003)
5. Ferng, H.W., Chao, C.C., Peng, C.C.: Path-wise performance in a tree-type network: per-stream loss probability, delay, and delay variance analysis. Perform. Eval. **64**, 55–75 (2007)
6. Heindl, A.: Decomposition of general tandem networks with $MMPP$ input. Perform. Eval. **44**, 5–23 (2001)
7. Gnedenko, B.W., Konig, D.: Handbuch der Bedienungstheorie. Akademie Verlag, Berlin (1983)
8. Lucantoni, D.M.: New results on the single server queue with a batch Markovian arrival process. Commun. Stat.-Stoch. Models **7**, 1–46 (1991)
9. Breuer, L., Dudin, A.N., Klimenok, V.I., Tsarenkov, G.V.: A two-phase $BMAP/G/1/N \rightarrow PH/1/M-1$ system with blocking. Autom. Rem. Control **65**, 117–130 (2004)

10. Gomez-Corral, A.: A tandem queue with blocking and Markovian arrival process. Queueing Syst. **41**, 343–370 (2002)
11. Gomez-Corral, A.: On a tandem G-network with blocking. Adv. Appl. Probab. **34**, 626–661 (2002)
12. Gomez-Corral, A., Martos, M.E.: Performance of two-station tandem queues with blocking: the impact of several flows of signals. Perform. Eval. **63**, 910–938 (2006)
13. Gomez-Corral, A., Martos, M.E.: A matrix-geometric approximations for tandem queues with blocking and repeated attempt. Oper. Res. Lett. **30**, 360–374 (2002)
14. Klimenok, V.I., Breuer, L., Tsarenkov, G.V., Dudin, A.N.: The $BMAP/G/1/N \to PH/1/M - 1$ tandem queue with losses. Perform. Eval. **61**, 17–40 (2005)
15. Klimenok, V., Kim, C.S., Tsarenkov, G.V., Breuer, L., Dudin, A.N.: The $BMAP/G/1 \to \cdot/PH/1/M$ tandem queue with feedback and losses. Perform. Eval. **64**, 802–818 (2007)
16. Kim, C.S., Klimenok, V., Taramin, O.: A tandem retrial queueing system with two Markovian flows and reservation of channels. Comput. Oper. Res. **37**, 1238–1246 (2010)
17. Klimenok, V.I., Taramin, O.S.: Tandem service system with batch Markov flow and repeated calls. Autom. Rem. Control **71**, 1–13 (2010)
18. Graham, A.: Kronecker Products and Matrix Calculus with Applications. Ellis Horwood, Chichester (1981)
19. Klimenok, V.I., Dudin, A.N.: Multi-dimensional asymptotically quasi-Toeplitz Markov chains and their application in queueing theory. Queueing Syst. **54**, 245–259 (2006)
20. Cinlar, E.: Introduction to Stochastic Process. Prentice-Hall, N.J. (1975)
21. Dudin, A.N., Klimenok, V.I., Tsarenkov, G.V.: Software "Sirius++" for performance evaluations of modern communication networks. In: Amborski, K., Meuth, H. (eds.) Proceedings of the 16th European Simulation Multiconference, Darmstadt, 3–5 June 2002, pp. 489–493. SCS, Netherlands (2002)