Tandem queueing system with infinite and finite intermediate buffers and generalized phase-type service time distribution

Chesoong Kim, Alexander Dudin, Olga Dudina, Sergey Dudin

1. Introduction

Tandem queueing systems are very important part of the queueing theory that takes into account the possibility that a customer may need service from several sequentially arranged servers. The overwhelming majority of the results in the theory of tandem queues is obtained for systems with stationary Poisson arrival process and exponential service time distribution. Tandem queueing systems are very important in modeling various real-world systems such as computer networks, communication systems, manufacturing systems, and healthcare systems. In this paper, we analyze a tandem queueing system with two types of customers. Type 1 customers arrive to the first stage while type 2 customers arrive to the second stage directly. The service time at the first stage has an exponential distribution. The service times of type 1 and type 2 customers at the second stage have a phase-type distribution with different parameters. During a waiting period in the intermediate buffer, type 1 customers can be impatient and leave the system. The ergodicity condition and the steady-state distribution of the system states are analyzed. Some key performance measures are calculated. The Laplace–Stieltjes transform of the sojourn time distribution of type 2 customers is derived. Numerical examples are presented.
and consider a generalized phase-type distribution of the service times at the second stage instead of considering two separate phase-type distributions for each type of customers, that greatly facilitates the investigation of the system.

The queueing system under consideration is quite general and can be applied for modeling many real-world systems. Let us introduce several examples of such systems:

(i) Contact-center with an IVR (Interactive Voice Response). Type 1 customers can be interpreted as calls and type 2 customers are the text requests (e-mails). The first stage server represents an IVR and the second stage server is a contact center operator. E-mail requests arrive to the second stage directly and are always patient. Callers are firstly serviced by an IVR. If the caller cannot solve his (her) problem by using an IVR, he (she) can request to connect with an operator and move to the second stage. Type 1 customers can be impatient. The service times of e-mail requests and calls are different.

(ii) Manufacturing system. The system processes two types of details. Type 1 details need some preprocessing while type 2 details do not require such a preprocessing. After the preprocessing, type 1 details should be processed during a limited time (e.g., if the preprocessing includes heating of details, type 1 details should be processed before they become cool).

(iii) Database. The system processes queries from external users, who need preliminary identification and are non-patient, along with the requests from the internal users to retrieve or update information.

(iv) Information transmission system. More important information should be encrypted prior transmission and transmitted more urgent than some less important information which is sent without encryption.

(v) Medical system. Emergency surgery center has several rooms for anesthesiology and operation theatre. Depending on the type of injury, some arriving patients need more urgent treatment after a preliminary general anesthetic, while others may wait longer time and do not need general anesthetic at all.

The paper is organized as follows. In Section 2, the mathematical model is described. In Section 3, the process of the system states is considered. The ergodicity condition and the stationary state distribution are analyzed in Section 4. The expressions for the main performance measures of the system are given in Section 4. The Laplace–Stieltjes transform of the sojourn time distribution of an arbitrary type 2 customer is presented in Section 5. Section 6 contains numerical examples. Section 7 concludes the paper.

2. Mathematical model

We consider a tandem queueing system with two types of customers and two intermediate buffers. The structure of the system under study is presented in Fig. 1.

Customers arrive to the system according to the MMAP. The customers in the MMAP are heterogeneous and have different types. The arrival of customers is directed by the stochastic process \( v_t, t \geq 0 \), which is an irreducible continuous-time Markov chain with the state space \( \{0, 1, \ldots, W\} \). The sojourn time of this chain in the state \( v \) is an exponentially distributed with the positive finite parameter \( \lambda^{(v)} \). When the sojourn time in the state \( v \) expires, with probability \( p_{v,v}^{(0)} \), the process \( v_t \) jumps to the state \( v' \) without generation of a customer, \( v, v' \in \{0, W\}, v \neq v' \), and with probability \( p_{v,v}^{(0)} \), the process \( v_t \) jumps to the state \( v' \) with a generation of a type 1 customer, \( t = 1, 2 \), \( v, v' \in \{0, W\} \). The notation \( v, v' \in \{0, W\} \) means that the parameter \( v \) takes values in the set \( \{0, 1, \ldots, W\} \).

The behavior of the MMAP is completely characterized by the matrices \( D_0, D_0^{(1)} \), \( l = 1, 2 \), defined by the entries \( (D_0^{(1)})_{v,v'} = \lambda^{(v)} p_{v,v'}^{(0)} \), \( v, v' \in \{0, W\}, l = 1, 2 \), and \( (D_0)_{v,v} = -\lambda^{(v)} \), \( v = 0, W \). \( (D_0)_{v,v'} = \lambda^{(v)} p_{v,v'}^{(0)} \), \( v, v' \in \{0, W\}, v \neq v' \). The matrix \( D_1(1) = D_0 + D_0^{(1)} + D_0^{(1)} \) represents the generator of the process \( v_t, t \geq 0 \).

The average total arrival intensity \( \lambda \) is defined by \( \lambda = \theta (D_1(1) + D_2(1)) \) where \( \theta \) is the invariant vector of the stationary distribution of the Markov chain \( v_t, t \geq 0 \). The vector \( \theta \) is the unique solution to the system \( D_0(1) = 0, \theta e = 1 \). Hereinafter \( e \) denotes a column vector consisting of 1’s, and \( 0 \) is a zero row vector. The average arrival intensity \( \lambda \) of type 1 customers is defined by \( \lambda_1 = \theta D_1(1) e, l = 1, 2 \).

The squared integral (without differentiating the types of customers) coefficient of variation \( c_{var} \) of the intervals between customer arrivals is defined by \( c_{var} = 2(\theta D_0(1))^{-1} e - 1 \). The squared coefficient of variation \( c_{var}^{(l)} \) of inter-arrival times of type \( l \) customers is defined by

\[
c_{var}^{(l)} = 2\lambda_1 \theta (D_1(1) - D_0(1))^{-1} e - 1, \quad l \neq 1, \quad \lambda_1, l = 1, 2.
\]

The integral coefficient of correlation \( c_{cor} \) of two successive intervals between arrivals is defined by

\[
c_{cor} = (\lambda \theta (D_0(1))^{-1} (D_1(1) - D_0(1))^{-1} e - 1) / c_{var}.
\]

The coefficient of correlation \( c_{cor}^{(l)} \) of two successive intervals between type \( l \) customers’ arrivals is computed by

\[
c_{cor}^{(l)} = \left( \lambda_1 \theta (D_0(1) + D_0^{(1)} - 1) D_1(1) (D_0(1) + D_0^{(1)} - 1) e - 1 \right) / c_{var}^{(l)},
\]

\( l \neq 1, \quad \lambda_1, l = 1, 2 \).

We assume that type 1 customers arrive to the first stage, while type 2 customers arrive to the second stage directly, not entering the first stage. The first stage is described by an R-server queueing system without buffer. The service time for each server at this stage has an exponential distribution with the parameter \( \mu \).

After receiving service at the first stage, type 1 customer proceeds to the second stage of the tandem with probability \( q \), \( 0 \leq q \leq 1 \), or leaves the system forever (is lost) with the complementary probability. The second stage represents an N-server queue with a finite buffer of capacity \( K \) for type 1 customers (buffer 1) and an infinite buffer for type 2 customers (buffer 2).

If there is a free server at the second stage during an arbitrary type 1 customer arrival epoch, the customer is admitted to this buffer. Otherwise, type 1 customer leaves the system forever (is lost). If all servers are busy and buffer 1 is not full during an arbitrary type 1 customer arrival epoch, the customer is admitted to buffer 2.
We suggest that type 1 customers have priority, i.e., if there is at least one type 1 customer in buffer 1 during the service completion epoch, the next service will be provided to type 1 customer. Otherwise, type 2 customer is taken for service if he (she) is presenting in buffer 2. For customers of the same type the service discipline is First-In-First-Out.

At the second stage, type 1 customers can be impatient, i.e., the customer leaves buffer 1 after an exponentially distributed time described by the parameter $\alpha$, $0 < \alpha < \infty$, after arrival, due to a lack of service.

The service time of type 1 customer for the second stage server has a phase-type distribution $PH_l$ with an irreducible representation $(\beta_l, S_l)$, $l = 1, 2$. This service time can be interpreted as the time until the underlying Markov process $m_l(t)$, $t \geq 0$, with the finite state space $\{1, \ldots, M_l, M_l + 1\}$ reaches the single absorbing state $M_l + 1$ conditioned on the fact that the initial state of this process is selected among the states $\{1, \ldots, M_l\}$ according to the probabilistic row vector $\underline{p}_l$, $l = 1, 2$. The transition rates of the process $m_l(t)$ within the set $\{1, \ldots, M_l\}$ are defined by the sub-generator $S_l$ and the transition rates into the absorbing state (what leads to service completion) are given by the entries of the column vector $S^0_l = -S_l \underline{e}$, $l = 1, 2$. Note that the representation $(\beta_l, S_l)$ is irreducible if the matrix $S_l + S^0_l \underline{p}_l$ is irreducible, $l = 1, 2$.

The service time distribution function has the form $A_l(x) = 1 - \beta_l e^{-x}$, the Laplace-Stieljes transform (LST) $s^{(l)}_x = e^{-sx} dA_l(x)$ of this distribution is $\beta_l(s - S_l)^{-1} s^{(l)}_x$, $\text{Re} s > 0$, $l \geq 1, 2$. The mean service time of type 1 customer at the second stage is calculated by $b^{(l)} = \beta_l (-S_l)^{-1} \underline{e}$. $l = 1, 2$. The squared coefficient of variation of type 1 customer is given by $c^{(l)}_{\text{var}} = b^{(l)}_2 / (b^{(l)}_1)^2 - 1$ where $b^{(l)}_2 = 2 \beta_l (-S_l)^{-2} \underline{e}$.

3. The process of system states

Let us note that the queueing system under consideration is complicated for analysis. So, the task of the optimal choice of the process of the system states is extremely important.

The behavior of the queueing system under study can be described in terms of the following regular irreducible continuous-time Markov chain:

$$c^{(l)}_t = \left\{ i_l, r_l, k_l, n_l, v_l, o^{(n1)}_l, \ldots, o^{(N)}_l, s^{(m1)}_t, \ldots, s^{(\min(i_l, N) - n_l)}_t \right\}, \quad t \geq 0,$$

where $i_l$ is the number of customers at the first stage, $i_l \geq 0$, $r_l$ is the number of customers at the first stage, $r_l \geq 0$, $k_l$ is the number of type 1 customers in buffer 1, $k_l = 0$, $\max(0, \min(i_l - N, K))$, $n_l$ is the number of type 1 customers in service, $n_l = 0$, $\min(i_l, N)$, $v_l$ is the state of the underlying process of the MMAP, $v_l = 0$, $\omega^{(m)}_l$ is the state of the mth server that serves type 1 customer, $m = 1, T_k$, $\omega^{(m)}_l = \sum_{k=1}^{M_l} \omega^{(k)}_l$, $\omega^{(k)}_l$ is the state of kth server that serves type 2 customer, $l = 1, \min(i_l, N) - n_l$, $\omega^{(k)}_l = \sum_{k=1}^{M_l} \omega^{(k)}_l$, during the epoch $t$, $t \geq 0$.

It is evident that the Markov chain $c^{(l)}_t$ has a huge dimension of the state space even for small dimensions of underlying MMAP and PH processes. Computation of the performance measures of the queueing system under consideration even on advanced computer seems to be extremely difficult task. To reduce the dimension of the stochastic process $c^{(l)}_t$ we use the method proposed by Ramaswami (1985) and Ramaswami and Lucchini (1985), for reducing the dimension of the state space of systems with a phase-type service time distribution. The main idea of this method is the following. Instead of the consideration of the components $\omega^{(n1)}_l, \ldots, \omega^{(N)}_l$ and $s^{(m1)}_t, \ldots, s^{(\min(i_l, N) - n_l)}_t$, it is proposed to consider the components $\omega^{(n1)}_l, \ldots, \omega^{(M_l)}_l$ and $s^{(l1)}_t, \ldots, s^{(M_l)}_t$, which mean the following:

- $o^{(m)}_l$ is the number of servers at the phase $m$ of service of type 1 customers, $m = \sum_{l=1}^{M_l} \omega^{(l)}_l$.
- $c^{(l)}_l$ is the number of servers at the phase $l$ of service of type 2 customers, $l = \sum_{l=1}^{M_l} \omega^{(l)}_l$.

Then to reduce the dimension of the process of system states instead of the Markov chain $c^{(l)}_t$ we should consider the Markov chain $c^{(l)}_t = \{i_l, r_l, k_l, n_l, v_l, o^{(n1)}_l, \ldots, o^{(M_l)}_l, s^{(l1)}_t, \ldots, s^{(M_l)}_t\}, \quad t \geq 0$.

Despite the fact that the Markov chain $c^{(l)}_t$ has more suitable dimension of the state space for investigation in comparison to Markov chain $c^{(l)}_t$, its complicated structure is not conducive to the analysis.

In order to greatly simplify the investigation of the queue under study, instead of considering the service times of type 1 and type 2 customers we propose to consider a generalized processing time having distribution which we call as a generalized phase-type distribution with an irreducible representation $(\beta^g, S^g)$.

The time having such a distribution can be interpreted as the time until the underlying Markov process $\eta_t, t \geq 0$, with the finite state space $\{1, \ldots, M, M + 1\}$, where $M = M_1 + M_2$, reaches the single absorbing state $M + 1$. The initial state of this process is selected among the states $\{1, \ldots, M\}$ depending on the type of a customer who is chosen for service. If an arbitrary type 1 customer is chosen for service, the initial state of this process is selected according to the probabilistic row vector $\beta^g = (\beta_{g1}, \beta_{g2})$ and, if the type 2 customer is chosen for service, the initial state is selected according to the probabilistic row vector $\beta^{g2} = (\beta_{g2}, \beta_{g1})$. The transition rates of the process $\eta_t, t \geq 0$, are given by the entries of the column vector $S^g = (S_{g1}, S_{g2})$ and the transition rates into the absorbing state (which lead to the service completion) are given by the entries of the column vector $S^g = -S^g$.

If we use a generalized phase-type distribution, we do not consider the service process of each type of customers separately. Using of a generalized phase-type distribution instead of consideration of two separate phase-type distributions does not change the dimension of the state space of the Markov chain, but allows us to eliminate from the consideration of the component $l_t$ (the number of type 1 customers in service) that greatly simplifies the investigation of the system under study and allows us to construct the generator, obtain the ergodicity condition, and find sojourn time distribution more or less easily.

So, instead of the Markov chain $c^{(l)}_t$ we assume that the behavior of the system under study is described in terms of the regular irreducible continuous-time Markov chain $c_t = \{i, r, k, n, v, \eta^{(n1)}_l, \ldots, \eta^{(M_l)}_l\}, \quad t \geq 0,$

where $\eta^{(m)}_l$ is the number of servers at the phase $m$ of generalized service, $m = \sum_{l=1}^{M_l} \omega^{(l)}_l = \sum_{l=1}^{M_l} \min(i_l, N)$, $\sum_{m=1}^{M_l} \eta^{(m)}_l = \min(i_l, N)$, during the epoch $t$, $t \geq 0$.

4. Ergodicity condition and stationary probabilities

For further use throughout this paper, we introduce the following notation:

- $I$ is the identity matrix; $\oplus$ and $\odot$ indicate the symbols of Kronecker sum and product of matrices, respectively;
- $W = W + 1$, $R = R + 1$, $K = K + 1$, $K_t = \max(0, \min(i_l, N)) + 1$, $t \geq 0$;
- $T_t = \left( \frac{i + M - 1}{M - 1} \right) = \left( \frac{i - M}{\max(i, N)} \right), \quad i = \overline{0, N};$
\[ Q_{i,j} = -\mu C_i \otimes I_{E_{i-N+1} \otimes I_{\tau}} + I_{E_i} \otimes D_0 \otimes I_{\tau} + I_{E_i} \otimes \Gamma \]
\[ + (A_{\min(i,N)}(N,S) + A_{\min(i,N)}(N)) + (I_2 + I_0) \otimes I_2 \otimes D^{(1)} \otimes I_{\tau} \]
\[ + (1 - \eta) \mu C_i \otimes I_{E_{i-N+1} \otimes I_{\tau}} - \lambda \otimes C_i \otimes I_{\tau} \]
\[ 0 \leq i < N + K, \]
\[ Q_{i,i} = Q_i = -\mu C_i \otimes I_{E_{i-N+1} \otimes I_{\tau}} + I_2 \otimes D_0 \otimes I_{\tau} + I_2 \otimes \Gamma + (A_{\min(N,S)} + A^{(N)}) \]
\[ + (I_0 + I_2) \otimes I_2 \otimes D^{(1)} \otimes I_{\tau} + (1 - \eta) \mu C_i \otimes I_{\tau} \]
\[ - \lambda \otimes C_i \otimes I_{\tau} + \mu C_i \otimes I_{\tau}, \quad i > N + K, \]
The entries of the blocks $Q_{i,j}$, $i \geq j$, define the intensities of the transitions of the Markov chain $\zeta$, $t \geq 0$, that decrease the number of customers at the second stage by one. This can happen if (a) an arbitrary customer completes service or (b) type 1 customer leaves the system due to impatience.

If there are no customers in buffer 1 and buffer 2, i.e., $i \leq N$, only event (a) is possible. Their intensities are given by the entries of the matrix $I_{Q_{i}} \odot E_{i-N} \odot l_{1}$, if there are type 1 customers in buffer 1 and, by the entries of the matrix $l_{2} \odot E_{i-N} \odot l_{2}$, otherwise. The intensities of event (b) are given by the entries of the matrix $l_{2} \odot C_{i} \odot E_{i-N} \odot l_{1}$.

By analogy, if the number of customers in the buffers exceeds $K$, i.e., $i > N + K$, then both events are also possible and their intensities are given by the entries of the matrices $l_{R} \odot I_{R} \odot I_{W} \odot l_{1} \odot l_{1} \odot l_{1} \odot l_{2} \odot l_{2}$, respectively. The intensities of event (b) are given by the entries of the matrix $l_{R} \odot C_{i} \odot C_{i} \odot l_{R} \odot l_{1}$.

The non-diagonal entries of the blocks $Q_{i,j}$, $i > j$, define the intensities of the transitions of the chains $\zeta_{i}$, $t \geq 0$, that do not lead to the change of the number of customers at the second stage. If the number of customers at the second stage $i < N + K$, the following events are possible: (a) the transition of the process $\tau_{i}$, without generation of customers; (b) the transition of the process $\tau_{i}$, with generation of type 1 customer; (c) loss of type 1 customer after service at the first stage, and (d) the transition of the process $\{\eta^{(1)}, \ldots, \eta^{(M)}\}$, $t \geq 0$, which does not lead to the service completion at the second stage. The intensities of events (a), (b), (c) and (d) are given by the non-diagonal entries of the matrices $l_{R} \odot I_{R} \odot D_{0} \odot I_{R} \odot l_{R} \odot l_{R} \odot D_{1} \odot I_{R} \odot l_{R} \odot (1 - \mu)C_{R} \odot I_{R} \odot l_{R}$ and $l_{R} \odot A_{i}(N, S)$, respectively.

The diagonal entries of the matrix $Q_{i,i}$ are negative and the modulus of each entry defines the total intensity of leaving the corresponding state of the chain $\zeta_{i}$, $t \geq 0$. The diagonal entries of the matrix $-\mu C_{R} \odot I_{R} \odot D_{0} \odot I_{R} \odot l_{R} \odot l_{R} \odot D_{1} \odot I_{R} \odot l_{R} + l_{R} \odot I_{R} \odot D_{1} \odot I_{R} \odot l_{R} + l_{R} \odot C_{R} \odot A_{i}(N, S) + (1 - \mu)C_{R} \odot I_{R} \odot l_{R} + l_{R} \odot l_{R}$ define the total intensity of leaving the corresponding states of this process given that $i$ servers are busy. For $i > N + K$, the blocks $Q_{i,j}$ do not depend on $i$ and are equal to the matrix $Q_{1}$.

Since the probability that more than one arrival can occur and more than one customer can leave the system during the interval of an infinitesimal length is negligible, the blocks $Q_{i,j}$ are zero matrices for $|i - j| > 1$, thus the generator $Q$ has a block-tridiagonal structure.

**Corollary 1.** The Markov chain $\zeta$, $t \geq 0$, belongs to the class of continuous-time quasi-birth-and-death processes, see, e.g., Neuts (1981).

It follows from Neuts (1981) that the necessary and sufficient ergodicity condition of the quasi-birth-and-death process is the fulfillment of the inequality

$$y_{Q_{i}} > y_{Q_{i+1}}.$$  \hspace{1cm} (1)

where the row vector $y$ is the unique solution to the following system of linear algebraic equations

$$y_{Q_{0} + Q_{1} + Q_{2}} = 0, \quad ye = 1.$$  

Let us suppose that the vector $y$ has the form $(y_{0}, y_{1}, \ldots, y_{g})$. Note that the matrix $Q_{0} + Q_{1} + Q_{2}$ has a block-tridiagonal structure and we propose the following numerically stable algorithm for calculation of the sub-vectors $y_{r}$, $r = 0, R$.

**Step 1.** Calculate the matrices $Q_{r,2}$ and $Q_{r,1}$ which are the diagonal, the up-diagonal and the sub-diagonal blocks of the matrix $Q_{0} + Q_{1} + Q_{2}$ respectively:

$$Q_{r,2} = I_{R} \odot (D_{0} + \delta_{R} D_{1}^{(1)} + D_{1}^{(2)}) \odot I_{R} + I_{R} \odot (A_{i}(N, S) + A^{(1)}) + I_{R} \odot I_{R} \odot l_{1} + l_{1} \odot I_{R} \odot l_{1} - \mu I_{R} \odot l_{1} + \frac{1}{2}C_{R} I_{R} \odot I_{W} \odot l_{1}, \quad r = 0, R.$$  

$$Q_{r,1} = I_{R} \odot D_{1}^{(1)} \odot I_{R}, \quad r = 0, R - 1,$$

$$Q_{r,0} = \mu I_{R} \odot l_{1}, \quad r = 1, R, \ldots, 0.$$  

where the Kronecker delta.

**Step 2.** Calculate the matrices $A_{r}$ using the recurrent formulas

$$A_{r} = -Q_{r,1}A_{r+1}Q_{r,2}^{-1}, \quad r = 0, R - 2, R - 3, \ldots 0.$$  

**Step 3.** Calculate the matrices $F_{r}$ using the recurrent formulas

$$F_{0} = I, \quad F_{r} = F_{r-1}A_{r-1}^{-1}, \quad r = 1, R.$$  

**Step 4.** Calculate the sub-vector $y_{0}$ as the unique solution to the following system

$$y_{0} - A_{0} y_{0} = 0, \quad y_{0} \sum_{r=0}^{R} F_{r} e = 1.$$  

**Step 5.** Calculate the sub-vectors $y_{r} = y_{r} F_{r}, \quad r = 1, R$.

If ergodicity condition (1) of the Markov chain $\zeta$, is fulfilled, then the stationary probabilities of the system states $\pi(i, r, k, v)$, $i > 0, \quad r = 0, R, \quad k = 0, \max\{0, \min\{i - N, K\}\}, \quad v = 0, W, \quad h^{(1)}, \ldots, h^{(M)}$, then let us form the row vectors $\pi_{i}(i, r, k, v)$ of these probabilities enumerated in the reverse lexicographic order of the components $h^{(0)}, \ldots, h^{(M)}$. Then let us form the row vectors

$$\pi_{i}(i, r, k) = \left(\pi_{i}(i, r, k, 0, \pi_{i}(i, r, k, 1), \ldots, \pi_{i}(i, r, k, W)\right),$$

$$\pi_{i}(r) = \left(\pi_{i}(i, r, 0), \pi_{i}(i, r, 1), \ldots, \pi_{i}(i, r, \max\{0, \min\{i - N, K\}\})\right),$$

$$r = 0, R, \quad i = 0, \ldots, R.$$  

It is well known that the probability vectors $\pi_{i}$, $i > 0$, satisfy the following system of linear algebraic equations:

$$(\pi_{0}, \pi_{1}, \ldots, \pi_{g}) = 0, \quad (\pi_{0}, \pi_{1}, \ldots, \pi_{g}) e = 1$$

where $Q$ is the infinitesimal generator of the Markov chain $\zeta$, $t \geq 0$. To solve this system the numerically stable algorithm that takes
5. Performance measures

As soon as the vectors \( \pi_i \), \( i \geq 0 \), have been calculated, we are able to find various performance measures of the system.

The probability \( P^{(1)}_{\text{los}} \) that an arbitrary type 1 customer will be lost at the first stage is computed by

\[
P^{(1)}_{\text{los}} = 1 - \sum_{i=0}^{\infty} \pi(i, R) \left( I_{k_i} \otimes D_1^{(1)} \otimes I_{\text{les}(i)} \right) e.
\]

The average number \( N^{(1)} \) of busy servers at the first stage is computed by

\[
N^{(1)} = \sum_{r=1}^{\infty} \sum_{i=0}^{\infty} \pi(i, r) e.
\]

The average intensity \( \lambda^{(1)}_\text{out} \) of flow of customers, who receive service at the first stage, is computed by

\[
\lambda^{(1)}_\text{out} = N^{(1)} \mu.
\]

The average number \( N^{(2)} \) of busy servers at the second stage is computed by

\[
N^{(2)} = \sum_{i=0}^{\infty} \min\{i, N\} \pi(i, r) e.
\]

The average number \( N^{(2)}_{\text{buffer}} \) of type 1 customers in buffer 1 is computed by

\[
N^{(2)}_{\text{buffer}} = \sum_{i=0}^{\infty} \sum_{r=0}^{\infty} \sum_{k=0}^{\min\{i, N-K\}} k \pi(i, r, k) e.
\]

The average number \( N^{(2)}_{\text{buffer}} \) of type 2 customers in buffer 2 is computed by

\[
N^{(2)}_{\text{buffer}} = \sum_{i=0}^{\infty} \sum_{r=0}^{\infty} \sum_{k=0}^{\min\{i, N-K\}} (i - N - k) \pi(i, r, k) e.
\]

The average number \( L^{\text{system}} \) of customers in the system is computed by

\[
L^{\text{system}} = \sum_{r=0}^{\infty} \sum_{i=0}^{\infty} (i + r) \pi(i, r) e.
\]

The average intensity \( \lambda^{(2)}_\text{out} \) of flow of customers, who receive service at the second stage, is computed by

\[
\lambda^{(2)}_\text{out} = \sum_{i=1}^{\infty} \pi(i, R_{\text{buffer}}) \otimes \Gamma_{\text{max}(0, N-i)} (N, S) e.
\]

The probability \( P^{(2)}_{\text{los}} \) that an arbitrary type 1 customer will be lost at the second stage is computed by

\[
P^{(2)}_{\text{los}} = 1 - \lambda^{(2)}_\text{out} / \lambda^{(2)}_\text{out}.
\]

The probability \( P^{(2, \text{ent})}_{\text{los}} \) that an arbitrary type 1 customer will be lost at the entrance to the second stage is computed by

\[
P^{(2, \text{ent})}_{\text{los}} = \sum_{i=k}^{\infty} \pi(i, r, k) e.
\]

The probability \( P^{(2, \text{imp})}_{\text{los}} \) that, after arrival to the second stage, an arbitrary type 1 customer will go to the buffer and leave it due to impatience is computed by

\[
P^{(2, \text{imp})}_{\text{los}} = \sum_{i=k}^{\infty} \pi(i, r, k) e.
\]

The average intensity \( \lambda^{(2)}_\text{out} \) of flow of type 1 customers, who receive service at the second stage, is computed by

\[
\lambda^{(2)}_\text{out} = \frac{1}{1 - P^{(2)}_{\text{los}} - P^{(2, \text{ent})}_{\text{los}}}.
\]

The average number \( N^{(1)}_\text{out} \) of flow of type 1 customers, who receive service at the second stage, is computed by

\[
N^{(1)}_\text{out} = \left( 1 - P^{(2)}_{\text{los}} \right) \lambda^{(2)}_\text{out}.
\]

6. Distribution of sojourn and waiting times of type 2 customer

Because type 1 customers have a priority, they leave the system due to impatience and restriction on the total number of type 1 customers at the second stage, the upper bound of their sojourn time is more or less clear. So, we concentrate on the consideration of the sojourn time of type 2 customers.

We will derive the distribution of an arbitrary type 2 customer’s waiting and sojourn time in terms of the LST. Let \( W(x) \) be the distribution function of the sojourn time of an arbitrary type 2 customer in the system under study and \( w(s) = \int_0^s e^{-st} dW(x) \), \( \Re s > 0 \), be its LST.

To derive the expression for the LST \( w(s) \) we use the method of collective marks (method of additional event, method of catastrophes) (see, e.g., Kesten & Runenburg (1956) and van Danzig (1955)). Let us tag an arbitrary type 2 customer and keep track of its staying in the system. We interpret the variable \( s \) as the intensity of some virtual stationary Poisson flow of catastrophes. Thus, \( w(s) \) has the meaning of the probability that no catastrophe arrives during the sojourn time of the tagged type 2 customer.

Let \( w(s, n, r, k, v, \eta(1), \ldots, \eta(N)) \) be the probability that catastrophe will not arrive during the tagged type 2 customer’s sojourn time in the system conditioned on the fact that, during the given moment, the position of the tagged customer in buffer 2 is \( n, n \geq 1 \), the number of customers at the first stage is \( r, r = 0 \), the number of type 1 customers in buffer 1 is equal to \( k, k = 0 \), the state of the process \( v, v \), the states of the processes \( \eta(1), \ldots, \eta(N) \) are \( \eta(1), \ldots, \eta(N) \), respectively, \( t \geq 0 \).

Let us enumerate the probabilities \( w(s, n, r, k, v, \eta(1), \ldots, \eta(M)) \) in the lexicographic order of components as indicated above and form the column vectors \( w(s, n, r, k) \) from these probabilities.

**Theorem 1.** The LST \( w(s) \) of distribution of an arbitrary type 2 customer’s sojourn time in the system is computed by

\[
w(s) = \lambda^{(1)}_2 \sum_{i=0}^{N-1} \pi(i, R) \otimes D_1^{(2)} \otimes I_{r_i} e \beta_2(s - S_2) \frac{S_2^{(2)}}{w(s, i + r, k)}.
\]

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

- There is an idle server during the arrival epoch of the tagged customer and this customer immediately starts getting service. The probability of this event is \( \lambda^{(1)}_2 \sum_{i=0}^{N-1} \pi(i, R) \otimes D_1^{(2)} \otimes I_{r_i} e \). In this case, the probability that no catastrophe arrives during the sojourn time is equal to the probability that no catastrophe arrives during the service time of type 2 customer and is given is \( \beta_2(s - S_2) \). The probability of this event is

- All servers are busy during the arrival epoch and the tagged customer joins buffer 2. The probability of this event is \( \lambda^{(1)}_2 \sum_{i=0}^{N-1} \sum_{r=0}^{i+k} \sum_{k=0}^{i+k} \pi(i, r, k) \otimes D_1^{(2)} \otimes I_{r_i} e \). In this case, the probability that no catastrophe arrives during the sojourn time of the tagged type 2 customer under the fixed values of the components \( n, r, k \) is equal to \( w(s, i - N + k + 1, r, k) e \).
Using the law of total probability one can verify the validity of the theorem. □

So, for calculation of the LST $w(s)$ we must compute the vectors $w(s, n, r, k)$. Let us introduce the column vectors
\[
w(s, n, r) = (w(s, n, r, 0), \ldots, w(s, n, r, K))^T, \\
w(s, n) = (w(s, n, 0), \ldots, w(s, n, R))^T.
\]

**Theorem 2.** The column vectors $w(s, n)$, $n \geq 1$, are computed by
\[
w(s, 1) = (sl - \Omega)^{-1}a(s), \\
w(s, n) = (sl - \Omega)^{-1} \Psi w(s, n - 1), \quad n > 1,
\]
where
\[
\begin{align*}
\Omega &= -\mu C_s \otimes I_{m[R]} + I_{m[R]} \otimes \left( (D_0 + D_1^2) \otimes (A_N(N, S) + A'^N) \right) + \left( I_{m[R]} \otimes D_1^1 + I_{m[R]} \otimes A^N \otimes I_{m[R]} \right) + (q - I_{m[R]} \otimes I_{m[R]} \otimes I_{m[R]} \otimes I_{m[R]}), \\
\Psi &= I_{m[R]} \otimes I_{m[R]} \otimes L_2.
\end{align*}
\]
\[a(s) = \mathbf{e}_R \otimes \mathbf{e} \otimes ((I_{m[R]} \otimes L_2) \mathbf{e}) \beta_2 (sl - S_1)^{-1} S_2^2.
\]
and $\mathbf{e}$ is the column vector of size $R$ with all zero entries except the entry $(\mathbf{e})_0 = 1$.

**Proof.** Based on a probabilistic sense of the LST and the law of total probability, the vectors $w(s, n, r, k)$ can be found from the following system of algebraic equations:
\[
w(s, n, r, k) = [I + kx + r \mu I - D_0 \otimes (A_N(N, S) + A'^N)]^{-1} \\
\times \left( (1 - \delta_{rr}) D_1^1 \otimes I_{m[R]} w(s, n, r + 1, k) + \left( \delta_{rr} D_1^1 + D_1^2 \right) \otimes I_{m[R]} w(s, n, r, k) + \mu (I - q) \otimes I_{m[R]} w(s, n, r, 1, k) + (1 - \delta_{rr}) \mu r I_{m[R]} w(s, n, r - 1, k + 1) + \delta_{rr} \delta_{r0} (I_{m[R]} \otimes L_2) \mathbf{e}_R \beta_2 (sl - S_1)^{-1} S_2^2 \\
+ \delta_{rr} (1 - \delta_{rr}) I_{m[R]} \otimes L_2 w(s, n, r, 1, 0) + (1 - \delta_{rr}) \mu \beta_2 \otimes I_{m[R]} \otimes I_{m[R]} \otimes I_{m[R]} w(s, n, r, k - 1) \right), \\
\]
\[n \geq 1, \quad r = R, \quad k = R.
\]

Let us explain formula (2) in brief. The diagonal entries of the matrix in the square brackets in (2) are equal to the total intensity of the events which can happen after an arbitrary epoch: catastrophe, service completion, the transition of the underlying process of the MMAP, the transition of any underlying process of the generalized PH service processes, abandoning of type 1 customers from the buffer, and the service completion of type 1 customer at the first stage. The non-diagonal entries of the matrix in the square brackets in (2) are equal to the intensities of the transition of the MMAP underlying process without generation of a customer, the transition of one of generalized PH underlying processes which does not lead to the service completion (these events do not impact on the components $n$, $r$, and $k$). The first term in the round brackets in (2) corresponds to the case when type 1 customer is admitted at the first stage of the system. In this case the number of customers at the first stage (the component $r$) increases by one. The second term in the round brackets in (2) corresponds to the cases when type 1 customer is rejected at the first stage upon arrival and type 2 customer is admitted at the second stage. In this case the values of the components $n$, $r$, and $k$ do not change. The third term corresponds to the cases when type 1 customer leaves the system after the service completion at the first stage. In this case the number of customers at the first stage decreases by one. The fourth term corresponds to the case when type 1 customer is admitted to the second stage. In this case the number of customers at the first stage decreases by one, and the number of type 1 customers in buffer 1 (the component $k$) increases by one. The fifth term correspond to the case when during the service completion epoch at the second stage there are no type 1 customers in buffer 1, the tagged customer has the first position in buffer 2 and is chosen for service. The number $\delta_{rr} (I_{m[R]} \otimes L_2) \mathbf{e}_R \beta_2 (sl - S_1)^{-1} S_2^2$ defines the probability that catastrophe will not arrive during the service time of the tagged customer. The sixth term corresponds to the case when during the service completion epoch at the second stage there are no type 1 customers in buffer 1 and there are type 2 customers that arrive to the system before the tagged customer. In this case the position of the tagged customer in buffer 2 (component $n$) decreases by one. Finally, the seventh term corresponds to the case when type 1 customer leaves the buffer (starts service or leaves the buffer due to impatience).

System (2) can be rewritten into the matrix form as
\[
(-sl + \Omega) w(s, n) + (1 - \delta_{rr} \mu) \Psi w(s, n - 1) + \delta_{rr} a(s) = 0^T, \quad n \geq 1.
\]

It can be verified that the diagonal entries of the matrix $\Omega - sl$ dominate in all rows of this matrix. So the inverse matrix exists. Thus, based on (3) one can verify the validity of the theorem. □

**Corollary 2.** The average sojourn time $\varphi_0$ of an arbitrary type 2 customer is calculated by
\[
\varphi_0 = \lambda_2 + \frac{1}{\mu} \left( b_0^2 \sum_{i=0}^{N-1} \pi_i \left( \beta_1^1 \otimes I_{m[R]} \right) \mathbf{e} \right) - \mu \sum_{i=0}^{N-1} \sum_{r=0}^{K} \sum_{k=0}^{K} \pi(i, r, k) \left( D_1^2 \otimes I_{m[R]} \right) w(s, i - n - k + 1, r, k),
\]
where the column vectors $w(s, n, r, k)|_{s=0}$ are calculated as the blocks of the vector $w(s, n)|_{s=0}$ which can be computed as follows
\[
w(s, 1)|_{s=0} = \Omega^{-1} \mathbf{e} - a(s)|_{s=0};
\]
\[
w(s, n)|_{s=0} = \Omega^{-1} (t - \Psi) w(s, n - 1)|_{s=0}, \quad n > 1.
\]

Here $a(s)|_{s=0} = -b_0^2 \mathbf{e}_R \otimes \mathbf{e} \otimes ((I_{m[R]} \otimes L_2) \mathbf{e})$.

**Proof.** Formula for the calculation of the average sojourn time of an arbitrary type 2 customer is based on the definition $\varphi_0 = -w(s)|_{s=0}$ taking into account that $w(s, n)|_{s=0} = \mathbf{e}, \quad n \geq 1$. □

**Corollary 3.** The average waiting time $\varphi_{\text{wait}}$ of an arbitrary type 2 customer is calculated by
\[
\varphi_{\text{wait}} = \varphi_0 - b_1^2.
\]

7. Numerical examples

In Experiment 1, we investigate the impact of the coefficient of correlation in the arrival flow on the system performance measures.

For this purpose, let us introduce three MMAPs defined by the matrices $D_0$, $D_1^1$ and $D_1^2$. All these MMAPs have the same average total arrival intensity $\lambda$, the average intensity of type 1 customers $\lambda_1 = \frac{1}{2} \lambda$, the average intensity of type 2 customers $\lambda_2 = \frac{1}{2} \lambda$, but different coefficients of correlation. MMAP$^1$ defines the MMAP arrival process with the coefficient of correlation $\rho_{\text{corr}} = \lambda$. 

The first process, coded as MMAP\(^p\), is defined by the matrices
\[
D_0 = \lambda \begin{pmatrix} -1.35162 & 0 \\ 0 & -0.04384 \end{pmatrix}, \quad D_1^{(1)} = \lambda \begin{pmatrix} 1.00699 & 0.00673 \\ 0.01832 & 0.01457 \end{pmatrix},
\]
has the coefficients of correlation \(c_{\text{cor}}^{(1)} = 0.17\) and \(c_{\text{var}}^{(1)} = 12.34\), and the coefficients of variation \(c_{\text{var}}^{(1)} = 10.55\). Let us vary the average total arrival intensity \(\lambda\) in the interval \([1, 15]\) in steps of 0.1.

Figs. 2, 3 illustrate the dependence of the average number \(L\) of customers in the system, the average sojourn time \(V_{\text{imp}}\) of type 2 customers, the average waiting time of type 2 customers at the second stage, the average sojourn time of type 2 customers at the second stage, the average arrival intensity \(\lambda_{\text{in}}\) of flow of type 1 customers, who receive service at the first stage, the loss probability \(p_{\text{los}}^{(2)}\) of an arbitrary type 1 customer at the first stage, the average intensity \(\lambda_{\text{out}}^{(2)}\) of flow of customers, who receive service at the first stage, the loss probability \(p_{\text{los}}^{(2)}\) of an arbitrary type 1 customer at the second stage, and the probability \(P^{(2)}\) that an arbitrary type 1 customer will join buffer 1 and leave it due to impatience on the average arrival intensity \(\lambda\) for different MMAPs.

It is clearly evident from Figs. 2 and 3 that an increase in the average arrival intensity leads to an increase in the average intensities of output flows, the loss probabilities of type 1 customers and the average sojourn time of type 2 customers. So, with increasing average arrival intensity the quality of service becomes worse. Based on figures presented one can conclude that the coefficient of correlation in the arrival process has a profound impact on the system performance measures and taking into account the correlation in the arrival flow is extremely important for adequate modeling of the system performance. For example, when the coefficient of correlation increases, the average output intensities \(\lambda_{\text{out}}^{(1)}\) and \(\lambda_{\text{out}}^{(2)}\) decrease essentially, while the loss probability of type 1 customers at the first stage increases essentially. Additionally, the fulfillment of the ergodicity condition depends on the coefficient of correlation in the arrival flow. In the example under consideration, the ergodicity condition is not fulfilled for \(\lambda > 14.1\) in case of MMAP\(^p\) and for \(\lambda > 14.9\) in case of MMAP\(^p^2\). This finding can be explained as follows. As it was mentioned above, for arrival flows with small coefficients of correlation the loss probability of type 1 customers at the second stage is less than for arrival flows with high coefficients of correlation. Thus, for arrival flows with small coefficients of correlation...
more type 1 customers arrive to the second stage thereby increasing the load of the second stage and lead to an earlier violation of the ergodicity condition. It also explains the dependence of the average number $L$ of customers in the system and the average sojourn time $V_{soj}$ of an arbitrary type 2 customer on the average arrival intensity $\lambda$ for $\lambda > 13$. Note for example, that for the stationary Poisson arrival flows with $\lambda = 13$ the average sojourn time of an arbitrary type 2 customer $V_{soj}^{(2)} = 5.23$, while under $\lambda = 14$ the average sojourn time $V_{soj}^{(2)} = 327.71$, i.e., increasing the average arrival flow of 7.7% (change $\lambda$ from 13 to 14) leads to an increase in

Fig. 3. Loss probability of an arbitrary type 1 customer at the first stage, the average intensity of flow of customers, which receive service at the first stage, the loss probability of an arbitrary type 1 customer at the second stage and the probability that an arbitrary type 1 customer will go to the buffer and leave it due to impatience as functions of the average arrival intensity for different MMAPs.

Fig. 4. Dependence of the average waiting time of type 2 customers, the average number of type 1 customers in the buffer 1 and the loss probabilities $P_{loss}^{1}$ and $P_{loss}^{2}$ on the intensity of impatience $\alpha$ for PH service processes with different coefficients of variation.
the average sojourn time $V_{\text{soj}}$ by more than 62 times. Prediction of such behavior without corresponding calculations is unrealistically difficult task, so this proves the necessity of the presented research.

The numerical results also show that the following equalities hold true:

$$V_{\text{soj}} - b_1^k = V_{\text{soj}} = \frac{N_{\text{buffer}}}{\rho_2}.$$  
$$N_{\text{buffer}} = \frac{p_{1,\text{imp}}}{\lambda_{\text{out}}}.$$  

The first equality is the famous Little formula. The second equality can be explained as follows. Let us rewrite it in the form $dN_{\text{buffer}} = \lambda_{\text{out}} p_{1,\text{imp}}$. The left and the right sides of the equality define the intensity of type 1 customers leaving buffer 1 due to impatience, so the equality is correct.

In Experiment 2, let us show the effect of the coefficient of variation in the service processes of type 1 and type 2 customers at the second stage on the main performance measures of the system. To this end, we consider three cases of PH service processes of type 1 and type 2 customers.

The first case corresponds to the exponential service time distributions of type 1 and type 2 customers. We assume that the service time of type 1 customers is defined by the vector $\mathbf{p}_1 = (1)$ and the matrix $\mathbf{S}_1 = (-1)$. The service time of type 2 customers is defined by the vector $\mathbf{p}_2 = (1)$ and the matrix $\mathbf{S}_2 = (-0.5)$. These processes have the same coefficient of variation $c_{\text{var}} = 1$.

In the second case, we assume that the service time distribution of type 1 customers is characterized by the vector $\mathbf{p}_1 = (1, 0)$ and the matrix $\mathbf{S}_1 = (-2, 0, -2)$. The service time distribution of type 2 customers is characterized by the vector $\mathbf{p}_2 = (1, 0)$ and the matrix $\mathbf{S}_2 = (-1, 0, -1)$. These processes have the same coefficient of variation $c_{\text{var}} = 0.5$.

In the third case, we assume that the service time distribution of type 1 customers is characterized by the vector $\mathbf{p}_1 = (0, 1.0, 0.9)$ and the matrix $\mathbf{S}_1 = (-0.2339, 0.19389, -2.54192)$, and the service time distribution of type 2 customers is characterized by the vector $\mathbf{p}_2 = (0.1, 0.9, 0)$ and the matrix $\mathbf{S}_2 = (-0.11659, 0.0056, -0.69994, -1.27096)$. These processes have the same coefficient of variation $c_{\text{var}} = 5$.

Note that in all cases under consideration the average service time of type 1 customers $b_1^k = 1$ and the average service time of type 2 customers $b_2^k = 0.5$, but the service processes have different values of the coefficient of variation $c_{\text{var}}$.

Let us fix the number of servers at the first stage $R = 5$, the service intensity at the first stage $\mu = 0.8$, the probability $q = 0.5$, the number of servers at the second stage $N = 4$, the buffer capacity for type 1 customers $K = 5$. We assume that arrivals are defined by the MMAP$^k$-1 arrival process presented in the first experiment with the arrival intensity $\lambda = 1$, the coefficient of correlation $c_{\text{corr}} = 0.4$ and the coefficient of variation $c_{\text{var}} = 12.39$.

Let us vary the intensity of impatience $\alpha$ in the interval $[0.01, 4]$. Fig. 4 illustrates the dependence of the average waiting time $V_{\text{wait}}$ of type 2 customers, the average number $N_{\text{buffer}}$ of type 1 customers in buffer 1, the probability $p_{1,\text{out}}$ that an arbitrary type 1 customer will be lost at the entrance to the second stage, and the loss probability $P_{\text{loss}}$ of an arbitrary type 1 customer at the second stage on the intensity of impatience $\alpha$ for PH service processes with different coefficients of variation.

As it is seen from Fig. 4, with growth of the intensity of impatience $\alpha$ the average sojourn time of type 2 customers decreases. This finding can be explained as follows. When the intensity of impatience increases, the loss probability of type 1 customers at the second stage $P_{\text{loss}}$ also increases. Because type 1 customers have a priority over type 2 customers, the loss of type 1 customers decreases the average waiting time (and also the average sojourn time) of type 2 customers. For the same reason the average number of type 1 customers in buffer 1 essentially decreases with an increase in the intensity of impatience which in turn leads to a decrease in the loss probability of type 1 customers due to buffer 1 overflow.

Another important conclusion from Fig. 4 is that the coefficient of variation in the service process affects the system performance measures. So, for adequate prediction of the system operation it is necessary to take into account the variation in the service process.

8. Conclusion

A tandem queueing system with Marked Markovian arrival flow and generalized phase-type service time distribution is studied. The process of the system states is analyzed, and the ergodicity condition is derived. Some key performance measures are obtained. The Laplace–Stieltjes transform of the sojourn time distribution of type 2 customers is derived. The numerical results show the importance of taking into account the correlation in the arrival flow and the variance of the service time.

Acknowledgements

This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (Grant No. 2011-0015214).

The authors are very thankful to three anonymous referees for valuable remarks and suggestions. Their account led to improvement of presentation of the results.

References


Kim, C. S., Dudin, A. N., Klimenok, V. I., & Tsarenkov, G. V. (2010). Investigation of the $\text{BMAP/G/1/N} \rightarrow \cdot (\text{PH}/1/M$ tandem queue with retials and losses. Applied Mathematical Modelling, 34, 2926–2090.


Kim, C. S., Dudin, S. A., Tararmin, O. S., & Baek, J. (2013b). Queueing system $\text{MAP/PH1/1/N} \rightarrow \cdot \text{R with impatient heterogeneous customers as a model of call center. Applied Mathematical Modelling, 37, 958–976.}


