A tandem GI/PH/1 → •/PH/1/0 queue with blocking

Valentina Klimenoka, Chesoong Kim, Olga Dudina, Alexander Dudin

Department of Applied Mathematics and Computer Science, Belarusian State University, Minsk, Belarus
Department of Industrial Engineering, Sangji University, Wonju, Kangwon 220-702, South Korea

ABSTRACT

Tandem queues are widely used in mathematical modeling of random processes describing the operation of manufacturing systems, supply chains, computer and telecommunication networks. Although there exists a lot of publications on tandem queueing systems, analytical research on tandem queues with non-Markovian input is very limited. In this paper, the results of analytical investigation of two-node tandem queue with arbitrary distribution of inter-arrival times are presented. The first station of the tandem is represented by a single-server queue with infinite waiting room. After service at the first station, a customer proceeds to the second station that is described by a single-server queue without a buffer. Service times of a customer at the first server have PH (Phase-type) distributions. A customer, who completes service at the first server and meets a busy second server, is forced to wait at the first server until the second server becomes available. During the waiting period, the first server becomes blocked, i.e., not available for service of customers. We calculate the joint stationary distribution of the system states at the embedded epochs and at arbitrary time. The Laplace–Stieltjes transform of the sojourn time distribution is derived. Key performance measures are calculated and numerical results presented.

1. Introduction

Tandem queueing systems are the simple queueing networks consisting of finite number of nodes in series. Such systems can be used for modeling real-life networks having a linear topology as well as for validation of general decomposition algorithms in networks (see, e.g., [1–4]). In case all nodes of a tandem are modeled by the $M/M/1$ queue, the tandem can be considered as elementary Jackson network and a joint distribution of the nodes has a product form. In case of more complicated arrival and service processes, calculation of joint stationary distribution becomes a difficult problem which is often not amenable to analytical treatment because of its complexity. This is particularly true for queues where the input process or (and) service processes at the nodes are non-Markovian.

Tandem queueing systems have found much interest in the literature. An extensive survey of early papers on tandem queues can be seen in [5]. Most of these papers are devoted to exponential queueing models in steady state. Over the last two decades, interesting analytical results have been obtained in investigations of complicated two-node tandem queues with a batch Markovian arrival process. For references see, e.g., [6–8].

However, such queues are of limited use for modeling of real-life networks with non-Markovian flows, e.g., renewal flows with deterministic, uniform, log-normal, Weibull etc. distribution of inter-arrival time intervals. To the best of our knowledge, there are a few papers dealing with analytical study of tandem queues with non-Markovian input. Papers by Avi-Itzhak...
and co-authors (see, e.g. [9]) are devoted to non-stochastic queues with blocking, arbitrary input flow and regular service time. The paper [10] deals with a two-node tandem queue with renewal input, exponentially distributed service time and infinite waiting rooms at both stations. Analogous tandems are studied in the publications by Knessl and Tier (see, e.g. [11]). The authors investigate joint distribution of the queue length at the stations under overload condition using a diffusion approximation method. In the paper [12] the asymptotic behavior of the steady-state measures of the two-node tandem with renewal input and infinite waiting rooms at both stations is analyzed.

The queueing model under study takes into account the main feature of the service process in tandem queues – blocking after service. Blocking mechanism is used in tandem queues with finite intermediate buffer or without such a buffer. In our case, there is no buffer between the servers, so that a customer having completed processing at the first server and meeting a busy second server is forced to wait at Station 1 occupying the first server until the second server becomes available. Thus, the first server becomes blocked or not available for service of customers that arrive at Station 1. Tandem queues without an intermediate buffer arise, e.g., in modeling production lines where two technological operations must follow one another immediately to prevent degradation in the viscosity of a material.

In queues with blocking, service processes at both stations are strongly mutually dependent. The main contribution of this paper is that we are able to take into account this dependence in tandem queue $GI^{\text{PH}}=\text{PH}=1$ through the analysis and formal description of the output flow from Station 1. This allows us to construct the embedded $GI/M/1$-type Markov chain, which describes the process of the system states at arrival epochs, avoiding the “complex overhead computations” mentioned in the well-known monograph by Neuts [13]. Note that, to the best of our knowledge, there are no analogous results in the literature even for the case of exponentially distributed service time at the servers. Further, we derive the ergodicity condition of the Markov chain and calculate its stationary distribution using the algorithm given in [13]. Based on this distribution we derive the stationary distribution of the system states at an arbitrary time and the Laplace–Stieltjes transforms of the distribution of the sojourn time at the stations and in the whole system. A number of important performance measures are also derived. All proposed algorithms are realized as computer program in the frameworks of software “Sirius +”, see [14]. Using this software we carried out numerical work to analyze the behavior of key performance measures. Some of the numerical results are presented in the paper.

2. Model description

We consider $GI/PH=1 \rightarrow \bullet/PH=1/0$ tandem queue. Station 1 of the tandem is represented by the $GI/PH=1$ queue. The inter-arrival times in the input flow at Station 1 are independent random variables with general distribution $A(t)$ and finite first moment $a_1 = \int_0^\infty \text{td}A(t)$. After getting the service at Station 1, a customer has to be served at Station 2 that is represented by a single-server queue without a buffer. In case the customer completes the service at the first server and meets the second server being idle, he/she occupies it immediately. Otherwise, the customer is forced to wait at Station 1 occupying the server space until the second
server becomes available. Thus, the first server becomes blocked or not available for service of customers. We assume that a customer who caused the blocking stays at the first server until the end of the blocking.

The service times of a customer at both stations have PH distribution.

Service time having PH distribution with an irreducible representation \((\beta, S)\) can be interpreted as time until the underlying Markov chain \(m_t, t \geq 0\), with the finite state space \(\{1, \ldots, M, e\}\) reaches the single absorbing state \(e\) conditional that the initial state of this process is selected among the states \(\{1, \ldots, M\}\) according to probabilistic row vector \(\beta\). Transition rates of the process \(m_t\), within the set \(\{1, \ldots, M\}\) are defined by the sub-generator \(S\) and transition rates into the absorbing state (which lead to service completion) are given by the entries of the column vector \(S_e = -Se\). Here and in the sequel \(e\) is a column vector consisting of 1’s. For more information about PH distribution see, e.g. [13].

We assume that service process at the rth, \(r = 1, 2\), server has PH\(^{(1)}\) distribution with an irreducible representation \((\beta_r, S^{(r)})\) and is governed by the Markov chain \(m_r^{(r)}, t \geq 0\), with the state space \(\{1, \ldots, M_r, e^{(r)}\}\) where the state \(e^{(r)}\) is an absorbing one. The mean service times are calculated by \(b_i^{(r)} = \beta_r (-S^{(r)})^{-1} e\). \(r = 1, 2\).

The goal of the paper is to get the stationary distribution of the system, the main performance measures and the stationary distribution of the sojourn time.


Let \(\Omega_2, k = 1, 2, 3\), be the sets of pairs formed by the states of the PH\(^{(1)}\), PH\(^{(2)}\) service processes as follows:

\[
\Omega_1 = \{(m^{(1)}, m^{(2)}), m^{(1)} = T, m^{(2)} = M_r\},
\]

\[
\Omega_2 = \{(m^{(1)}, e^{(2)}), m^{(1)} = T, m^{(2)} = M_r\}, \quad \Omega_3 = \{(e^{(1)}, m^{(2)}), m^{(2)} = M_r\}.
\]

The notation \(m^{(1)} = T, M_r\) means that the parameter \(m^{(1)}\) takes the values in the set \(\{1, 2, \ldots, M_r\}\).

The process \(\xi_t, t \geq 0\), of the system states at an arbitrary time \(t\) is defined by the state space \(\{(0, m^{(2)}), m^{(2)} = 1, \ldots, M_r, e^{(2)}; (i, m), i \geq 1, m \in (\Omega_1 \cup \Omega_2 \cup \Omega_3)\}\).

The states of this process mean the following:

- if \(\xi_0 = (0, m^{(2)})\) at time \(t\) then Station 1 is empty and the service process at the server of Station 2 is in the phase \(m^{(2)}\);
- if \(\xi_t = (i, m), i \geq 1, m = (m^{(1)}, m^{(2)}) \in \Omega_1\) then \(i\) customers stay at Station 1 and servers of both stations are busy with PH\(^{(1)}\), PH\(^{(2)}\) in the phases \(m^{(1)}\), \(m^{(2)}\) respectively;
- if \(\xi_t = (i, m), i \geq 1, m = (m^{(1)}, e^{(2)}) \in \Omega_2\) then \(i\) customers stay at Station 1, the server of Station 1 is busy with PH\(^{(1)}\) in the phase \(m^{(1)}\) and the server of Station 2 is idle;
- if \(\xi_t = (i, m), i \geq 1, m = (e^{(1)}, m^{(2)}) \in \Omega_3\) then \(i\) customers (including the blocked customer) stay at Station 1, the server of Station 1 is blocked and the server of Station 2 is busy with PH\(^{(2)}\) in the phase \(m^{(2)}\).

Fig. 2. The system performance measures as functions of the system load for different variation in the input process.
It is clear that the process \( q_t \) is non-Markovian if the input flow does not possess the Markov property. We intend to study the stationary distribution of this process through the investigation of the embedded Markov chain.

Let \( t_n \) denote the instant of the \( n \)th arrival at Station 1, \( n \geq 1 \). To be able to construct an embedded Markov chain we need to know the distribution of the number of customers proceeding to Station 2 during an inter-arrival time. However, calculation of such a distribution is a difficult problem, because, due to the possibility of blocking, we do not know a priori how long a customer occupies the first server before he/she proceeds to the second server. This time may consist of only service time or service time and blocking time depending on whether or not the second server is free at the epoch of the service completion at the first server. Furthermore, the duration of the blocking time, if any, depends on the state of the \( \text{PH}^{(2)} \) service process at the second server at the instant of service completion at the first server.

To solve the problem, we first investigate the output flow from Station 1 in the busy period at this station. To this end, we consider the process \( m_t; t \geq 0 \), with the state space \( \{X_1; X_2; X_3\} \). In the busy period at Station 1, the process \( m_t; t \geq 0 \), behaves as a Markovian one. Moreover, in such a period the output flow from Station 1 (which, at the same time, is the input flow at Station 2) can be described in terms of a MAP (Markovian Arrival Process) with underlying process \( m_t; t \geq 0 \). Such arrival process was introduced as a versatile Markovian point process (VMPP) by Neuts in the 70th. Ramaswami presented in [15] detailed analysis of a single server queue with such an arrival process and general service time distribution. The original development of VMPP contained extensive notations; however these notations were simplified greatly in [16] and ever since this process bears the name Markovian arrival process (MAP). Currently, the mathematical model of a MAP is well known and widely used in modern queueing systems theory. Detailed description of a MAP can be found, e.g., in [17,16].

In brief, the arrivals in a MAP occur only at the epochs of transitions of a continuous time Markov chain with a finite state space which is called the underlying process of the MAP. Let \( \mathcal{N} \) be the state space dimension. Then the MAP is completely defined by the \( \mathcal{N} \)-size square matrices \( D_0, D_1 \). The matrix \( D_0 \) governs the transitions of the chain corresponding to no arrival while the matrix \( D_1 \) governs those corresponding to an arrival.

To start a formal description of output flow from Station 1 in terms of a MAP, we arrange the states of the process \( m_t \) as follows. Pairs within each set \( \Omega_k \), \( k = 1, 2, 3 \), are listed in the lexicographic order and then the obtained sets are arranged in increasing order of \( k \). As a result, pairs from the set \( \Omega_1 \) are assigned numbers 1,\( \ldots, M_1M_2 \), pairs from \( \Omega_2 \) – numbers \( M_1M_2 + 1,\ldots,M_1M_2 + M_1 \), and pairs from \( \Omega_3 \) – numbers \( M_1M_2 + M_1 + 1,\ldots,M_1M_2 + M_1 + M_2 \).

Remark 1. In what follows we will refer to the pair that appears on the \( m \)th position in the resulting list as the \( m \)th state of the process \( m_t; t \geq 0 \).

Lemma 1. In the busy period at Station 1, the output flow from this station is the MAP with the underlying process \( m_t; t \geq 0 \), and the transition intensity matrices
\[
D_0 = \begin{pmatrix}
S^{(1)} \otimes S^{(2)} & I_{M_1} \otimes S^{(2)} & S^{(1)} & I_{M_2} \\
0 & S^{(1)} & 0 & S^{(2)}
\end{pmatrix}, \quad D_1 = -D_0 e^b^{(s)},
\]

where \(S^{(r)} = -S^{(r)} e, r = 1, 2, b^{(s)} = (b_1 \otimes b_2, \theta_{M_1+M_2}). \otimes \) and \(\otimes\) are symbols of Kronecker’s sum and product respectively, \(I_d\) is an identity matrix of size \(d\); \(\theta_d\) is a row vector of zeros of size \(d\).

**Proof.** The blocks of the matrix \(D_0\) define the rates of the process \(m_t, t \geq 0\), transitions within the sets \(\Omega_k, k = 1, 2, 3\), and between them. Such transitions do not lead to generation of a customer in the output flow from Station 1. The more detailed description of the blocks of the matrix \(D_0\) is as follows:

- the block \(S^{(1)} \otimes S^{(2)}\) governs the transitions corresponding to a change of the phase of the service process at one of the busy servers without the service completion;
- the blocks \(I_{M_1} \otimes S^{(2)}\) and \(S^{(1)} \otimes I_{M_2}\) govern the transitions corresponding to the service completion at one of the busy servers (the second and the first one respectively) without generation of a customer in output flow;
- the entries of the second (third) block row of \(D_0\) define the transition rates from the states corresponding to busy first server and idle second server (blocked first server and busy second server). The non-zero transition rates are combined into diagonal blocks \(S^{(1)}\) and \(S^{(2)}\) that govern the transitions of the phase of the service at the first and at the second server respectively without the service completion.

Entries of the column vector \(-D_0 e\) corresponding to the states \(M_1M_2 + 1, \ldots, M_1M_2 + M_1\) of the underlying process define the rates of the transitions corresponding to the service completion at the first server given this server was busy and the second server was idle. Entries of this vector with numbers \(M_1M_2 + M_1 + 1, \ldots, M_1M_2 + M_1 + M_2\) define the rates of the transitions corresponding to the service completion at the second server given this server was busy and the first server was blocked. As a result of such transitions a customer, who has been occupying the first server, moves to the second server and the phase of the \(PH^{(2)}\) service process is installed according to the probabilistic row vector \(b_2\). Simultaneously, the first server is occupied by a customer waiting in the queue at Station 1 and the initial phase of \(PH^{(1)}\) service processes is installed according to probabilistic vector \(b_1\). Thus, the matrix \(D_1 = -D_0 e^{b^{(s)}}\) governs the transitions of the process \(m_t, t \geq 0\), corresponding to generation of a customer in the output flow from Station 1. \(\square\)

---

**Fig. 4.** The system performance measures as functions of the system load for different service time variation at Station 1.
Corollary 1. In the busy period at Station 1, time intervals between the events in output flow from this station, beginning from the second interval, are i.i.d. random values having the PH distribution with underlying process $m_\nu$, sub-generator $D_0$ and stochastic vector $p^{(g)}$. Each of these intervals consists of the service time of a customer at the first server and possible blocking time and can be considered as generalized service time of the customer at Station 1.

Corollary 2. In the busy period at Station 1, the distribution of the number of customers moving from Station 1 to Station 2 during an inter-arrival time is defined as the distribution of the number of events in the MAP defined by Lemma 1.

Now we are able to construct the embedded Markov chain. Let $n_\nu$ be the number of customers at Station 1 (including the blocked customer, if any) at the epoch $t_n = 0$, $m_n$ be the state of the process $m_\nu$, $t \geq 0$, at the epoch $t_n$, $n \geq 1$.

It is easy to see that the process $(\nu n, m)_{t \in \mathbb{R}}$ is an irreducible Markov chain with the state space \{(0, m), m = 1, \ldots, K_0; (i, m), i > 0, m = 1, \ldots, K\} where $K_0 = M_1M_2 + M_1$ and $K = M_1M_2 + M_1 + M_2$.

Theorem 1. The transition probability matrix of the chain $(\nu n, n \geq 1$, has the following block structure:

$$P = \begin{pmatrix} B_0 & A_0 & 0 & 0 & \ldots \\ B_1 & A_1 & A_0 & 0 & \ldots \\ B_2 & A_2 & A_1 & A_0 & \ldots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix},$$

where

$$A_l = \int_0^\infty P(l, t) dA(t), \quad B_l = \int_0^\infty \int_0^t P(l, x) (-D_0) e^{D_0 t} (t-x) dA(t), \quad l \geq 0,$$

$$B_0 = I B_0, \quad \Lambda_0 = I \Lambda_0, \quad \Phi'(y) = (\beta_1 \otimes \beta_2 e^{D_1 y}, \beta_1 (1 - \beta_2 e^{D_1 y} e)),$$

$$I = \begin{pmatrix} I_{M_1M_2} & 0 & 0 \\ 0 & I_{M_1} & 0_{M_1 \times M_2} \end{pmatrix}.$$ 

matrices $P(l, t), l \geq 0$, are defined by $\sum_{l=0}^{\infty} P(l, t) z^l = e^{D_0 + D_1 z I}$ and $0_{M_1 \times M_2}$ is the matrix of size $M_1 \times M_2$ consisting of zeroes.

Proof. To clarify the expressions for transition probability matrices $A_l$ and $B_l$, we register the following probabilistic interpretation of the vectors and matrices appearing in these expressions.
The \((m, m')\)th entry of the matrix \(P(l, x)\) is the conditional probability that, in the time interval \((0, x)\) belonging to the busy period at Station 1, \(l\) customers move from this station to Station 2 and the state of the underlying process of the MAP at the moment \(x\) is \(m'\) conditional that the state of this process at the moment 0 was \(m\).

The \(m'\)th entry of the vector \((-D_0)\mathbf{e}\mathbf{dx}\) is the probability that a customer, which has been occupying the first server at time \(x\), moves to the second server in the interval \((x, x + dx)\) conditional that the MAP was in the state \(m'\) at the moment \(x\).

The \(r\)th entry of the vector \(\beta_2 e^{\mathbf{C}_0} \mathbf{y}\) is the probability that a processing of a customer at the second server is not completed till the moment \(y\) and the process \(m'^{(2)}\), \(t \geq 0\), is in the state \(r\) at that moment conditional that the customer started the service at time 0. The scalar \(1 - \beta_2 e^{\mathbf{C}_0} \mathbf{y}\) is the probability that the second server completes the service of a customer until the time \(y\) conditional that the service started at time 0.

Thus, given the first server is idle at time 0 and the second server starts the processing of a customer at that time, the vector \(\mathbf{b}_2 e^{\mathbf{S}} \mathbf{y}\) installs, in accordance with the vector \(\mathbf{b}_1\), the state of service process for the future customer at the first server and, if the second server continues the service of the customer, defines the state of its service process at time \(y\).

Taking into account the above probabilistic interpretations, it is not difficult to derive expressions (1) for the matrices \(A_l\) and \(B_l\).

The matrix \(\mathbf{I}\) is used to delete the rows of the matrices \(B_0\) and \(A_0\) that correspond to states \((0, m)\), \(m = K_0 + 1, K\), of the chain \(\zeta_n\). The chain cannot enter such states since this would mean that the first server is idle just after arrival of a customer into empty Station 1.

**Corollary 3.** The process \(\zeta_n\), \(n \geq 1\), is the Markov chain of the GI/M/1 type.

**Proof.** The process \(\zeta_n\), \(n \geq 1\), is an irreducible aperiodic Markov chain whose transition probability matrix \(P\) has the block lower-Hessenberg structure where the blocks \(P_{ij}, i, j > 0\), are represented as functions of \(i - j\). So, this chain belongs to the class of GI/M/1 type Markov chains which was extensively studied by Neuts [13].

**Theorem 2.** Stationary distribution of the Markov chain \(\zeta_n\), \(n \geq 1\), exists if and only if the inequality

\[
\rho = a_1^{-1} b_1^{(g)} < 1
\]

is fulfilled. Here \(b_1^{(g)} = \beta^{(g)} (-D_0)^{-1} e\) is the mean value of generalized service time.

**Proof.** Let \(A(z) = \sum_{l=0}^{\infty} A_l z^l, |z| \leq 1\), be the generating function of the transition probability matrices \(A_l\), \(l \geq 0\). It follows from [13], that the necessary and sufficient condition for the stationary distribution existence is the fulfillment of the inequality \(z A(1) e > 1\).

![Fig. 6.](image-url) The system performance measures as functions of the system load for different service time variation at Station 2.
where the vector \( \mathbf{x} \) is the unique solution of the system
\[
\mathbf{x} A(1) = \mathbf{x}, \quad \mathbf{xe} = 1.
\] (4)

It our case \( A(z) = \int_0^\infty e^{(s+b)z}f(t)dt \).

Let the vector \( \mathbf{x} \) be of the form \( \mathbf{x} = (b)\ast^{-1}b\ast(D_0)^{-1} \). By the direct substitution, we verify that such a vector provides the unique solution of system (4). Substituting the vector \( \mathbf{x} \) and the expression for \( A'(1) \) into inequality (3) and taking into account the relations \( \beta \ast(D_0)^{-1}(D_0 + D) = 0 \), \( (D_0 + D)\mathbf{e} = \mathbf{0} \), we reduce this inequality to the form (2).

Denote the stationary state probabilities of the chain \( \xi_n \) by \( \pi(0,m), \quad m = 1,2,\ldots, \pi(i,m), i > 0, \quad m = 1,2,\ldots, \pi(i,0) \). Introduce the notation for row vectors of these probabilities
\[
\pi_0 = (\pi(0,1), \pi(0,2), \ldots, \pi(0,K_0)), \quad \pi_i = (\pi(i,1), \pi(i,2), \ldots, \pi(i,K)), \quad i > 0.
\]

**Theorem 3.** The stationary probability vectors \( \pi_i, \quad i \geq 0 \), are calculated as follows:
\[
\pi_i = \pi_i \mathcal{R}^i, \quad i \geq 2,
\] (5)
where the matrix \( \mathcal{R} \) is the minimal non-negative solution of the matrix equation
\[
\mathcal{R} = \sum_{j=0}^{\infty} \mathcal{R}^j A_j
\]
and the vector \( (\pi_0, \pi_1) \) is the unique solution of the system
\[
(\pi_0, \pi_1) = (\pi_0, \pi_1)\mathcal{T}, \quad \pi_0 \mathbf{e} + \pi_1 (I - \mathcal{R})^{-1} \mathbf{e} = 1
\]
where
\[
\mathcal{T} = \left( \begin{array}{cc}
\mathcal{B}_0 & \mathcal{A}_0 \\
\sum_{j=1}^{\infty} \mathcal{R}^{j-1} \mathcal{B}_j & \sum_{j=1}^{\infty} \mathcal{R}^{j-1} A_j
\end{array} \right).
\]
The proof of the theorem is immediate from the results of [13].

4. Stationary distribution at an arbitrary time

In this section, we calculate the stationary distribution of the non-Markovian process \( \xi_t, \quad t \geq 0 \), of the system states at an arbitrary time based on the stationary distribution of the embedded Markov chain \( \xi_{n,t} \), \( n \geq 1 \). Enumerate the states of the process \( \xi_t, \quad t \geq 0 \), corresponding to the fixed value \( i \) of the number of customers at Station 1 in the lexicographic order assuming that the vectors \( \mathbf{m} \in (\Omega_1 \cup \Omega_2 \cup \Omega_3) \) are arranged as above (before Remark 1) and refer to these states as level \( i \).

Let \( \mathbf{p}_i \) be the vector of steady-state probabilities the states that belong to the level \( i \).

**Theorem 4.** The vectors \( \mathbf{p}_i, \quad i \geq 0 \), are calculated as follows:
\[
\mathbf{p}_0 = a_0^{-1}(\pi_0 \mathbf{l} \Phi_0 + \pi_1 \sum_{j=1}^{\infty} \mathcal{R}^{j-1} \Phi_1),
\] (6)
\[
\mathbf{p}_1 = a_1^{-1}[\mathbf{p}_0 \mathbf{l} - \pi_1 (I - \mathbf{e} \beta \ast)(-D_0)^{-1},
\] (7)
\[
\mathbf{p}_i = a_i^{-1}[\mathbf{p}_{i-1} \mathbf{l} - \mathcal{R} (I - \mathbf{e} \beta \ast)(-D_0)^{-1}, \quad i > 1,
\] (8)
where
\[
\Phi_1 = \int_0^\infty \int_0^t P(l,x)(-D_0)\mathbf{e} \beta \ast(l-x) (1 - A(t))dt, \quad \mathcal{S}^{(2)} = \left( \begin{array}{c}
\mathcal{S}_0^{(2)} \\
\mathcal{S}_1^{(2)}
\end{array} \right).
\]

**Proof.** Using the definition of semi-regenerative processes given in [18], it can be verified that the process \( \xi_t \) is a semi-regenerative one with the embedded Markov renewal process \( \{\xi_{n,t}, t \geq 0\} \), \( n \geq 1 \). The stationary distribution \( \mathbf{p}_i, \quad i \geq 0 \), exists if the process \( \{\xi_{n,t}, n \geq 1, \) is irreducible aperiodic recurrent and the value \( a_1 \) of the mean inter-arrival time at Station 1 is finite. All these conditions hold if inequality (2) is satisfied.

By the ergodic theorem for semi-regenerative processes, see [18], the stationary distribution of the process \( \xi_t, \quad t \geq 0 \), can be related to the stationary distribution of the embedded Markov chain \( \xi_{n,t}, n \geq 1 \).
Before we can use the theorem we need to define some probabilistic functions associated with the process $\xi_t$, $t \geq 0$. Introduce the matrices $K_i(l, i) = 0, t \geq 0$. The nontriviality of matrices $K_i(l, i)$ is due to the conditional probability that, given that time 0 is an instant of an arrival at Station 1 and the embedded Markov chain $\xi_t$ is in the state $(l, m)$ at that time, the next arrival at Station 1 occurs later than $t$, and the process $\xi_t$ is in the $m$th state from the level $i$ at time $t$.

By the ergodic theorem, the vectors $p_i$, $i \geq 0$, are expressed in terms of the stationary distribution $\pi_i$, $i \geq 0$, of the embedded Markov chain $\xi_t$ as follows:

$$p_i = a_i^{-1} \sum_{l=0}^{\infty} \pi_l \int_0^\infty K_i(l, i) dt, \quad i \geq 0. \tag{9}$$

Taking into account the probabilistic sense of matrices clarified in the proof of Lemma 1, we easy derive the expressions for the matrices $K_i(l, i)$ and, substituting them into (9), rewrite the relations (9) as follows:

$$p_0 = a_1^{-1} \left[ \pi_0 \int_0^\infty \int_0^t P(0, x)(-D_0) e^{S^0(t-x)}(1 - A(t)) dt + \sum_{l=0}^{\infty} \pi_l \int_0^t P(l, x)(-D_0) e^{S^0(t-x)}(1 - A(t)) dt \right], \tag{10}$$

$$p_1 = a_1^{-1} \left[ \pi_0 \int_0^\infty P(0, t)(1 - A(t)) dt + \sum_{l=1}^{\infty} \pi_l \int_0^t P(l, t)(1 - A(t)) dt \right], \tag{11}$$

$$p_i = a_1^{-1} \sum_{l=1}^{\infty} \pi_l \int_0^\infty P(l - i + 1, t)(1 - A(t)) dt, \quad i > 1. \tag{12}$$

Substituting expressions (5) for the vector $p_i$, $i > 0$, into (10)--(12), after algebraic transformations we get formulas (6)--(8). \(\square\)

**Corollary 4.** The following relation holds true:

$$\sum_{i=1}^{\infty} p_i = a_1^{-1} [\pi_0 (I - e^{S^0}) + \beta e^0] (-D_0)^{-1}. \tag{9}$$

**Proof.** This formula is obtained from (6)--(8) as a result of equivalent transformations of the expression for vector $\sum_{i=1}^{\infty} p_i$. \(\square\)

### 5. Stationary performance measures

As soon as the vectors $\pi_i, p_i$, $i \geq 0$, have been calculated, we are able to find various performance measures of the system. Below we present some of them. Nontrivial performance measures will be given with brief explanations.

- **Mean number of customers at Station 1 at the arrival epoch and at an arbitrary time**

  $$L^{(s)} = \pi_1 (I - R)^{-1} e,$$

  $$L = L^{(s)} + a_1^{-1} [\pi_0 (I - R)^{-1} (-D_0)^{-1} e].$$

- **Probability that an arbitrary customer will be successfully served at both stations without waiting and blocking**

  $$P_{\text{succ}} = \pi_0 \begin{pmatrix} (S^{(1)} \oplus S^{(2)})^{-1} (I_{M_1} \otimes S^{(2)}) & 0 \\ 0 & I_{M_1} \end{pmatrix} e.$$

- **Probability that the server of Station 1 is blocked at an arbitrary time**

  $$P_{\text{block}} = \sum_{i=1}^{\infty} p_i \text{diag} \{0_{M_1/M_2+1}, I_{M_2} \} e. \tag{13}$$

The brief explanation of formula (13) is as follows. The diagonal matrix $\text{diag} \{0_{M_1/M_2}, 0_{M_1}, I_{M_2} \}$ selects the entries of the vector $\sum_{i=1}^{\infty} p_i$ corresponding to the joint probability that, at an arbitrary time, the first server is blocked and the second server is busy with the $PH^{(2)}$ service process taking values in the set $\{1, \ldots, M_2 \}$. As a result, we get the row vector of such joint probabilities. Multiplying this vector by $e$, we obtain the sum of its entries that evidently gives the probability $P_{\text{block}}$. 
Remark 2. The analogous reasonings are used to derive formulas for the probabilities $p_{\text{busy}}^{(1,2)}, p_{\text{busy, idle}}^{(1,2)}, p_{\text{idle, busy}}^{(1,2)}, p_{\text{idle}}^{(1,2)}$ that are listed below.

- Probability that both servers are busy at an arbitrary time
  \[ p_{\text{busy}}^{(1,2)} = \sum_{i=1}^{\infty} p_i \text{diag} \{ I_{M_1, M_2}, 0_{M_1, M_2} \} \mathbf{e}. \]

- Probability that the server of Station 1 is busy and the server of Station 2 is idle at an arbitrary time
  \[ p_{\text{busy, idle}}^{(1,2)} = \sum_{i=1}^{\infty} p_i \text{diag} \{ 0_{M_1, M_2}, I_{M_1}, 0_{M_2} \} \mathbf{e}. \]

- Probability that the server of Station 1 is idle and the server of Station 2 is busy at an arbitrary time
  \[ p_{\text{idle, busy}}^{(1,2)} = p_0 \text{diag} \{ I_{M_1}, 0 \} \mathbf{e}. \]

- Probability that both stations are empty at an arbitrary time
  \[ p_{\text{idle}}^{(1,2)} = p_0 \text{diag} \{ 0_{M_2}, 1 \} \mathbf{e}. \]

6. Sojourn time distribution

Let $V_1(t)$ and $V(t)$ be the distribution functions of sojourn time at Station 1 and in the tandem respectively. Let also $V_1'(u) = \int_0^u e^{-ut} dV_1(t)$ and $V'(u) = \int_0^u e^{-ut} dV(t)$ be the Laplace–Stieltjes transforms of these distributions.

**Theorem 5.** The Laplace–Stieltjes transform of the stationary distribution of the sojourn time is calculated as follows:

(i) at Station 1
\[ V_1'(u) = [\pi_0 I + \pi_1 (I - R \varphi_1(u))^{-1} \varphi_1(u)](ul - D_0)^{-1} (-D_0) \mathbf{e}; \]  
\[ (14) \]

(ii) in the whole system
\[ V'(u) = V_1'(u) \varphi_2(u). \]

Here $\varphi_1(u) = \beta_1'(ul - D_0)^{-1} (-D_0) \mathbf{e}$ and $\varphi_2(u) = \beta_2(u - S^{(2)})^{-1} S_0^{(2)}$ are the Laplace–Stieltjes transforms of distribution of generalized service time and service time at Station 2 respectively.

**Proof.** (i) The sojourn time of a tagged customer, which arrives at Station 1 when $i$ customers stay at this station, consists of the residual time until the customer in service proceeds to the second server, the generalized service times of $i - 1$ customers staying in the queue and the generalized service time of the tagged customer.

Let $\pi_i(u)$ be the row vector of size $K$ which $m$th entry is the Laplace–Stieltjes transform of the sojourn time distribution of a customer at Station 1 under condition that, at the beginning of this time, $i$ customers stay at the station and the underlying process of the MAP defined by Lemma 1 is in the state $m$.

Taking into account the structure of the sojourn time and using Lemma 1 and Corollary 1, we easily derive the following expression for the vectors $\mathbf{v}_i(u)$:
\[ \mathbf{v}_i(u) = \int_0^\infty e^{-ut} e^{\theta(t)} (-D_0) \mathbf{e} dt (\varphi_1(u))^i, \quad i \geq 0, \]  
\[ (15) \]

where the Laplace–Stieltjes transform $\varphi_1(u)$ of the generalized service time distribution is defined in the statement of the theorem.

The Laplace–Stieltjes transform $V_1'(u)$ is calculated from the formula of total probability
\[ V_1'(u) = \pi_0 I \mathbf{v}_0(u) + \sum_{i=1}^{\infty} \pi_i \mathbf{v}_i(u). \]  
\[ (16) \]

Substituting the expressions for $\mathbf{v}_i(u)$ from (15) and the expressions for $\pi_i, \quad i \geq 1,$ from (5) into (16), after some evident transformations we obtain formula (14).

(ii) The sojourn time distribution in the whole system is a convolution of the sojourn time distribution at Station 1 and the service time distribution at Station 2. Then the Laplace–Stieltjes transform $V'(u)$ is the product of the Laplace–Stieltjes transforms $V_1'(u)$ and $\varphi_2(u)$.
Corollary 5. The mean sojourn time is calculated as follows:

(i) at Station 1

\[ \nu_1 = [\pi_0 I + \pi_1 (I - R)^{-1}] (-D_0)^{-1} e + Lb_1^{(0)}; \]

(ii) in the whole system: \( \nu = \nu_1 + b_1^{(2)}. \)

Here \( b_1^{(0)} \) and \( b_1^{(2)} \) are the mean values of generalized service time and service time at Station 2 respectively.

7. Numerical examples

Experiment 1. In this experiment, we intend to show that the variation in the input process has a great impact on the performance measures of the system. To this end, we consider five input processes with the same value \( a_1 = 0.1 \) of the mean inter-arrival time but different values of the coefficient of variation \( c_{var}. \)

The first process is coded as \( D \) and corresponds to constant inter-arrival times. The second process is coded as \( U \) and corresponds to the uniform distribution in the interval \([0.05, 0.15]\). The third process is coded as \( E \) and corresponds to the Erlangian distribution of order 4 with parameter 40. The fourth process is coded as \( M \) and corresponds to the exponential distribution. The fifth process is coded as \( HM \). It corresponds to hyper-exponential distribution of order 2 defined by the mixing probabilities \((0.05, 0.95)\) and the intensities \((0.62, 49)\). The coefficients of variation of the processes \( D, U, E, M \) and \( HM \) are equal to 0, 0.28, 0.5, 1, 5 respectively. Phase-type service time distribution at Station 1 is the Erlangian distribution of order 2 with parameter 20. Phase-type service time distribution at Station 2 is defined by the irreducible representation \( (\beta, S) \)

\[ \beta = (0.2, 0.8), S = \begin{pmatrix} -9 & 2.7 \\ 4.5 & -18 \end{pmatrix}. \]

Figs. 1 and 2 show the system characteristics as functions of the system load \( \rho \) for five input flows introduced above. The value of \( \rho \) is varied by means of scaling the mean value \( a_1 \) of inter-arrival time in the interval \([0.145, 1.2]\). Note that the coefficients of variation do not change under such a scaling.

Based on Fig. 1, one can conclude that the increase of the system load has negative impact on key performance measures of both stations of the tandem. More important conclusion from Figs. 1 and 2 is that all performance measures are very sensitive with respect to the variation in the input process. Among the dependences presented in Fig. 2, the dependence of probability \( P_{\text{idle-fuss}} \) on \( \rho \) looks the most unpredictable. For small values of the system load \( \rho \) (less than 0.5), this probability increases while for the larger values of \( \rho \) it decreases. Such an effect is explained as follows. When the load \( \rho \) grows, the probability that Station 1 is idle decreases while the probability that Station 2 is busy increases. It appears that, for small values of \( \rho \), the latter probability dominates while for the greater values of \( \rho \) the first probability dominates. Thus, the joint probability that Station 1 is idle and Station 2 is busy has a shape presented in Fig. 2. The behavior of this probability as well as some other characteristics of the system cannot be predicted correctly without computer realization of the results obtained in the present paper.

Experiment 2. In this experiment, we show the impact of the coefficient of variation in the service processes at Station 1 and Station 2 for different values of the system load \( \rho \). Inter-arrival time intervals are assumed to be constant, i.e., the input flow at Station 1 is regular. We consider three service processes at Station 1 given by distributions \( PH_1, PH_2, PH_3 \) defined as follows:

- \( PH_1 \) distribution is exponential distribution with parameter 10;
- \( PH_2 \) distribution is defined by the irreducible representation \( (\beta, S) \) where \( \beta = (0.2, 0.8), S = \begin{pmatrix} -9 & 2.7 \\ 4.5 & -18 \end{pmatrix}; \)
- \( PH_3 \) distribution corresponds to hyper-exponential distribution of order 2. This distribution is defined by the mixing probabilities \((0.05, 0.95)\) and the intensities \((-0.62, -49)\).

For all these three distributions, the mean service time is equal to 0.1. Coefficients of variation are equal to 1, 1.13 and 5, correspondingly.

Phase-type service time distribution \( PH_4 \) at Station 2 is Erlangian distribution of order 2 with parameter 20.

Figs. 3 and 4 show the dependence of the main system characteristics on the value of the system load \( \rho \) under different service time distributions at Station 1.

Now we consider three service processes at Station 2 given by distributions \( PH_1, PH_2, PH_3 \) defined above. Phase-type service time distribution at Station 1 is \( PH_4 \). Figs. 5 and 6 show the dependence of the main system characteristics on the value of the system load \( \rho \) under different service time distributions at Station 2.

Looking at Figs. 3–6 one can observe that the system performance measures are sensitive with respect to the service time variation at both stations. The growth of variation has strong negative effect on the sojourn time. Concerning the effect of the service time variation at Station 1 and Station 2 on the values of probabilities \( P_{\text{succ}} \) and \( P_{\text{block}} \). It is worth to note the following. The essential (from 1-1.13 to 5) growth of the service time variation at Station 1 has negative effect on the probability \( P_{\text{block}} \) while the analogous growth of the service time variation at Station 2 has positive effect on the probability \( P_{\text{succ}} \). At the same time, the effect of service time variation at Station 1 on the probability \( P_{\text{succ}} \) and the effect of service time variation at Station 2 on the probability \( P_{\text{block}} \) depends also on the system load.
8. Conclusion

In this paper, the $\text{GI}/\text{PH}/1 \rightarrow \bullet/\text{PH}/1/0$ tandem queue with blocking is studied. The necessary and sufficient condition for the existence of the stationary distribution is derived and the algorithms for calculating the steady state probabilities are presented. Expressions for the important performance measures of the system are obtained. The Laplace–Stieltjes transforms of the distribution of the sojourn time at both stations as well as at the whole system are derived. Formulas for the mean values of these times are derived. The numerical examples are presented. They illustrate the effect of variation in the input flow on the system performance measures. The results of this paper can be applied to areas such as capacity planning, performance evaluations, and optimization of real-life tandem queues and two-node networks.

Acknowledgments

This research was supported by 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).

References