European Journal of Operational Research 236 (2014) 177–189

Contents lists available at ScienceDirect

European Journal of Operational Research

journal homepage: www.elsevier.com/locate/ejor

Stochastics and Statistics

A tollbooth tandem queue with heterogeneous servers

Qi-Ming He [a,*], Xiuli Chao [b]

[a] Department of Management Sciences, University of Waterloo, Waterloo, Ontario N2L 3G1, Canada
[b] Department of Industrial and Operations Engineering, University of Michigan, Ann Arbor, MI 48109-2117, USA

ARTICLE INFO

Article history:
Received 4 May 2013
Accepted 23 January 2014
Available online 1 February 2014

Keywords:
Tandem queue
Matrix-analytic methods
Traffic management

ABSTRACT

We study a tandem queueing system with K servers and no waiting space in between. A customer needs service from one server but can leave the system only if all down-stream servers are unoccupied. Such a system is often observed in toll collection during rush hours in transportation networks, and we call it a tollbooth tandem queue. We apply matrix-analytic methods to study this queueing system, and obtain explicit results for various performance measures. Using these results, we can efficiently compute the mean and variance of the queue lengths, waiting time, sojourn time, and departure delays. Numerical examples are presented to gain insights into the performance and design of the tollbooth tandem queue. In particular, it reveals that the intuitive result of arranging servers in decreasing order of service speed (i.e., arrange faster servers at downstream stations) is not always optimal for minimizing the mean queue length or mean waiting time.

1. Introduction

Tollbooths are used to collect tolls for road usage on toll roads. Due to service delays, queues can build up in front of tollbooths. It is commonly observed in some metropolitan cities, e.g., Seoul and Beijing, that tandem tollbooths are set up during rush hours to reduce delays. In this paper, we present a queueing model to study the flows of the traffic going through tollbooths, and develop an algorithmic study for its performance analysis. We note that when all tollbooths serve traffic in parallel, instead of in series, the tollbooth system can be modeled as a queueing network with customer jockeying and has been analyzed extensively in the literature (e.g., Disney & Mitchell, 1971; Zhao & Grassmann, 1995).

When tollbooths are in series, each car/truck has to go through every tollbooth but receive service from one and only one of them. Such a queueing system has two distinct features: (i) cars/trucks that completed service at one station may not be able to leave the system immediately; and (ii) servers may be idle even if there are waiting customers. When a car/truck is in service in a tollbooth, it blocks all cars/trucks at the upstream tollbooths. We shall refer to this type of queueing system as tollbooth tandem queue.

The tollbooth tandem queue is similar to the traditional tandem queue with no waiting space between servers or with blocking (e.g., Melamed, 1986; Manitz, 2008; Papadopoulos & O'Kelly, 1993; Van Houdt & Alfa, 2005), but different in that each customer receives service from one and only one server. Note that when there are always customers waiting in front of the tollbooth, the service process of the tollbooth queue is similar to that of a batch service queue (Chaudhry & Templeton, 1983), with service time of a batch being the maximum of processing times of the customers in the batch at their respective tollbooths. To the author's knowledge, the added features in the service process distinguish the tollbooth tandem queue from all the existing models in the queueing literature.

In the transportation research literature there have been only numerical and simulation studies of tollbooth tandem queues, see for example, Hall and Daganzo (1983), Rubenstein (1983), Hong, Kim, Kho, Kim, and Yang (2009), Gu and Li (2011), Gu, Cassidy, and Li (2012). These studies suggest that tandem tollbooths can improve queueing performance significantly. However, there has been no systematic queueing study on the tollbooth tandem queue problem.

In this paper, we apply matrix-analytic methods to investigate the tollbooth tandem queue. We refer to Neuts (1981) and Latouche and Ramaswami (1999) for more details on matrix-analytic methods. Closed form solutions are obtained for a number of performance measures. For example, *PH*-representations are constructed for the waiting time and sojourn time of an arbitrary customer. We also obtain special performance measures such as the utilizations of servers, percentages of customers served by servers, and departure delays.

Using the results derived in this paper, we analyze several examples to gain insight into the tollbooth tandem queue. It is intuitive that arranging servers in the descending order of their service speed (i.e., faster servers are arranged at the most downstream stations)

* Corresponding author. Tel.: +1 519 888 4567.
E-mail addresses: q7he@uwaterloo.ca (Q.-M. He), xchao@umich.edu (X. Chao).

http://dx.doi.org/10.1016/j.ejor.2014.01.049
0377-2217/

should reduce the chance of blocking and customer delays. However, we found that that is not always true. Many of the performance measures, including the mean waiting time, mean sojourn time, mean number of customers waiting, and mean total number of customers in system, may not be minimized by this arrangement of servers (though in many cases it does).

In addition to its potential application in toll road management, the tollbooth tandem queue may also find applications in modeling traffic in gas stations (e.g., Brewer & Henwood, 1980; Daskin, Shladover, & Sobel, 1976; Teimoury, Yazdi, Haddadi, & Fathi, 2011). For such cases, the tollbooth tandem queue can approximate the real system well if the order of departures of cars/trucks is, more or less, consistent with the order of arrivals.

The remainder of the paper is organized as follows. In Section 2, the tollbooth tandem queue of interest is formally defined. A Markov chain formulation for the queue is presented in Section 3. A matrix-geometric solution is given in Section 4. A number of performance measures are obtained in Section 5. Distributions of the waiting time, sojourn time, and departure delay are obtained in Section 6. Section 7 collects some special properties including the phase-type representations of the waiting time and sojourn time. The special case of identical serial servers is analyzed in Section 8 and a numerical analysis is presented in Section 9. Section 10 concludes the paper.

2. Definition of the tollbooth tandem queue

We consider a tollbooth tandem queueing model with K servers and no waiting space in between (see Fig. 1 for a case with $K = 3$). Customers join a single queue in front of server K waiting for service by one (and only one) of the K servers. Customers enter service on a first-come-first-in basis. If server k is available, then a waiting customer enters server k if all *upstream servers* $\{k+1, \ldots, K\}$ are empty and server $k-1$ is occupied. If a customer at server k completes its service, it can leave the system only if all *downstream servers* $\{1, \ldots, k-1\}$ are empty; otherwise, the customer occupies server k until all servers $\{1, \ldots, k-1\}$ become empty. Accordingly, if a customer occupies server k, in service or waiting to depart, all customers at servers $\{k+1, \ldots, K\}$, if there is any, cannot depart; thus there may be batch departures from the system.

To define the tollbooth tandem queue, we specify the customer arrival process and the distributions of the service times for individual servers.

Markovian arrival process (***MAP***). Customers arrive according to Markovian arrival process (D_0, D_1), where D_0 and D_1 are square matrices of order m_a, D_0 has negative diagonal elements and nonnegative off-diagonal elements, and D_1 is a nonnegative matrix. Let $D = D_0 + D_1$, which is an infinitesimal generator. We assume that D is irreducible. Then D defines an irreducible continuous time Markov chain (CTMC). Let $I_a(t)$ be the state (phase) of the CTMC associated with D, at time t. Then $\{I_a(t),\ t \geqslant 0\}$ is an irreducible CTMC, called the underlying Markov chain. Let $\boldsymbol{\theta}_a$ be the stationary distribution of $\{I_a(t),\ t \geqslant 0\}$. Then $\boldsymbol{\theta}_a$ is the unique solution to linear system $\boldsymbol{\theta}_a D = 0$ and $\boldsymbol{\theta}_a \mathbf{e} = 1$, where $\mathbf{e}$ is the column vector with all elements being one. The (average) arrival rate can be obtained as $\lambda = \boldsymbol{\theta}_a D_1 \mathbf{e}$. For more about *MAP*s, readers are referred to Neuts (1979) and Lucantoni (1991).

Phase-type (***PH***) **service times.** The service time V_k of server k has a *PH*-distribution with *PH*-representation $(\boldsymbol{\alpha}_k, T_k)$ of order m_k, for $1 \leqslant k \leqslant K$. We assume $\boldsymbol{\alpha}_k \mathbf{e} = 1$, i.e., the service time is positive with probability one. Let $\{I_{s,k}(t),\ t \geqslant 0\}$ be the state of the underlying CTMC associated with $(\boldsymbol{\alpha}_k, T_k)$. Then $\{I_{s,k}(t),\ t \geqslant 0\}$ has $m_k + 1$ states $\{0, 1, 2, \ldots, m_k\}$ and infinitesimal generator

$$Q_k = \begin{pmatrix} 0 & 0 \\ \mathbf{T}_k^0 & T_k \end{pmatrix}. \tag{1}$$

We call the state 0 the *absorption state*. Then V_k is the time until absorption of the underlying Markov chain into state 0, if the distribution of $I_{s,k}(0)$ is $(0, \boldsymbol{\alpha}_k)$. The (average) service rate is given by $\mu_k = 1/E[V_k] = -1/\left(\boldsymbol{\alpha}_k T_k^{-1} \mathbf{e}\right)$. Recall that $\mathbf{T}_k^0 = -T_k \mathbf{e}$. Then $T + \mathbf{T}_k^0 \boldsymbol{\alpha}_k$ is an infinitesimal generator. We assume that $T + \mathbf{T}_k^0 \boldsymbol{\alpha}_k$ is irreducible. Then the *PH*-representation $(\boldsymbol{\alpha}_k, T_k)$ is called *PH*-irreducible. Denote by $f_k^*(s)$ the Laplace–Stieltjes transform (LST) of the service time V_k: $f_k^*(s) = \boldsymbol{\alpha}_k (sI - T_k)^{-1} \mathbf{T}_k^0$, for $s \geqslant 0$, where I is the identity matrix whose order is determined by the context. For more about *PH*-distributions, readers are referred to Neuts (1975, 1981).

We are interested in investigating the following quantities: The number of waiting customers in system q_w, the number of customers occupying a server q_s, the total number of customers in the system q_{tot}, customer waiting time W_q, customer sojourn time W_s, and customer departure delay W_d (i.e., the amount of time a customer spends waiting for departure after finishing service).

3. Maximum of independent random variables

In the tollbooth tandem queue, if K customers enter the K servers for service simultaneously, then the time for all K customers to leave the system (i.e., the time for the customer at server K to leave the system) is the maximum of all K service times. This fact plays a key role in the construction of a Markov chain for the tollbooth tandem queue and in the analysis of performance measures. It is well-known that the maximum of independent *PH*-random variables is again a *PH*-random variable and its *PH*-representation can be constructed (Neuts, 1981). To find the solutions for the tollbooth tandem queue, we need the exact representation of this random variable. Thus, we first present two methods for constructing a *PH*-representation of the maximum of K independent *PH*-random variables. Both representations will be used in our subsequent analysis.

Define $V_{\max,k} = \max\{V_1, \ldots, V_k\}$, for $1 \leqslant k \leqslant K$. Next, we construct a *PH*-representation for $V_{\max,k}$. Assume that the (independent) underlying Markov chains $\{I_{s,k}(t), t \geqslant 0\}$, for $1 \leqslant k \leqslant K$, are all initialized at time zero. Then $V_{\max,k}$ can be interpreted as the first time that all underlying Markov chains $\{I_{s,j}(t),\ t \geqslant 0\}$, for $1 \leqslant j \leqslant k$, are absorbed. Thus, we consider CTMC $\{(I_{s,1}(t), I_{s,2}(t), \ldots, I_{s,k}(t)),\ t \geqslant 0\}$. The number of states, the initial probability distribution, and the infinitesimal generator of the CTMC $\{(I_{s,1}(t), I_{s,2}(t), \ldots, I_{s,k}(t)), t \geqslant 0\}$ are given by, for $1 \leqslant k \leqslant K$,

$$\begin{aligned} &m_{\max,\langle k\rangle} = \prod_{j=1}^{k} (m_j + 1) - 1; \\ &(0, \boldsymbol{\alpha}_{\max,\langle k\rangle}) \equiv (0, \boldsymbol{\alpha}_1) \otimes (0, \boldsymbol{\alpha}_2) \otimes \cdots \otimes (0, \boldsymbol{\alpha}_k); \\ &Q_{\langle k\rangle} = \begin{pmatrix} 0 & 0 \\ \mathbf{T}^0_{\max,\langle k\rangle} & T_{\max,\langle k\rangle} \end{pmatrix} \equiv Q_1 \oplus Q_2 \oplus \cdots \oplus Q_k, \end{aligned} \tag{2}$$

where "$\otimes$" is for Kronecker product and "$\oplus$" is for Kronecker sum, i.e., $Q_1 \oplus Q_2 = Q_1 \otimes I + I \otimes Q_2$. Denote by $\Psi_{\max,\langle k\rangle}$ the set of states obtained by removing state $(0, \ldots, 0)$ from the state space $\{(i_1, i_2, \ldots, i_k): 0 \leqslant i_j \leqslant m_j, 1 \leqslant j \leqslant k\}$ of $\{(I_{s,1}(t), I_{s,2}(t), \ldots, I_{s,k}(t)), t \geqslant 0\}$. Then $\Psi_{\max,\langle k\rangle}$

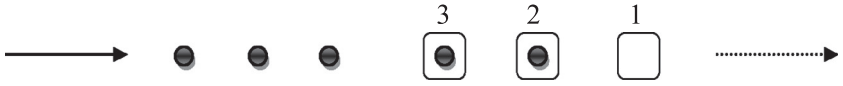

Fig. 1. A tollbooth tandem queue with $K = 3$.

has $m_{\max,\langle k\rangle}$ states. For any state in $\Psi_{\max,\langle k\rangle}$, at least one of the underlying Markov chains $\{I_{s,j}(t), t \geqslant 0\}, j = 1, 2, \ldots, k$, is not in its absorption state. Thus $V_{\max,k}$ is the absorption time into the state $(0, \ldots, 0)$ of the CTMC $\{(I_{s,1}(t), I_{s,2}(t), \ldots, I_{s,k}(t)), t \geqslant 0\}$. By Eq. (2), it is readily seen that $(\boldsymbol{\alpha}_{\max,\langle k\rangle}, T_{\max,\langle k\rangle})$ of order $m_{\max,\langle k\rangle}$ is a *PH*-representation for $V_{\max,k}$. Denote by $f^*_{\max,k}(s)$ the LST of $V_{\max,k}$, for $1 \leqslant k \leqslant K$.

Let $\{\lambda_{k,i}, 1 \leqslant i \leqslant m_k\}$ be the eigenvalues of T_k, counting multiplicities, for $1 \leqslant k \leqslant K$. We assume that $\lambda_{k,1}$ is the eigenvalue of T_k with the largest real part, which is in fact real.

Proposition 3.1. *The eigenvalues of $T_{max,\langle K\rangle}$ are given by, counting multiplicities,*

$$\{\lambda_{k_1,i_1} + \cdots + \lambda_{k_j,i_j}, 1 \leqslant k_1 < \cdots < k_j \leqslant K, 1 \leqslant i_1 \leqslant m_{k_1}, \ldots, 1 \leqslant i_j \leqslant m_{k_j}, 1 \leqslant j \leqslant K\}. \tag{3}$$

In particular, the eigenvalue of $T_{\max,\langle K\rangle}$ with the largest real part is $\max\{\lambda_{k,1}, k = 1, 2, \ldots, K\}$.

Proof. It is clear that the eigenvalues of Q_k are $\{0\} \cup \{\lambda_{k,i}, i = 1, 2, \ldots, m_k\}$. Similarly, all eigenvalues of $Q_{\langle K\rangle}$ are the eigenvalues of $T_{\max,\langle K\rangle}$, except for eigenvalue zero. By Horn and Johnson (1991) (Problem 19 on Page 251), Eq. (3) gives all the eigenvalues of $Q_{\langle K\rangle}$, except for eigenvalue zero. Consequently, Eq. (3) gives all the eigenvalues of $T_{\max,\langle K\rangle}$. This completes the proof of Proposition 3.1. □

The order $m_{\max,\langle K\rangle}$ increases rapidly in K. For special cases, an equivalent *PH*-representation with fewer states can be obtained for $V_{\max,K}$. Considering the (independent) underlying Markov chains $\{\{I_{s,k}(t), t \geqslant 0\}, k = 1, 2, \ldots, K\}$, $V_{\max,K}$ can be interpreted as (i) the time until one of the K underlying Markov chains is absorbed, (ii) plus the time until one of the $K-1$ remaining underlying Markov chains is absorbed, ..., and (iii) plus the time until the last remaining underlying Markov chain is absorbed. Based on this interpretation, an alternative *PH*-representation can be obtained for $V_{\max,K}$. However, the alternative *PH*-representation has the same order as that of $(\boldsymbol{\alpha}_{\max,\langle K\rangle}, T_{\max,\langle K\rangle})$. On the other hand, if all random variables $\{V_1, \ldots, V_K\}$ are also identically distributed with the same *PH*-representation $(\boldsymbol{\alpha}, T)$ of order m, many states of the underlying Markov chain $Q_{\langle K\rangle}$ are redundant. Thus, by combining the redundant states, a smaller *PH*-representation can be obtained for $V_{\max,K}$. Define, $T_{[1]} = T$, and, for $2 \leqslant k \leqslant K$,

$$\begin{aligned} T_{[k]} &= T_{[k-1]} \otimes I_m + I_{m^{k-1}} \otimes T; \\ T^0_{[k]} &= \sum_{j=1}^{k} I_{m^{j-1}} \otimes \mathbf{T}^0 \otimes I_{m^{k-j}}, \end{aligned} \tag{4}$$

where I_n is the identity matrix of order n and $\mathbf{T}^0 = -T\mathbf{e}$. Note that $T_{[k]}$ can be interpreted as the subgenerator for $\{(I_{s,1}(t), I_{s,2}(t), \ldots, I_{s,k}(t)), t \geqslant 0\}$ such that none of the subprocesses is absorbed; and $T^0_{[k]}$ is for the absorption of one of the subprocesses. We define $\Omega_{[k]} = \{(i_1, i_2, \ldots, i_k): 1 \leqslant i_j \leqslant m, 1 \leqslant j \leqslant k\}$, $\boldsymbol{\alpha}_{[1]} = \boldsymbol{\alpha}$, $\boldsymbol{\alpha}_{[k]} = \boldsymbol{\alpha}_{[k-1]} \otimes \boldsymbol{\alpha}$, $\boldsymbol{\alpha}_{\max,[K]} = (0, \ldots, 0, \boldsymbol{\alpha}_{[K]})$;

$$T_{\max,[K]} = \begin{matrix} \Omega_{[1]} \\ \Omega_{[2]} \\ \vdots \\ \Omega_{[K-1]} \\ \Omega_{[K]} \end{matrix} \begin{pmatrix} T_{[1]} & & & & \\ T^0_{[2]} & T_{[2]} & & & \\ & \ddots & \ddots & & \\ & & T^0_{[K-1]} & T_{[K-1]} & \\ & & & T^0_{[K]} & T_{[K]} \end{pmatrix} \quad \text{and} \tag{5}$$

$$\mathbf{T}^0_{\max,[K]} = -T_{\max,[K]}\mathbf{e} = \begin{matrix} \Omega_{[1]} \\ \Omega_{[2]} \\ \vdots \\ \Omega_{[K]} \end{matrix} \begin{pmatrix} \mathbf{T}^0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}.$$

The maximum random variable $V_{\max,K}$ has a *PH*-distribution with *PH*-representation $(\boldsymbol{\alpha}_{\max,[K]}, T_{\max,[K]})$ given in Eq. (5) of order $m_{\max,[K]} = m + m^2 + \cdots + m^K$, which can be significantly smaller than $m_{\max,\langle K\rangle} = (m+1)^K - 1$.

4. Matrix-geometric solutions for the tollbooth tandem queue

In this section, we show that the tollbooth tandem queue can be formulated and analyzed using a *GI/M/1* type Markov chain (e.g., Latouche & Ramaswami, 1999; Neuts, 1981). To construct the *GI/M/1* type Markov chain, we first observe that, at any time in the tollbooth tandem queue, all occupied servers (either a customer in service or waiting to depart) are next to each other. Thus, we distinguish the case with a waiting queue and the case with no waiting queue. If there is a waiting queue, then the Kth server must be occupied. If there is no waiting queue, the K servers can be divided into three groups $\{1, 2, \ldots, k_1\}$, $\{k_1 + 1, \ldots, k_2\}$, and $\{k_2 + 1, \ldots, K\}$. Servers $\{1, 2, \ldots, k_1\}$ are empty; servers $\{k_1 + 1, \ldots, k_2\}$ are all occupied and server $k_1 + 1$ is still in service; and servers $\{k_2 + 1, \ldots, K\}$ are empty and there is no waiting customer in the system if this set of servers is not empty. To present the Markov chain formulation explicitly, we define system variables:

- $q_w(t)$: The number of waiting customers in the system at time t;
- $I_a(t)$: The state of the Markovian arrival process at time t;
- $I_k(t)$: The state of the underlying CTMC of the service at server k, which is $I_{s,k}(t)$, if server k is working; Otherwise, $I_k(t) = 0$, at time t; for $1 \leqslant k \leqslant K$.
- $I(t)$: The maximum index of the occupied servers at time t (i.e., the index of the first occupied server seen by an arrival). If $q_w(t) = 0$, $I(t)$ takes values $\{0,1, \ldots, K\}$, where 0 means that the system is empty. If $q_w(t) > 0$, we must have $I(t) = K$.

Then it is easy to see that $\{(q_w(t), I(t), I_1(t), \ldots, I_K(t), I_a(t)), t \geqslant 0\}$ is a CTMC. We first analyze this CTMC and then find system performance measures. Our analysis consists of three steps: (i) Construct the infinitesimal generator of the CTMC and find its matrix-geometric solution; (ii) Based on the matrix-geometric solution, find the distributions of queue lengths at an arbitrary time, arrival epochs, and service completion epochs (Section 5); and (iii) Find the distributions of waiting time, sojourn time, and departure delays (Section 6).

If $q_w(t) > 0$, server K must be occupied, i.e., the server is either in service or with a finished customer whose departure is delayed because a downstream service has not been completed. When $q_w(t) = 0$, server K can be either idle or occupied. For convenience, we distinguish the two cases for $q_w(t) = 0$ by introducing a new state -1 for $q_w(t)$. If server K is unoccupied, then we set $q_w(t) = -1$. We shall continue to call $q_w(t)$ the waiting queue length with the understanding that, if $q_w(t) = -1$, then the waiting queue length is zero.

We shall call $q_w(t)$ the level variable and $(I(t), I_1(t), \ldots, I_K(t), I_a(t))$ the phase variable. The variable $q_w(t)$ can increase by at most one if a customer arrives, and can decrease by at most K if one service is completed and all servers becomes available to waiting customers. It is then readily seen that $\{(q_w(t), I(t), I_1(t), \ldots, I_K(t), I_a(t)), t \geqslant 0\}$ is a *GI/M/1* type Markov chain.

For level -1 of the *GI/M/1* type Markov chain, $I(t)$ takes values $\{0, 1, \ldots, K-1\}$. If $I(t) = 0$, there is no customer in the system and we must have $(I_1(t), \ldots, I_K(t)) = (0, \ldots, 0)$. If $I(t) = k \geqslant 1$, then $(I_1(t), \ldots, I_k(t))$ takes values in the set $\Psi_{\max,\langle k\rangle}$, and $(I_{k+1}(t), \ldots, I_K(t)) = (0, \ldots, 0)$. A state in level -1 with $I(t) = k$ indicates that at least one of the servers $\{1, 2, \ldots, k\}$ is working, the kth server is occupied, and servers $\{k+1, \ldots, K\}$ are available to customers. The number of states of $(I_1(t), \ldots, I_K(t))$ in level -1 is $m_{-1,\langle K-1\rangle} = 1 + m_{\max,\langle 1\rangle} + \cdots + m_{\max,\langle K-1\rangle}$.

For level $q_w(t) \geqslant 0$, we must have $I(t) = K$, and $(I_1(t), \ldots, I_K(t))$ takes values in the set $\Psi_{\max,\langle K\rangle}$, which has $m_{\max,\langle K\rangle}$ states. The structure of level $q_w(t) \geqslant 0$ is simpler than that of level -1, since the Kth server must be occupied for such levels.

The infinitesimal generator of $\{(q_w(t), I(t), I_1(t), \ldots, I_K(t), I_a(t)), t \geqslant 0\}$ is given by

$$Q = \begin{array}{c} -1 \\ 0 \\ 1 \\ \vdots \\ K-1 \\ K \\ K+1 \\ \vdots \end{array} \begin{pmatrix} A_{-1,-1} & A_{-1,0} & & & & & & \\ A_{0,-1} & A_1 & A_0 & & & & & \\ A_{1,-1} & 0 & A_1 & A_0 & & & & \\ \vdots & \vdots & \ddots & \ddots & \ddots & & & \\ A_{K-1,-1} & 0 & \cdots & 0 & A_1 & A_0 & & \\ 0 & A_{K+1} & 0 & \cdots & 0 & A_1 & A_0 & \\ \vdots & & A_{K+1} & 0 & \cdots & 0 & A_1 & A_0 \\ \vdots & & & \ddots & \ddots & \ddots & \ddots & \ddots \end{pmatrix}, \tag{6}$$

where

$$A_{-1,-1} = \begin{pmatrix} 0 & \boldsymbol{\alpha}_1 & & & \\ & 0 & \left(0_{m_{\max,\langle 1\rangle}\times m_2} \quad I_{m_{\max,\langle 1\rangle}} \otimes (0, \boldsymbol{\alpha}_2)\right) & & \\ & & \ddots & \ddots & \\ & & & 0 & \left(0_{m_{\max,\langle K-2\rangle}\times m_{K-1}} \quad I_{m_{\max,\langle K-2\rangle}} \otimes (0, \boldsymbol{\alpha}_{K-1})\right) \\ & & & & 0_{m_{\max,\langle K-1\rangle}\times m_{\max,\langle K-1\rangle}} \end{pmatrix} \otimes D_1 \tag{7}$$

$$+ I_{m_{-1,\langle K\rangle}} \otimes D_0 + \begin{pmatrix} 0 & & & \\ \mathbf{T}^0_{\max,\langle 1\rangle} & T_{\max,\langle 1\rangle} & & \\ \vdots & & \ddots & \\ \mathbf{T}^0_{\max,\langle K-2\rangle} & & & T_{\max,\langle K-2\rangle} \\ \mathbf{T}^0_{\max,\langle K-1\rangle} & & & \end{pmatrix} \otimes I_{m_a};$$

$$A_{-1,0} = \begin{pmatrix} 0_{m_{-1,\langle K-2\rangle}\times m_K} & 0_{m_{-1,\langle K-2\rangle}\times(m_{\max,\langle K\rangle}-m_K)} \\ 0_{m_{\max,\langle K-1\rangle}\times m_K} & I_{m_{\max,\langle K-1\rangle}} \otimes (0, \boldsymbol{\alpha}_K) \end{pmatrix} \otimes D_1;$$

$$A_{0,-1} = \left(\mathbf{T}^0_{\max,\langle K\rangle} \quad 0_{m_{\max,\langle K\rangle}\times(m_{-1,\langle K\rangle}-1)}\right) \otimes I_{m_a};$$

$$A_{j,-1} = \left(0_{m_{\max,\langle K\rangle}\times m_{-1,\langle j-1\rangle}} \quad \mathbf{T}^0_{\max,\langle K\rangle}\boldsymbol{\alpha}_{\max,\langle j\rangle} \quad 0_{m_{\max,\langle K\rangle}\times(m_{-1,\langle K\rangle}-m_{-1,\langle j\rangle})}\right) \otimes I_{m_a}, \quad 1 \leqslant j \leqslant K-1;$$

$$A_0 = I_{m_{\max,\langle K\rangle}} \otimes D_1;$$
$$A_1 = I_{m_{\max,\langle K\rangle}} \otimes D_0 + T_{\max,\langle K\rangle} \otimes I_{m_a};$$
$$A_{K+1} = \left(\mathbf{T}^0_{\max,\langle K\rangle}\boldsymbol{\alpha}_{\max,\langle K\rangle}\right) \otimes I_{m_a}.$$

Note that (i) $m_{-1,\langle 0\rangle} = 1$; and (ii) in the above matrices, "0" also represents a block of zeros, whose order is determined by the context. The order of "0" is provided whenever it is necessary. First, we find a condition for the ergodicity of the CTMC.

Theorem 4.1. *The CTMC $\{(q_w(t), I(t), I_1(t), \ldots, I_K(t), I_a(t)), t \geqslant 0\}$ is irreducible. The CTMC is positive recurrent if and only if $\lambda E[V_{max,K}] = -\lambda\boldsymbol{\alpha}_{max,\langle K\rangle}(T_{max,\langle K\rangle})^{-1}\mathbf{e} < K$.*

Proof. A proof can be given by applying Neuts condition to the ergodicity of $GI/M/1$ type Markov chains (Neuts, 1981). Details are omitted. □

Next, we present the matrix-geometric solution for the stationary distribution $\boldsymbol{\pi}$ of the CTMC, which is divided into $\boldsymbol{\pi} = (\boldsymbol{\pi}_{-1}, \boldsymbol{\pi}_0, \boldsymbol{\pi}_1, \ldots)$ according to the level variable $q_w(t)$. It is well-known that, if the CTMC is irreducible and positive recurrent, then $\boldsymbol{\pi}$ is the unique solution to linear system $\boldsymbol{\pi}Q = 0$ and $\boldsymbol{\pi}\mathbf{e} = 1$. Assume that $-\lambda\boldsymbol{\alpha}_{\max,\langle K\rangle}(T_{\max,\langle K\rangle})^{-1}\mathbf{e} < K$. By Neuts (1981), the stationary distribution $\boldsymbol{\pi}$ is given by $\{\boldsymbol{\pi}_{-1}, \boldsymbol{\pi}_0 R^n, n = 0, 1, 2, \ldots\}$, where $\{\boldsymbol{\pi}_{-1}, \boldsymbol{\pi}_0\}$ is the unique solution to

$$\begin{aligned} &\boldsymbol{\pi}_{-1}A_{-1,-1} + \boldsymbol{\pi}_0\sum_{k=0}^{K-1} R^k A_{k,-1} = 0; \\ &\boldsymbol{\pi}_{-1}A_{-1,0} + \boldsymbol{\pi}_0 A_1 + \boldsymbol{\pi}_0 R^K A_{K+1} = 0; \\ &\boldsymbol{\pi}_{-1}\mathbf{e} + \boldsymbol{\pi}_0(I-R)^{-1}\mathbf{e} = 1, \end{aligned} \tag{8}$$

and the matrix R is the minimal nonnegative solution to

$$A_0 + RA_1 + R^{K+1}A_{K+1} = 0. \tag{9}$$

The computation of R can be done by using existing algorithms such as the monotonic iterative algorithm (Neuts, 1981) and more efficient algorithms (Bini, Meini, Steffe, & Van Houdt, 2006). To check the accuracy of computation, the following property can be used:

$$\boldsymbol{\theta}_a = \boldsymbol{\pi}_{-1}(\mathbf{e} \otimes I_{m_a}) + \boldsymbol{\pi}_0(I-R)^{-1}(\mathbf{e} \otimes I_{m_a}). \tag{10}$$

In the next two sections, we use the stationary distribution $\{\boldsymbol{\pi}_{-1}, \boldsymbol{\pi}_0 R^n, n = 0, 1, 2, \ldots\}$ to find the distributions of queue lengths, waiting time, sojourn time, and departure delays.

5. Queue length analysis of the tollbooth tandem queue

We consider two types of queue lengths in this section: (i) the number of customers waiting in queue and (ii) the number of customers at servers (in service or waiting to depart). We consider the queue lengths (i) at an arbitrary time, (ii) seen by an arbitrary arrival, and (iii) right after a service completion. We also consider the number of waiting batches seen by an arbitrary arrival.

Proposition 5.1 (*Waiting queue length at an arbitrary time*). *In steady-state, let q_w be the number of customers waiting in queue.*

(i) *The distribution of q_w is $P\{q_w = 0\} = \boldsymbol{\pi}_{-1}\mathbf{e} + \boldsymbol{\pi}_0\mathbf{e}$, $P\{q_w = n\} = \boldsymbol{\pi}_0 R^n\mathbf{e}$, $n \geqslant 1$.*
(ii) *The probability generating function of q_w is $E[z^{q_w}] = \boldsymbol{\pi}_{-1}\mathbf{e} + \boldsymbol{\pi}_0(I - zR)^{-1}\mathbf{e}$, for $0 \leqslant z \leqslant 1$.*
(iii) *The mean and variance of q_w are*

$$\begin{aligned} &E[q_w] = \boldsymbol{\pi}_0 R(I-R)^{-2}\mathbf{e}; \\ &Var[q_w] = 2\boldsymbol{\pi}_0 R^2(I-R)^{-3}\mathbf{e} - (\boldsymbol{\pi}_0 R(I-R)^{-2}\mathbf{e})^2. \end{aligned} \tag{11}$$

To find the number of customers in service or waiting to depart, we need to decompose the vector $\boldsymbol{\pi}_{-1}$ according to the variable $I(t)$: $\boldsymbol{\pi}_{-1} = (\boldsymbol{\pi}_{-1,\langle 0\rangle}, \boldsymbol{\pi}_{-1,\langle 1\rangle}, \ldots, \boldsymbol{\pi}_{-1,\langle K-1\rangle})$. We also need to find the number of customers in service or waiting to depart for states in $\Psi_{\max,\langle k\rangle}$, for $1 \leqslant k \leqslant K$. Define vectors $\{\mathbf{u}_{\max,\langle k,j\rangle}, j = 0,1,\ldots,k\}$ as follows: $\mathbf{u}_{\max,\langle k,\ 0\rangle} = 0$ (a column vector with all elements being zero), and, for $1 \leqslant j \leqslant k \leqslant K$,

$$\begin{pmatrix} 0 \\ \mathbf{u}_{\max,\langle k,j\rangle} \end{pmatrix} \equiv \left(\begin{pmatrix} 1 & \\ & 0_{m_1\times m_1} \end{pmatrix} \otimes \cdots \otimes \begin{pmatrix} 1 & \\ & 0_{m_{k-j}\times m_{k-j}} \end{pmatrix} \otimes \begin{pmatrix} 0 & \\ & I_{m_{k-j+1}} \end{pmatrix} \otimes I_{m_{k-j+2}+1} \otimes \cdots \otimes I_{m_k+1} \right) \mathbf{e}. \tag{12}$$

Elements in the vector $\mathbf{u}_{\max,\langle k,j\rangle}$ is either zero or one. That an element in $\mathbf{u}_{\max,\langle k,j\rangle}$ is one indicates that the corresponding state in $\Psi_{\max,\langle k\rangle}$ has j customers either in service or waiting to depart. By the interpretation, we must have $\mathbf{u}_{\max,\langle k,0\rangle} + \mathbf{u}_{\max,\langle k,1\rangle} + \cdots + \mathbf{u}_{\max,\langle k,k\rangle} = \mathbf{e}$.

For convenience we introduce notation $\boldsymbol{\pi}_{\geqslant 0} = \boldsymbol{\pi}_0 + \boldsymbol{\pi}_1 + \boldsymbol{\pi}_2 + \cdots = \boldsymbol{\pi}_0(I-R)^{-1}$. Also define $p_{-1,\langle k,j\rangle} = \boldsymbol{\pi}_{-1,\langle k\rangle}(\mathbf{u}_{\max,\langle k,j\rangle} \otimes \mathbf{e})$, for $1 \leqslant j \leqslant k \leqslant K$, where the order of $\mathbf{e}$ is m_a that represents the number of states of the Markovian arrival process.

Proposition 5.2 (*Server queue length at an arbitrary time*). *In steady-state, let q_s be the number of customers occupying a server.*

(i) *The distribution of q_s is given by*

$$P\{q_s = k\} = \sum_{j=k}^{K-1} p_{-1,\langle j,k\rangle} + \boldsymbol{\pi}_0(I-R)^{-1}(\mathbf{u}_{\max,\langle K,k\rangle} \otimes \mathbf{e}), \quad k = 0,1,2,\ldots,K. \tag{13}$$

(ii) *The probability generating function of q_s is*

$$E[z^{q_s}] = \sum_{j=1}^{K-1}\sum_{k=1}^{j} z^k p_{-1,\langle j,k\rangle} + \boldsymbol{\pi}_0(I-R)^{-1}\left(\sum_{k=1}^{K} z^k \mathbf{u}_{\max,\langle K,k\rangle}\right) \otimes \mathbf{e}. \tag{14}$$

We can similarly obtain the mean and variance of q_s. Details are omitted.

Customers in the queueing system are either waiting in the queue or occupying a server (in service or waiting to depart). Thus, we can use Propositions 5.1 and 5.2 to find the total number of customers in the system.

Proposition 5.3 (*Total queue length at an arbitrary time*). *In steady-state, let $q_{tot} = q_w + q_s$ be the total number of customers in the queueing system.*

(i) *The distribution of q_{tot} is given by*

$$P\{q_{tot} = n\} = \begin{cases} \sum_{j=n}^{K-1} p_{-1,\langle j,n\rangle} + \sum_{j=1}^{n} \boldsymbol{\pi}_0 R^{j-1}(\mathbf{u}_{\max,\langle K,n-j+1\rangle} \otimes \mathbf{e}), & \text{for } 0 \leqslant n \leqslant K-1; \\ \boldsymbol{\pi}_0 R^{n-K}\sum_{j=1}^{K} R^{K-j}(\mathbf{u}_{\max,\langle K,j\rangle} \otimes \mathbf{e}), & \text{for } n \geqslant K. \end{cases} \tag{15}$$

(ii) *The probability generating function of q_{tot} is*

$$E[z^{q_{tot}}] = \sum_{j=0}^{K-1}\sum_{k=1}^{j} z^k p_{-1,\langle j,k\rangle} + \boldsymbol{\pi}_0 \sum_{j=1}^{K-1} R^{j-1}\left(\sum_{n=j}^{K-1} z^n \mathbf{u}_{\max,\langle K,n\rangle}\right) \otimes \mathbf{e} + \boldsymbol{\pi}_0 z^K (I - zR)^{-1}\left(\sum_{j=1}^{K} R^{K-j}\mathbf{u}_{\max,\langle K,j\rangle} \otimes \mathbf{e}\right) \tag{16}$$

We can similarly obtain the mean and variance of q_{tot}. Details are omitted.

Next, we find the queue length seen by an arbitrary arrival.

Proposition 5.4 (*The queue length seen by an arbitrary arrival*). *In steady state, the distribution of $\{(q_w(t), I(t), I_1(t), \ldots, I_K(t)), t \geqslant 0\}$ seen by an arbitrary arrival is given by $\{\boldsymbol{\pi}_{a,n},\ n = -1, 0, 1, 2, \ldots\}$, where $\boldsymbol{\pi}_{a,n} = \boldsymbol{\pi}_n(I \otimes (D_1\mathbf{e}))/\lambda$, for $n = -1, 0, 1, 2, \ldots$.*

Proof. By their definition, $\boldsymbol{\pi}_n(I \otimes (D_1\mathbf{e}))$ includes all the arrival rates when the waiting queue length is n, at an arbitrary arrival. Thus, we must have $\lambda = \boldsymbol{\pi}_{-1}(I \otimes D_1)\mathbf{e} + \boldsymbol{\pi}_0(I \otimes D_1)\mathbf{e} + \boldsymbol{\pi}_1(I \otimes D_1)\mathbf{e} + \cdots = \boldsymbol{\pi}_{-1}(I \otimes D_1)\mathbf{e} + \boldsymbol{\pi}_0(I-R)^{-1}(I \otimes D_1)\mathbf{e}$, i.e., the total arrival rate. By Markov renewal theory (e.g., Grassmann & Tavakoli, 2007), the conclusion follows. This completes the proof of Proposition 5.4. □

Similar to Propositions 5.1, 5.2, and 5.3, based on Proposition 5.4, the distributions, means, and variances of the waiting queue length, server queue length, and total queue length seen by an arbitrary arrival can be obtained. Details are omitted.

Proposition 5.5. *In steady state, let $N_{a,w}$ be the number of batches of size K to be served just before the arrival of a customer, excluding the batch that is currently in service and the batch of the arriving customer. The distribution of $N_{a,w}$ is given by*

$$P\{N_{a,w} = n\} = \begin{cases} \sum_{j=-1}^{K-1} \boldsymbol{\pi}_{a,j}\mathbf{e}, & \text{for } n = 0; \\ \lambda^{-1}\boldsymbol{\pi}_0 R^{nK}(I-R)^{-1}(I-R^K)(I \otimes D_1)\mathbf{e}, & \text{for } n \geqslant 1. \end{cases} \tag{17}$$

The probability generating function and mean of $N_{a,w}$ are given by, for $0 \leqslant z \leqslant 1$,

$$E[z^{N_{a,w}}] = \sum_{j=-1}^{K-1} \boldsymbol{\pi}_{a,j}\mathbf{e} + z\lambda^{-1}\boldsymbol{\pi}_0 R^K (I - zR^K)^{-1}(I-R)^{-1}(I-R^K)(I \otimes D_1)\mathbf{e};$$
$$E[N_{a,w}] = \lambda^{-1}\boldsymbol{\pi}_0 R^K (I-R^K)^{-1}(I-R)^{-1}(I \otimes D_1)\mathbf{e}. \tag{18}$$

Proof. Using the distribution of $N_{a,w}$, we obtain

$$E[z^{N_{a,w}}] = \sum_{j=-1}^{K-1} \boldsymbol{\pi}_{a,j}\mathbf{e} + \sum_{n=1}^{\infty} z^n \left(\sum_{j=0}^{K-1} \boldsymbol{\pi}_{a,nK+j}\mathbf{e}\right) = \sum_{j=-1}^{K-1} \boldsymbol{\pi}_{a,j}\mathbf{e} + \frac{z\boldsymbol{\pi}_0 R^K}{\lambda} \sum_{n=1}^{\infty} z^{n-1} R^{K(n-1)} \left(\sum_{j=0}^{K-1} R^j \mathbf{e}\right)(I \otimes D_1)\mathbf{e}, \tag{19}$$

which leads to Eq. (18). This completes the proof of Proposition 5.5. □

The variance of $N_{a,w}$ can be found. Details are omitted.

6. Waiting time, sojourn time, and departure delays

To find the distributions of waiting time and sojourn time, the key is to find the phases of the underlying Markov chains of the service times, right after an arbitrary arrival. First, we find the distribution of the waiting time of an arbitrary arrival.

Let $q_w(t-)$ be the queue length just before time t. If a customer arrives at time t, $(I(t), I_1(t), \ldots, I_K(t))$ may change only if $q_w(t-) = -1$. We consider the process $\{(q_w(t-), I(t), I_1(t), \ldots, I_K(t)),\ t \geqslant 0\}$. By Markov renewal theory, the distribution of $\{(q_w(t-), I(t), I_1(t), \ldots, I_K(t)),\ t \geqslant 0\}$ right after an arbitrary arrival is given by $\{\mathbf{p}_{a,n},\ n = -1, 0, 1, \ldots\}$, where

$$\mathbf{p}_{a,-1}=\frac{1}{\lambda}\boldsymbol{\pi}_{-1}\begin{pmatrix}\boldsymbol{\alpha}_1 & & & \\ & (0,I_{m_{\max,\langle 1\rangle}}\otimes(0,\boldsymbol{\alpha}_2)) & & \\ & & \ddots & \\ & & & (0,I_{m_{\max,\langle K-1\rangle}}\otimes(0,\boldsymbol{\alpha}_K))\end{pmatrix}\otimes(D_1\mathbf{e});$$
$$\mathbf{p}_{a,n}=\lambda^{-1}\boldsymbol{\pi}_n(I\otimes(D_1\mathbf{e})),\quad n\geqslant 0. \tag{20}$$

Note that the definitions of $\{\mathbf{p}_{a,n},\ n=-1,0,1,\ldots\}$ and $\{\boldsymbol{\pi}_{a,n},\ n=-1,0,1,\ldots\}$ are different. On the other hand, since the service process does not change when a customer arrives, except for $q_w(t)=-1$, we actually have $\mathbf{p}_{a,n}=\boldsymbol{\pi}_{a,n}$, for $n\geqslant 0$.

If an arriving customer sees $n=jK, jK+1,\ldots,$ or $jK+K-1$ customers waiting, the waiting time of the customer is $V_{r,n}+V_{\max,K,1}+\cdots+V_{\max,K,j}$, where all the random variables are independent, $\{V_{\max,K,1},\ldots,V_{\max,K,j}\}$ are i.i.d.r.v.s with the same distribution as $V_{\max,K}$, for $j\geqslant 0$, and the random variable $V_{r,n}$ has a *PH*-distribution with *PH*-representation $(\mathbf{p}_{a,n}/\mathbf{p}_{a,n}\mathbf{e},\ T_{\max,\langle k\rangle})$, for $n\geqslant 0$. Note that $j=N_{a,w}$ for this case. Using the distribution of $V_{r,n}$, we find the distribution of the remaining completion time of all customers in service at an arbitrary arrival instant.

Proposition 6.1. *In steady state, let V_r be the remaining completion time of the current service(s) in all servers at an arbitrary arrival instant, i.e., the time until server K becomes available. Then V_r is PH-distributed with PH-representation $(\mathbf{p}_{a,\geqslant 0},\ T_{max,\langle k\rangle})$, where $\mathbf{p}_{a,\geqslant 0}=\lambda^{-1}\boldsymbol{\pi}_0(I-R)^{-1}(I\otimes(D_1\mathbf{e}))$. Note that $P\{V_r=0\}=\mathbf{p}_{a,-1}\mathbf{e}=1-\mathbf{p}_{a,\geqslant 0}\mathbf{e}$.*

Using the distribution of $V_{r,n}$, we also find the distribution of the waiting time. Next proposition shows that the waiting time has a matrix-exponential distribution (e.g., Asmussen & Bladt, 1996).

Proposition 6.2. *In steady state, let W_q be the waiting time of an arbitrary customer. Then $W_q=V_r+V_{\max,K,1}+\cdots+V_{\max,K,N_{a,w}}$ and W_q has a matrix-exponential distribution with matrix-exponential (ME) representation*

$$\begin{aligned}&\boldsymbol{\alpha}_w=\phi\left(\left(I_{m_{\max,\langle K\rangle}}\otimes(D_1\mathbf{e})\right)\right);\\ &T_w=I\otimes T_{\max,\langle K\rangle}+(R^K)'\otimes\left(\mathbf{T}^0_{\max,\langle K\rangle}\boldsymbol{\alpha}_{\max,\langle K\rangle}\right);\\ &\mathbf{T}^0_w=\lambda^{-1}\left(\boldsymbol{\pi}_0\left(\sum_{k=0}^{K-1}R^k\right)\right)'\otimes\mathbf{T}^0_{\max,\langle K\rangle},\end{aligned} \tag{21}$$

where $\phi(X)$ denote the direct-sum of matrix X (i.e., $\phi(X)$ is obtained by stringing out the elements of all the row vectors in X into a single row vector, starting from row 1), and X′ is the transpose of X. The cumulative distribution function and the mean of W_q can be obtained as

$$\begin{aligned}&P\{X\leqslant t\}=1-\boldsymbol{\alpha}_w\exp\{T_wt\}\left(-T_w^{-1}\mathbf{T}^0_w\right);\\ &E[W_q]=-\mathbf{p}_{a,\geqslant 0}T^{-1}_{\max,\langle K\rangle}\mathbf{e}+E[N_{a,w}]E[V_{\max,K}]=\boldsymbol{\alpha}_wT_w^{-2}\mathbf{T}^0_w.\end{aligned} \tag{22}$$

In addition, we must have $\mathbf{p}_{a,-1}\mathbf{e}+\boldsymbol{\alpha}_w(-T_w)^{-1}\mathbf{T}^0_w=1$.

Proof. First note that, if $q_w(t)=-1$, then $W_q=0$. By the definition of W_q, conditioning on the number of customers waiting in queue and the phases of services at the arrival epoch, we obtain, for $s\geqslant 0$,

$$\begin{aligned}E[e^{-sW_q}]&=\mathbf{p}_{a,-1}\mathbf{e}+\sum_{n=0}^{\infty}\left(\sum_{k=0}^{K-1}\mathbf{p}_{a,nK+k}\right)(sI-T_{\max,\langle K\rangle})^{-1}\mathbf{T}^0_{\max,\langle K\rangle}\left(f^*_{\max,K}(s)\right)^n\\ &=\mathbf{p}_{a,-1}\mathbf{e}+\frac{\boldsymbol{\pi}_0\left(\sum_{k=0}^{K-1}R^k\right)}{\lambda}\sum_{n=0}^{\infty}R^{nK}(I\otimes(D_1\mathbf{e}))(sI-T_{\max,\langle K\rangle})^{-1}\mathbf{T}^0_{\max,\langle K\rangle}\\ &\quad\times\left(f^*_{\max,K}(s)\right)^n=\mathbf{p}_{a,-1}\mathbf{e}+\frac{\boldsymbol{\pi}_0\left(\sum_{k=0}^{K-1}R^k\right)}{\lambda}\sum_{n=0}^{\infty}R^{Kn}(I\otimes(D_1\mathbf{e}))\\ &\quad\times((sI-T_{\max,\langle K\rangle})^{-1}\mathbf{T}^0_{\max,\langle K\rangle}\boldsymbol{\alpha}_{\max,\langle K\rangle})^n(sI-T_{\max,\langle K\rangle})^{-1}\mathbf{T}^0_{\max,\langle K\rangle}\\ &=\mathbf{p}_{a,-1}\mathbf{e}+\phi\left(\sum_{n=0}^{\infty}R^{Kn}(I\otimes(D_1\mathbf{e}))\left((sI-T_{\max,\langle K\rangle})^{-1}\mathbf{T}^0_{\max,\langle K\rangle}\boldsymbol{\alpha}_{\max,\langle K\rangle}\right)^n\right)\\ &\quad\times\left(\frac{\boldsymbol{\pi}_0\left(\sum_{k=0}^{K-1}R^k\right)}{\lambda}\right)'\otimes((sI-T_{\max,\langle K\rangle})^{-1}\mathbf{T}^0_{\max,\langle K\rangle})\\ &=\mathbf{p}_{a,-1}\mathbf{e}+\phi((I\otimes(D_1\mathbf{e})))\left(\sum_{n=0}^{\infty}\left(R^{K'}\otimes\left((sI-T_{\max,\langle K\rangle})^{-1}\mathbf{T}^0_{\max,\langle K\rangle}\boldsymbol{\alpha}_{\max,\langle K\rangle}\right)\right)^n\right)\\ &\quad\times\left(\frac{\boldsymbol{\pi}_0\left(\sum_{k=0}^{K-1}R^k\right)}{\lambda}\right)'\otimes\left((sI-T_{\max,\langle K\rangle})^{-1}\mathbf{T}^0_{\max,\langle K\rangle}\right)\\ &=\mathbf{p}_{a,-1}\mathbf{e}+\phi((I\otimes(D_1\mathbf{e})))\left(I-R^{K'}\otimes\left((sI-T_{\max,\langle K\rangle})^{-1}\mathbf{T}^0_{\max,\langle K\rangle}\boldsymbol{\alpha}_{\max,\langle K\rangle}\right)\right)^{-1}\\ &\quad\times\left(\frac{\boldsymbol{\pi}_0\left(\sum_{k=0}^{K-1}R^k\right)}{\lambda}\right)'\otimes((sI-T_{\max,\langle K\rangle})^{-1}\mathbf{T}^0_{\max,\langle K\rangle})\\ &=\mathbf{p}_{a,-1}\mathbf{e}+\phi((I\otimes(D_1\mathbf{e})))\left(sI-I\otimes T_{\max,\langle K\rangle}-R^{K'}\otimes\left(\mathbf{T}^0_{\max,\langle K\rangle}\boldsymbol{\alpha}_{\max,\langle K\rangle}\right)\right)^{-1}\\ &\quad\times\left(\frac{\boldsymbol{\pi}_0\left(\sum_{k=0}^{K-1}R^k\right)}{\lambda}\right)'\otimes\mathbf{T}^0_{\max,\langle K\rangle}=\mathbf{p}_{a,-1}\mathbf{e}+\boldsymbol{\alpha}_w(sI-T_w)^{-1}\mathbf{T}^0_w.\end{aligned} \tag{23}$$

This completes the proof of Proposition 6.2. □

Higher moments and variance of W_q can be found easily. Details are omitted.

Next, we study the sojourn time of an arbitrary arrival. We decompose $\mathbf{p}_{a,-1}$ according to the variable $I(t)$: $\mathbf{p}_{a,-1}=(\mathbf{p}_{a,-1,\langle 1\rangle},\mathbf{p}_{a,-1,\langle 2\rangle},\ldots,\mathbf{p}_{a,-1,\langle K\rangle})$, where $\mathbf{p}_{a,-1,\langle j\rangle}$ is for the case where the $(j-1)$st server is not available, and the jth server is available and is initialized for service of the new arrival. If an arrival sees an empty system, its waiting time is zero and its sojourn time is v_1. If an arrival sees server k $(\leqslant K)$ available and server $k-1$ is not available, its waiting time is zero and its sojourn time has a *PH*-distribution with *PH*-representation $(\mathbf{p}_{a,-1,\langle k\rangle},T_{\max,\langle k\rangle})$. If an arrival sees $n=jK+k-1$ customers waiting, then its sojourn time is $V_{r,n}+V_{\max,K,1}+\cdots+V_{\max,K,j}+V_{\max,k}$.

Proposition 6.3. *In steady state, let W_s be the sojourn time of an arbitrary arrival. Then W_s has a matrix-exponential distribution with ME-representation $\left(\boldsymbol{\alpha}_S,T_S,\mathbf{T}^0_S\right)$ given as*

$$\begin{aligned}&\boldsymbol{\alpha}_S=(\mathbf{p}_{a,-1,\langle 1\rangle},\ \cdots,\ \mathbf{p}_{a,-1,\langle K\rangle},\ \boldsymbol{\alpha}_w);\\ &T_S=\begin{pmatrix}T_{\max,\langle 1\rangle} & & & \\ & \ddots & & \\ & & T_{\max,\langle K\rangle} & \\ \lambda^{-1}(\boldsymbol{\pi}_0)'\otimes\mathbf{T}^0_{\max,\langle K\rangle}\boldsymbol{\alpha}_{\max,\langle 1\rangle} & \cdots & \lambda^{-1}\left(\boldsymbol{\pi}_0R^{K-1}\right)'\otimes\mathbf{T}^0_{\max,\langle K\rangle}\boldsymbol{\alpha}_{\max,\langle K\rangle} & T_w\end{pmatrix};\\ &\mathbf{T}^0_S=\begin{pmatrix}\mathbf{T}^0_{\max,\langle 1\rangle}\\ \vdots\\ \mathbf{T}^0_{\max,\langle K\rangle}\\ 0\end{pmatrix}.\end{aligned} \tag{24}$$

The mean sojourn time can be calculated as $E[W_S]=\boldsymbol{\alpha}_ST_S^{-2}\mathbf{T}^0_S$. In addition, we must have $\boldsymbol{\alpha}_S(-T_S)^{-1}\mathbf{T}^0_S=1$.

Proof. By definition, we have, for $s \geqslant 0$,

$$E[e^{-sW_S}] = \sum_{k=1}^{K} \mathbf{p}_{a,-1,\langle k\rangle}(sI - T_{\max,\langle k\rangle})^{-1}\mathbf{T}^0_{\max,\langle k\rangle} + \sum_{n=0}^{\infty}\sum_{k=1}^{K} \mathbf{p}_{a,nK+k-1}(sI - T_{\max,\langle K\rangle})^{-1}\mathbf{T}^0_{\max,\langle K\rangle}\left(f^*_{\max,K}(s)\right)^n f^*_{\max,k}(s). \tag{25}$$

The rest of the proof is similar to that of Proposition 6.2. Details are omitted. This completes the proof of Proposition 6.3. □

Applying Little's law, we obtain the following relationships: $E[q_w] = \lambda E[W_q]$ and $E[q_{tot}] = \lambda\, E[W_S]$, which are useful for checking the accuracy of computation.

In the tollbooth tandem queue, a customer may have to stay in the server after its service completion, until all down-stream services are completed. We call this period of time the *departure delay*. To find the distribution of departure delay, we need to find the phases of the underlying Markov chains of the services at servers, right after an arbitrary service completion. Define, $T^0_{\max,\langle k,0\rangle} = 0$ (zero matrix), and, for $1 \leqslant j \leqslant k \leqslant K$,

$$\begin{pmatrix} 0_{1\times(m_{\max,\langle j-1\rangle}+1)} \\ T^0_{\max,\langle k,j\rangle} \end{pmatrix} \equiv I_{m_1+1}\otimes\cdots\otimes I_{m_{j-1}+1}\otimes\begin{pmatrix} 0 \\ \mathbf{T}^0_j \end{pmatrix}\otimes\mathbf{e}_{m_{j+1}}\otimes\cdots\otimes\mathbf{e}_{m_k};$$

$$\begin{pmatrix} 0_{1\times(m_{\max,\langle j-1\rangle}+1)} \\ U_{\max,\langle k,j\rangle} \end{pmatrix} \equiv \begin{pmatrix} 1 & 0 \\ 0 & 0_{(m_1+1)\times(m_1+1)} \end{pmatrix}\otimes\cdots\otimes \begin{pmatrix} 1 & 0 \\ 0 & 0_{(m_{j-1}+1)\times(m_{j-1}+1)} \end{pmatrix}\otimes\begin{pmatrix} 0 \\ \mathbf{T}^0_j \end{pmatrix}\otimes\mathbf{e}_{m_{j+1}}\otimes\cdots\otimes\mathbf{e}_{m_k}. \tag{26}$$

Elements in the matrix $T^0_{\max,\langle k,j\rangle}$ are the service completion rates at server j, for states in $\Psi_{\max,\langle k\rangle}$. Define

$$\begin{aligned} &\boldsymbol{\delta}_{d,-1,\langle k\rangle,j} \equiv \boldsymbol{\pi}_{-1,k}(T^0_{\max,\langle k,j\rangle}\otimes\mathbf{e}), \quad \text{for } 1\leqslant j\leqslant k\leqslant K-1;\\ &\boldsymbol{\chi}_{d,-1,\langle k\rangle,j} \equiv \boldsymbol{\pi}_{-1,k}(U_{\max,\langle k,j\rangle}\otimes\mathbf{e}), \quad \text{for } 1\leqslant j\leqslant k\leqslant K-1;\\ &\boldsymbol{\delta}_{d,n,j} = \boldsymbol{\pi}_n\left(T^0_{\max,\langle K,j\rangle}\otimes\mathbf{e}\right), \quad \text{for } 1\leqslant j\leqslant K, n\geqslant 0;\\ &\boldsymbol{\chi}_{d,n,j} = \boldsymbol{\pi}_n(U_{\max,\langle K,j\rangle}\otimes\mathbf{e}), \quad \text{for } 1\leqslant j\leqslant K, n\geqslant 0. \end{aligned} \tag{27}$$

Then $\boldsymbol{\delta}_{d,-1,\langle k\rangle,j}$ contains the service completion rates at server j, where the server k is occupied and servers $\{k+1,\ldots,K\}$ are available (and no waiting customers). Also, $\boldsymbol{\delta}_{d,n,j}$ contain the service completion rates at server j, where there are n customers waiting. Define

$$\begin{aligned} &\left(\gamma_{j,0},\boldsymbol{\gamma}_j\right) = \sum_{k=j}^{K-1}\boldsymbol{\delta}_{d,-1,\langle k\rangle,j} + \sum_{n=0}^{\infty}\boldsymbol{\delta}_{d,n,j} = \sum_{k=j}^{K-1}\boldsymbol{\delta}_{d,-1,\langle k\rangle,j} + \boldsymbol{\pi}_0(I-R)^{-1}\times\left(T^0_{\max,\langle K,j\rangle}\otimes\mathbf{e}\right), \quad 1\leqslant j\leqslant K;\\ &(\eta_{j,0},\boldsymbol{\eta}_j) = \sum_{k=j}^{K-1}\boldsymbol{\chi}_{d,-1,\langle k\rangle,j} + \sum_{n=0}^{\infty}\boldsymbol{\chi}_{d,n,j} = \sum_{k=j}^{K-1}\boldsymbol{\chi}_{d,-1,\langle k\rangle,j} + \boldsymbol{\pi}_0(I-R)^{-1}\times\left(U_{\max,\langle K,j\rangle}\otimes\mathbf{e}\right), \quad 1\leqslant j\leqslant K;\\ &\zeta_j = (\gamma_{j,0},\boldsymbol{\gamma}_j)\mathbf{e}, \quad 1\leqslant j\leqslant K;\\ &\xi_j = (\eta_{j,0},\boldsymbol{\eta}_j)\mathbf{e}, \quad 1\leqslant j\leqslant K. \end{aligned} \tag{28}$$

Proposition 6.4. *In steady state, let $W_{d,k}$ be departure delay of an arbitrary customer completed its service at server k, and let W_d be departure delay of an arbitrary customer.*

(i) *Then ζ_k is the actual average service completion rate of server k, for $1 \leqslant k \leqslant K$. We must have $\lambda = \zeta_1 + \cdots + \zeta_K$, which is useful for checking the accuracy of computation.*
(ii) *The percentage of customers served by server k is ζ_k/λ, for $1 \leqslant k \leqslant K$.*
(iii) *The probability that a customer has a zero departure delay is $(\xi_1+\cdots+\xi_K)/\lambda$.*
(iv) *Given that a customer completes its service at server k, its departure delay $W_{d,k}$ is zero, for k = 1 (i.e., $W_{d,1} = 0$), and has a PH-distribution with PH-representation $(\boldsymbol{\gamma}_k/(\boldsymbol{\gamma}_k\mathbf{e}), T_{max,\langle k-1\rangle})$, for $2 \leqslant k \leqslant K$.*
(v) *The departure delay W_d of an arbitrary customer has a PH-distribution with PH-representation*

$$\boldsymbol{\alpha}_{D_d} = \frac{1}{\lambda}(\boldsymbol{\gamma}_2,\boldsymbol{\gamma}_3,\ldots,\boldsymbol{\gamma}_K); \quad T_{D_d} = \begin{pmatrix} T_{\max,\langle 1\rangle} & & \\ & \ddots & \\ & & T_{\max,\langle K-1\rangle} \end{pmatrix}. \tag{29}$$

(vi) *We have $E[W_S] = E[W_q] + E[V_1]\zeta_1/\lambda + E[V_2]\zeta_2/\lambda + \cdots + E[V_K]\zeta_K/\lambda + E[W_d]$. (This relationship is useful for checking the accuracy of computation.)*

7. Further properties of the tollbooth tandem queue

In this section, we present some advanced properties associated with the distributions of waiting time and sojourn time obtained in Section 6. We show that the waiting time and sojourn time are actually *PH*-distributed.

Proposition 7.1. *The ME-representation $\left(\boldsymbol{\alpha}_w, T_w, \mathbf{T}^0_w\right)$ of the waiting time W_q has an equivalent PH-representation $\left(\boldsymbol{\alpha}_{w,PH}, T_{w,PH}, \mathbf{T}^0_{w,PH}\right)$, given by $\boldsymbol{\alpha}_{w,PH} = \boldsymbol{\alpha}_w\Lambda_w$, $T_{w,PH} = (\Lambda_w)^{-1}T_w\Lambda_w$, $\mathbf{T}^0_{w,PH} = (\Lambda_w)^{-1}\mathbf{T}^0_w$, where $\Lambda_w = diag((\boldsymbol{\pi}_0(I-R)^{-1})'\otimes\mathbf{e})/\lambda$. (Note: diag $((\boldsymbol{\pi}_0(I-R)^{-1})'\otimes\mathbf{e})$ is a diagonal matrix with the vector $(\boldsymbol{\pi}_0(I-R)^{-1})'\otimes\mathbf{e}$ on the diagonal.)*

Proof. It is easy to see, for $s \geqslant 0$,

$$\mathbf{p}_{a,-1}\mathbf{e} + \boldsymbol{\alpha}_w(sI - T_w)^{-1}\mathbf{T}^0_w = \mathbf{p}_{a,-1}\mathbf{e} + \boldsymbol{\alpha}_w\Lambda_w\left(sI - \Lambda_w^{-1}T_w\Lambda_w\right)^{-1}\Lambda_w^{-1}\mathbf{T}^0_w. \tag{30}$$

Note that $\boldsymbol{\alpha}_w$ and $\mathbf{T}^0_w$ are nonnegative, the diagonal elements of T_w are negative and the off-diagonal elements of $\mathbf{T}_w$ are nonnegative. Then $\left(\boldsymbol{\alpha}_{w,PH}, T_{w,PH}, \mathbf{T}^0_{w,PH}\right)$ have the same property. To prove the proposition, it is sufficient to show that $\boldsymbol{\alpha}_{w,PH}$ is a substochastic vector and $T_{w,PH}\mathbf{e} + \mathbf{T}^0_{w,PH} = 0$. First, by Eq. (10), we have

$$\boldsymbol{\alpha}_{w,PH}\mathbf{e} = \boldsymbol{\alpha}_w\Lambda_w\mathbf{e} = \frac{\phi(I_{m_{\max,\langle K\rangle}}\otimes(D_1\mathbf{e}))\left((\boldsymbol{\pi}_0(I-R)^{-1})'\otimes\mathbf{e}\right)}{\lambda} = \frac{\boldsymbol{\pi}_0(I-R)^{-1}(\mathbf{e}\otimes I_{m_a})D_1\mathbf{e}}{\lambda} < 1. \tag{31}$$

Second, we have

$$\begin{aligned} T_{w,PH}\mathbf{e}+\mathbf{T}^0_{w,PH} &= \Lambda_w^{-1}\left(T_w\lambda^{-1}(\boldsymbol{\pi}_0(I-R)^{-1})'\otimes\mathbf{e}+\mathbf{T}^0_w\right)\\ &= \Lambda_w^{-1}\Big(\lambda^{-1}(-I+(R^K)')(\boldsymbol{\pi}_0(I-R)^{-1})'\\ &\quad +\lambda^{-1}(\boldsymbol{\pi}_0(I-R)^{-1}(I-R^K))'\Big)\otimes\mathbf{T}^0_{\max,\langle K\rangle}\\ &= \Lambda_w^{-1}(0)\otimes\mathbf{T}^0_{\max,\langle K\rangle}=0. \end{aligned} \tag{32}$$

This completes the proof of Proposition 7.1. □

While the *ME*-representation and the *PH*-representation for waiting time W_q are both useful in computing the distribution function and moments of W_q, the *PH*-representation is of particular interest if W_q is used in further modeling of the system, due to its probabilistic interpretation.

Proposition 7.2. *The ME-representation* $\left(\boldsymbol{\alpha}_S, T_S, \mathbf{T}^0_S\right)$ *of the sojourn time* W_s *has an equivalent PH-representation* $\left(\boldsymbol{\alpha}_{S,PH}, T_{S,PH}, \mathbf{T}^0_{S,PH}\right)$, *given by* $\boldsymbol{\alpha}_{S,PH}=\boldsymbol{\alpha}_S\Lambda_S$, $T_{S,PH}=(\Lambda_S)^{-1}T_S\Lambda_S$, $\mathbf{T}^0_{S,PH}=(\Lambda_S)^{-1}\mathbf{T}^0_S$, *where*

$$\Lambda_S=\begin{pmatrix} I_{m_{-1,\langle K\rangle}-1} & 0\\ 0 & \Lambda_w\end{pmatrix}. \tag{33}$$

In addition, we must have $\boldsymbol{\alpha}_{S,PH}\mathbf{e}=1$.

Proof. The proposition can be proved in a way similar to that of Proposition 7.1. The proposition can also be proved by using the probabilistic interpretation of $\left(\boldsymbol{\alpha}_{w,PH}, T_{w,PH}, \mathbf{T}^0_{w,PH}\right)$. When the underlying Markov chain associated with $\left(\boldsymbol{\alpha}_{w,PH}, T_{w,PH}, \mathbf{T}^0_{w,PH}\right)$ is absorbed, the customer of interest enters a server. The customer enters server k with rate(s) $\Lambda_w^{-1}\lambda^{-1}(\boldsymbol{\pi}_0R^{k-1})'\otimes\mathbf{T}^0_{\max,\langle K\rangle}$ and the service of the (first) k servers is initialized with $\boldsymbol{\alpha}_{\max,\langle k\rangle}$, for $1\leqslant k\leqslant K$. Then $\left(\boldsymbol{\alpha}_{S,PH}, T_{S,PH}, \mathbf{T}^0_{S,PH}\right)$ is a *PH*-representation for W_s. Finally, we have

$$\boldsymbol{\alpha}_{S,PH}\mathbf{e}=\mathbf{p}_{a,-1}\mathbf{e}+\boldsymbol{\alpha}_{w,PH}\mathbf{e}=\mathbf{p}_{a,-1}\mathbf{e}+\mathbf{p}_{\geqslant 0}\mathbf{e}=1. \tag{34}$$

This completes the proof of Proposition 7.2. □

We note that the *PH*-representation of the sojourn time W_s can also be found by using results in Ozawa (2006). Matrix-exponential representations of the waiting time and sojourn time for the *GI/PH/1* queue can be found in Sengupta (1989).

The geometric solution $\{\boldsymbol{\pi}_{-1}, \boldsymbol{\pi}_0R^n, n=0,1,2,\ldots\}$ given in Section 4 has a geometric tail, which further implies that the queue length distributions given in Section 5 have a geometric decay as well. Let ρ_R be the Perron–Frobenius eigenvalue of R (i.e., the eigenvalue with largest modulus). Then ρ_R is the decay rate of the queue length distributions. Next, we find the decay rates for the waiting time and sojourn time distributions. Define

$$T_{w,R}=T_{\max,\langle K\rangle}+\rho_R^K\mathbf{T}^0_{\max,\langle K\rangle}\boldsymbol{\alpha}_{\max,\langle K\rangle}. \tag{35}$$

Denote by $\lambda_{w,R}$ the eigenvalue of $T_{w,R}$ with the largest real part. Since ρ_R is nonnegative and less than one, $\lambda_{w,R}$ is negative. Since the Perron–Frobenius eigenvalue is nondecreasing in the elements of the matrix, we must have $\max\{\lambda_{k,1}, k=1,2,\ldots,K\}<\lambda_{w,R}<0$.

Proposition 7.3. *That* $\lambda_{w,R}$ *is the eigenvalue of* T_w *with the largest real part. The waiting time distribution has decay rate* $-\lambda_{w,R}$, *i.e.,* $P\{W_q>t\}=c_wt^k\exp(\lambda_{w,R}t)+o(t^k\exp(\lambda_{w,R}t))$, *for some constants* c_w *and* k.

Proof. It can be shown that T_w is irreducible. Then $\lambda_{w,R}$ is the only eigenvalue with the largest real part. Consequently, the waiting time distribution has a decay rate $-\lambda_{w,R}$. This completes the proof of Proposition 7.3. □

Proposition 7.4. *That* $\lambda_{w,R}$ *is the eigenvalue of* T_S *with the largest real part. The sojourn time distribution has decay rate* $-\lambda_{w,R}$.

Proof. By Proposition 7.3 and Eq. (24), the eigenvalue of T_S with the largest real part is $\lambda_{w,R}$. Consequently, the waiting time distribution has a decay rate $-\lambda_{w,R}$. This completes the proof of Proposition 7.4. □

Lastly, we present a limiting result.

Proposition 7.5. *Assume that all the service times have the same probability distribution. Let* $\rho_K=-\lambda\boldsymbol{\alpha}_{max,\langle k\rangle}(T_{max,\langle k\rangle})^{-1}\mathbf{e}/K$. *Then we have* $\rho_K\to 0$ *as* $K\to\infty$.

Proof. It is sufficient to show $E[\max\{X_1,\ldots,X_K\}]/K\to 0$ as $K\to\infty$. Denote by $F_s(t)$ the distribution function of the service times. Then we have

$$\begin{aligned} \lim_{K\to\infty}\frac{E[\max\{X_1,\cdots,X_K\}]}{K} &= \lim_{K\to\infty}\int_0^\infty x(F_s(x))^{K-1}dF_s(x)\\ &= \int_0^\infty x\lim_{K\to\infty}(F_s(x))^{K-1}dF_s(x)=0. \end{aligned} \tag{36}$$

This completes the proof of Proposition 7.5. □

8. A special case

In this section, we consider the special case where the service times at the different servers are independent and identically distributed.

We first consider the simplest case of a tollbooth tandem queue where customers arrive according to a Poisson process and there are two identical servers of exponential service times. Indeed, one would hope to obtain a closed form solution for this extreme case. However, the following analysis shows that, even though the results for this very special case are simpler, a closed form solution is still not possible.

Suppose that the arrival rate is λ and the service rate is μ for both servers. Since the service rates of the two servers are the same, the transition blocks in the infinitesimal generator Q in Eq. (6) can be simplified to

$$\begin{aligned} &A_{-1,-1}=\begin{pmatrix}-\lambda & \lambda\\ \mu & -\lambda-\mu\end{pmatrix},\quad A_{-1,0}=\begin{pmatrix}0 & 0\\ 0 & \lambda\end{pmatrix},\\ &A_{0,-1}=\begin{pmatrix}\mu & 0\\ 0 & 0\end{pmatrix},\quad A_{1,-1}=\begin{pmatrix}0 & \mu\\ 0 & 0\end{pmatrix},\\ &A_0=\lambda I,\quad A_1=-\lambda I+\begin{pmatrix}-\mu & 0\\ 2\mu & -2\mu\end{pmatrix},\quad A_3=\begin{pmatrix}0 & \mu\\ 0 & 0\end{pmatrix}. \end{aligned} \tag{37}$$

For the new Markov chain associated $q_w(t)$, its state space is reduced but information on which server is in service is partially lost. For this case, the elements of $R=(r_{i,j})$ can be obtained by solving nonlinear equations:

$$\begin{aligned} &r_{1,2}=\frac{(\lambda+\mu)}{2\mu}r_{1,1}-\frac{\lambda}{2\mu};\quad r_{2,1}=\frac{2\mu}{\lambda+\mu}r_{2,2};\\ &r_{1,1}^2+r_{1,1}r_{2,2}+r_{1,1}-\frac{\lambda}{\lambda+\mu}r_{2,2}-\frac{\lambda}{\mu}=0;\\ &r_{2,2}^2+r_{1,1}r_{2,2}+r_{2,2}-\frac{\lambda(\lambda+\mu)}{2\mu^2}=0. \end{aligned} \tag{38}$$

It is seen from these equations that there is no closed form solution for $R=(r_{i,j})$, and neither is there a closed form solution for $\{\boldsymbol{\pi}_{-1},\boldsymbol{\pi}_0\}$, though they can be computed numerically rather easily.

We next consider the general tollbooth tandem queue but with identical servers. The orders of the transition blocks and matrix representations can be reduced significantly for this case (see Eq. (5)). Define $N(t)$ as the number of servers busy at time t. We consider the CTMC $\{(q_w(t), I(t), N(t), I_1(t), \ldots, I_{N(t)}(t), I_a(t)),\ t \geqslant 0\}$. In the construction of the infinitesimal generator (6) for the CTMC, we use $\left\{\boldsymbol{\alpha}_{\max,[k]}, T_{\max,[k]}, \mathbf{T}^0_{\max,[k]}\right\}$ given in Eqs. (4) and (5), instead of $\left\{\boldsymbol{\alpha}_{\max,\langle k\rangle}, T_{\max,\langle k\rangle}, \mathbf{T}^0_{\max,\langle k\rangle}\right\}$ given in Eq. (2). We also use $\boldsymbol{\alpha}_k$ to replace $(0, \boldsymbol{\alpha}_k)$ in Eq. (7). The distributions of the following random variables can be found similar to the heterogeneous case in Sections 4–6: (i) The queue length (i.e., the number of customers waiting in queue) at an arbitrary time, arrival, and service completion; (ii) The queue length in service at an arbitrary time, arrival, and service completion; (iii) The waiting time; (iv) The sojourn time; and (v) The departure delay. On the other hand, distributions of the total number of customers in the system cannot be found by using the simplified CTMC, since the total number of customers in the servers depends on which server(s) has completed its service. Information on which server(s) has completed service is partially lost for the simplified CTMC.

Specifically, the infinitesimal generator of $\{(q_w(t), I(t), N(t), I_1(t), \ldots, I_{N(t)}(t), I_a(t)),\ t \geqslant 0\}$ is given by

$$A_{-1,-1} = \begin{pmatrix} 0 & \boldsymbol{\alpha} & & & \\ & 0 & (0_{m_{\max,[1]}\times m},\ I_{m_{\max,[1]}}\otimes\boldsymbol{\alpha}) & & \\ & & \ddots & \ddots & \\ & & & 0 & (0_{m_{\max,[K-2]}\times m},\ I_{m_{\max,[K-2]}}\otimes\boldsymbol{\alpha}) \\ & & & & 0_{m_{\max,[K-1]}\times m_{\max,[K-1]}} \end{pmatrix} \otimes D_1$$

$$+ I_{m_{-1,[K]}} \otimes D_0 + \begin{pmatrix} 0 & & & \\ \mathbf{T}^0_{\max,[1]} & T_{\max,[1]} & & \\ \vdots & & \ddots & \\ \mathbf{T}^0_{\max,[K-2]} & & & T_{\max,[K-2]} \\ \mathbf{T}^0_{\max,[K-1]} & & & \end{pmatrix} \otimes I_{m_a}; \qquad (39)$$

$$A_{-1,0} = \begin{pmatrix} 0_{m_{-1,[K-2]}\times m} & 0_{m_{-1,[K-2]}\times(m_{\max,[K]}-m)} \\ 0_{m_{\max,[K-1]}\times m} & I_{m_{\max,[K-1]}}\otimes\boldsymbol{\alpha} \end{pmatrix} \otimes D_1;$$

$$A_{0,-1} = \left(\mathbf{T}^0_{\max,[K]} \quad 0_{m_{\max,[K]}\times(m_{-1,[K]}-1)}\right) \otimes I_{m_a};$$

$$A_{j,-1} = \left(0_{m_{\max,[K]}\times m_{-1,[j-1]}} \quad \mathbf{T}^0_{\max,[K]}\boldsymbol{\alpha}_{\max,[j]} \quad 0_{m_{\max,[K]}\times(m_{-1,[K]}-m_{-1,[j]})}\right) \otimes I_{m_a},$$
$$1 \leqslant j \leqslant K-1;$$

$$A_0 = I_{m_{\max,[K]}} \otimes D_1;$$
$$A_1 = I_{m_{\max,[K]}} \otimes D_0 + T_{\max,[K]} \otimes I_{m_a};$$
$$A_{K+1} = \left(\mathbf{T}^0_{\max,[K]}\boldsymbol{\alpha}_{\max,[K]}\right) \otimes I_{m_a},$$

where $m_{-1,[0]} = 1$, and $m_{-1,[j]} = 1 + m_{\max,[1]} + \cdots + m_{\max,[j]}$, for $j = 1, 2, \ldots, K-1$.

The distributions of waiting queue length, waiting time, and sojourn time can be obtained similar to the heterogeneous case. For the departure delay, we define

$$T^0_{\max,[kj]} = \begin{pmatrix} 0_{m_{\max,[j-1]}\times m_{\max,[j-1]}} \\ \left(0_{m^{j-1}\times m_{\max,[j-2]}}, I_{m^{j-1}\times m^{j-1}}\right)\otimes\mathbf{T}^0 \\ \vdots \\ \left(0_{m^{j-1}\times m_{\max,[j-2]}}, I_{m^{j-1}\times m^{j-1}}\right)\otimes\mathbf{T}^0\otimes\mathbf{e}_{m^{k-j}\times 1} \end{pmatrix}, \quad \text{for } 2\leqslant j\leqslant k;$$

$$\boldsymbol{\gamma}_{[j]} = \sum_{k=j}^{K-1}\boldsymbol{\pi}_{-1,[k]}\left(T^0_{\max,[kj]}\otimes\mathbf{e}_{m_a\times 1}\right) + \boldsymbol{\pi}_0(I-R)^{-1}\left(T^0_{\max,[Kj]}\otimes\mathbf{e}_{m_a\times 1}\right),$$
$$\text{for } 2\leqslant j\leqslant K. \qquad (40)$$

Note that $\boldsymbol{\pi}_{-1} = (\boldsymbol{\pi}_{-1,[0]}, \boldsymbol{\pi}_{-1,[1]}, \ldots, \boldsymbol{\pi}_{-1,[K-1]})$. Then the departure delay W_d of an arbitrary customer has a *PH*-distribution with *PH*-representation

$$\boldsymbol{\alpha}_{W_d} = \frac{1}{\lambda}\sum_{j=2}^{K}(\boldsymbol{\gamma}_{[j]}, 0_{m_{\max,[K-1]}-m_{\max,[j]}}), \quad T_{W_d} = T_{\max,[K-1]}. \qquad (41)$$

For this case, we must have $E[W_s] = E[W_q] + E[V_1] + E[W_d]$, which is useful for checking computation accuracy.

As seen the analysis above, the main advantage of using the simplified CTMC is the reduction in the orders of transition blocks in Q and the matrices $\{R, T_w, T_S\}$. Table 1 below summarizes the orders of the matrices for the two CTMCs $\{(q_w(t), I(t), I_1(t), \ldots, I_K(t), I_a(t)),\ t \geqslant 0\}$ and $\{(q_w(t), I(t), N(t), I_1(t), \ldots, I_{N(t)}(t), I_a(t)),\ t \geqslant 0\}$ for the case with identical servers. Clearly, orders for the formulation of identical servers are significantly smaller than that for the formulation in Section 4. The forth column in Table 1 gives the orders of matrices for the tollbooth tandem queue with identical servers, Poisson arrival process, and exponential service times, which gives an even simpler computational algorithm.

9. Numerical analysis

The method developed in this paper can be used not only to analyze various system performance measures of the tollbooth tandem system, but also to address interesting system design issues. For example, in the tollbooth application, an important question is, when the number of tollbooths increases, how effectively does it improve the customer delays?

Example 9.1. We consider a tollbooth tandem queue with Poisson arrival process and K identical and exponential servers. The arrival rate is $\lambda = 2$ and service rate is $\mu = 1$. In Table 2, we report the numerical results for mean number of waiting customers, mean number of customers in system, mean waiting time, mean sojourn time, mean number of batches of size K observed by an arrival, and mean departure delay, as a function of the K.

Table 1
Orders of transition blocks and matrices $\{R, T_w, T_S\}$.

Matrix	Section 4 formulation	Section 8 formulation	Poisson & Exponential
$A_{-1,-1}$	$(2 + K + ((m+1)^K - 1)/m)m_a$	$(1 + (K-1)m + \cdots + 2m^{K-2} + m^{K-1})m_a$	$1 + (K-1)K/2$
A_0, A_1, A_{K+1}, R	$((m+1)^K - 1)m_a$	$(m + \cdots + m^{K-2} + m^K)m_a$	K
T_w	$((m+1)^K - 1)^2 m_a$	$(m + \cdots + m^{K-2} + m^K)^2 m_a$	K^2
T_S	$(1 + K + ((m+1)^K - 1)/m)m_a + ((m+1)^K - 1)^2 m_a$	$(1 + (K-1)m + \cdots + 2m^{K-2} + m^{K-1})\ m_a + (m + \cdots + m^{K-2} + m^K)^2 m_a$	$1 + (K-1)K/2 + K^2$

Table 2
Performance measures for the Poisson arrival and exponential service case.

K	ρ	$E[q_w]$	$E[N_{a,w}]$	$E[W_q]$	$E[W_s]$	$E[W_d]$
5	0.9133	10.987	1.8860	5.4937	7.1803	0.6866
6	0.8167	4.0308	0.4424	2.0154	3.7368	0.7214
10	0.5858	0.8669	0.0146	0.4335	2.1919	0.7585
15	0.4424	0.3516	0.0006	0.1758	1.9341	0.7583
20	0.3598	0.1843	0	0.0922	1.8478	0.7556
25	0.3053	0.1054	0	0.0527	1.8067	0.7539
30	0.2663	0.0627	0	0.0313	1.7843	0.7530
40	0.2139	0.0234	0	0.0117	1.7638	0.7521
50	0.18	0.0090	0	0.0045	1.7563	0.7518
∞	0	0	0	0	1.7516	0.7516

Table 2 shows that, as K increases, the waiting queue length, customer waiting time, and sojourn time all decrease, as expected. In addition, the reduction in these performance measures is more significant when K is small. Thus, adding new tollbooths has a decreasing marginal effect. For example, the mean waiting time drops from 5.5 to 2.0 when the number of servers goes from 5 to 6, and then drops from 2.0 to 0.4 as four more servers are added. These numerical results show that for practical purposes, adding a small number of tandem booths would capture the most of the benefits. The mean queue length and mean sojourn time converge to their limits, and these limits correspond to the numbers for tollbooth tandem queue with infinite number of servers.

The impact of adding more tandem servers on departure delay does not have a clear pattern. For instance, it is seen from the numerical examples above that the departure delay first increases and then decreases. The mean departure delay also converges to a limit, which is the mean departure delay for the infinite-server system.

The second question we want to address in this section is, how does the order of servers in the tollbooth tandem system affect the system performance? That is, given a fixed number of servers with known service time distributions, what can be said about the arrangement of these servers that optimizes certain system performance measure? Next, we use four numerical examples to examine (i) the effect of mean service times and variances of service times on performance measures; and (ii) the arrangements of servers at which performance measures are optimized.

For the computation of general tollbooth tandem with non-Poisson arrivals and non-exponential processing times, we employ the following computational procedure:

1. Input system parameters: (m_a, D_0, D_1), K, and $\{(m_k, \boldsymbol{\alpha}_k, T_k), 1 \leqslant k \leqslant K\}$. Compute arrival rate λ and service rates $\{\mu_k, 1 \leqslant k \leqslant K\}$.
2. Use Eq. (2) to construct *PH*-representation $(m_{\max,\langle k\rangle}, \boldsymbol{\alpha}_{\max,\langle k\rangle}, T_{\max,\langle k\rangle})$ of $V_{\max,k} = \max\{V_1, \ldots, V_k\}$, for $1 \leqslant k \leqslant K$. Check system stability by Theorem 4.1.
3. Use Eq. (7) to construct transition blocks $\{A_{j,-1}, -1 \leqslant j \leqslant K-1, A_{-1,0}, A_0, A_1, A_{K+1}\}$ of the Markov chain Q.
4. Find $\{\boldsymbol{\pi}_{-1}, \boldsymbol{\pi}_0, R\}$ for the matrix-geometric solution using Eqs. (8) and (9).
5. Use Eqs. (11)–(19) to compute performance measures related to queue lengths, including the mean queue lengths $\{E[q_w], E[q_s], E[q_{tot}]\}$.
6. Construct the *ME*-representations of the waiting time and sojourn time using Eqs. (21) and (24). Use the *ME*-representations to compute the mean waiting time $E[W_q]$ and mean sojourn time $E[W_s]$.
7. Using Eqs. (26)–(28) to compute performance measures for individual servers including the percentages of customers served by servers $\{\zeta_1/\lambda, \ldots, \zeta_K/\lambda\}$.
8. Using Eq. (29) to construct a *PH*-representation of the departure delay. Use the *PH*-representation to compute the mean departure delay $E[W_d]$.

The base model for our comparison is presented in the following example.

Example 9.2. We consider a tollbooth tandem queue with three servers. The arrival process and service times are defined as follows:

$$
\begin{aligned}
&m_a = 2, \quad D_0 = \begin{pmatrix} -5 & 1 \\ 0 & -1 \end{pmatrix}, \quad D_1 = \begin{pmatrix} 4 & 0 \\ 0.1 & 0.9 \end{pmatrix};\\
&m_1 = 1, \quad \boldsymbol{\alpha}_1 = 1, \quad T_1 = -3, \quad \mathbf{T}_1^0 = 3;\\
&m_2 = 2, \quad \boldsymbol{\alpha}_2 = (0.5, 0.5), \quad T_2 = \begin{pmatrix} -2 & 1 \\ 1 & -4 \end{pmatrix}, \quad \mathbf{T}_2^0 = \begin{pmatrix} 1 \\ 3 \end{pmatrix};\\
&m_3 = 2, \quad \boldsymbol{\alpha}_3 = (0.8, 0.2), \quad T_3 = \begin{pmatrix} -3 & 1 \\ 1 & -1 \end{pmatrix}, \quad \mathbf{T}_3^0 = \begin{pmatrix} 2 \\ 0 \end{pmatrix}.
\end{aligned} \tag{42}
$$

By routine calculations, we obtain $\lambda = 1.2727$, $\mu_1 = 3$, $\mu_2 = 1.75$, $\mu_3 = 0.83333$, and $\rho = \lambda E[\max\{V_1, V_2, V_3\}]/3 = 0.6304$, for the queue. The mean queue lengths, mean waiting time, mean sojourn time, and mean departure delay are given in Table 3. Note that, in Tables 3–6, we use μ_i to represent the ith *PH*-distribution given in Eq. (42).

As shown in Table 3, the mean departure delay is small. This is due to the fact that server 1 is the faster server with mean service time 1/3, while servers 2 and 3 are slower; and server 2 is faster than server 3.

Table 3
Performance measures for Example 9.2.

Server			$E[q_w]$	$E[q_s]$	$E[q_{tot}]$	$E[W_q]$	$E[W_s]$	$E[W_d]$	ζ_1/λ	ζ_2/λ	ζ_3/λ
1	2	3									
μ_1	μ_2	μ_3	2.1675	0.8310	2.9985	1.7031	2.356	0.0866	0.5413	0.2617	0.1969

Table 4
Performance measures for Example 9.3.

Server			$E[q_w]$	$E[q_s]$	$E[q_{tot}]$	$E[W_q]$	$E[W_s]$	$E[W_d]$	ζ_1/λ	ζ_2/λ	ζ_3/λ
1	2	3									
μ_1	μ_2	μ_3	2.1675	0.8310	2.9985	1.7031	2.356	0.0866	0.5413	0.2617	0.1969
μ_1	μ_3	μ_2	2.3869	1.0711	3.4580	1.8755	2.717	0.2275	0.5197	0.2647	0.2156
μ_2	μ_3	μ_1	2.5494	1.3232	3.872	2.003	3.043	0.3435	0.4699	0.2895	0.2406
μ_2	μ_1	μ_3	2.3434	1.0259	3.3693	1.8412	2.647	0.1679	0.4931	0.2907	0.2162
μ_3	μ_1	μ_2	2.8902	1.7177	4.6079	2.2708	3.620	0.5837	0.4280	0.3129	0.2591
μ_3	μ_2	μ_1	2.8673	1.7591	4.6263	2.2528	3.635	0.6061	0.4253	0.3111	0.2636

Table 5
Performance measures for Example 9.4.

Server			$E[q_w]$	$E[q_s]$	$E[q_{tot}]$	$E[W_q]$	$E[W_s]$	$E[W_d]$	ζ_1/λ	ζ_2/λ	ζ_3/λ
1	2	3									
μ_1	μ_2	μ_3	17.333	1.0238	18.357	13.619	14.423	0.2322	0.5445	0.2480	0.2075
μ_1	μ_3	μ_2	16.081	1.2416	18.322	13.420	14.396	0.3714	0.5273	0.2519	0.2207
μ_2	μ_1	μ_3	11.015	1.1171	12.133	8.6553	9.5331	0.2358	0.4783	0.2970	0.2246
μ_2	μ_3	μ_1	10.039	1.3765	11.416	7.8884	8.9699	0.3850	0.4559	0.2938	0.2503
μ_3	μ_1	μ_2	12.178	1.8093	13.987	9.5689	10.990	0.6636	0.4168	0.3167	0.2665
μ_3	μ_2	μ_1	10.966	1.7986	12.764	8.6162	10.029	0.6457	0.4149	0.3135	0.2717

Table 6
Performance measures for Example 9.5.

	Server					$E[q_w]$	$E[q_s]$	$E[q_{tot}]$	$E[W_q]$	$E[W_s]$	$E[W_d]$
	1	2	3	4	5						
Min	μ_1	μ_2	μ_3	μ_4	μ_5	11.873	3.2779	15.151	11.873	15.151	1.2772
	μ_1	μ_2	μ_4	μ_5	μ_3	11.832	3.3056	15.138	11.832	15.138	1.3037
	μ_3	μ_5	μ_4	μ_2	μ_1	11.761	3.4572	15.218	11.761	15.218	1.4482
Max	μ_5	μ_4	μ_3	μ_2	μ_1	11.769	3.5083	15.278	11.769	15.278	1.4966
	μ_5	μ_3	μ_2	μ_1	μ_4	11.838	3.4570	15.295	11.838	15.295	1.4476
	μ_3	μ_1	μ_2	μ_4	μ_5	11.885	3.3234	15.208	11.885	15.208	1.3200

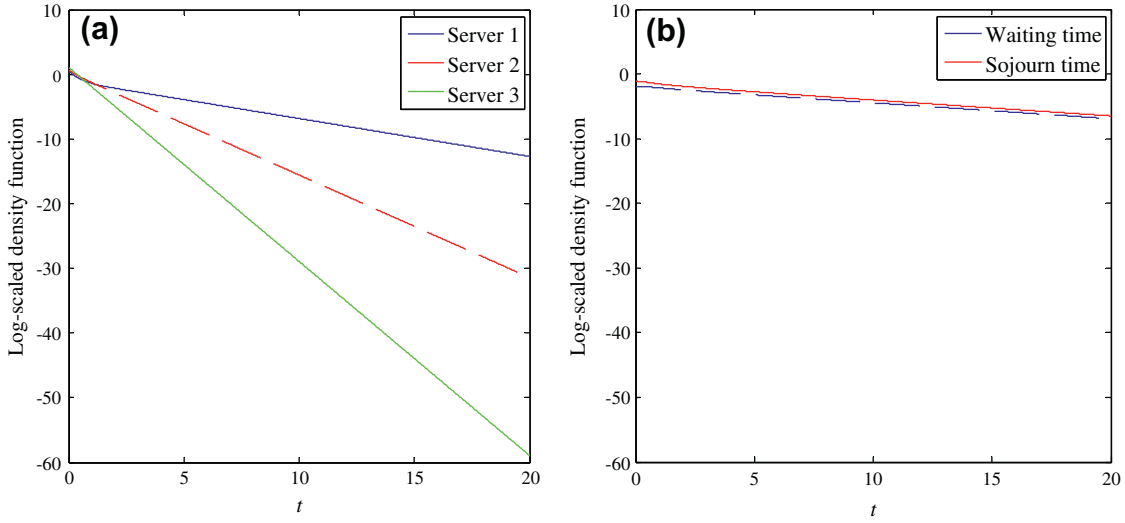

Fig. 2. Service times, waiting time, and sojourn time distributions for Example 9.2.

The log-scaled density functions of the service times (i.e., log ($f(t)$), where $f(t)$ is a density function) are plotted in Fig. 2(a), and the log-scaled density functions of the waiting time and sojourn time in Fig. 2(b). Fig. 2 demonstrates that the waiting time and sojourn time not only are significantly longer than the service times, but also have a much heavier tail. Thus, the waiting time and sojourn time are much more variable than service times.

To see the effect of the service speed (i.e., service rate) on performance measures, we switch the service times of servers.

Example 9.3 (*Example 9.2 continued*). By switching the three service times defined in Eq. (42) between the three servers, we generate six tollbooth tandem queues given below. Each of the six queues corresponds to an *arrangement* of the service times. Performance measures of the six tollbooth tandem queues are presented in Table 4.

Table 4 shows clearly that, if faster servers are used as the downstream servers (i.e., servers with a smaller index), the queue length, waiting time, and departure delay are all shorter. On the other hand, the actual workloads of the three servers are less balanced. The difference in the mean departure delay between the worst scenario and the best scenario is so significant that the configuration with the faster servers used in downstream should be considered in the design of such traffic systems.

For Examples 9.2 and 9.3, the standard deviations of the three service times are $\sigma_1 = 0.3333$, $\sigma_2 = 0.6061$, and $\sigma_3 = 1.5362$, respectively. It is readily seen that the standard deviations are ordered the same as the mean service times in Examples 9.2 and 9.3. Next, we change the standard deviation of service time 1 from $\sigma_1 = 0.3333$ to $\sigma_1 = 3.3334$, while the mean service time remains at $1/\mu_1 = 0.33333$. For this case, service time 1 has the smallest mean but the largest standard deviation. Under this condition, we examine the impact of the arrangement of the service times on performance measures.

Example 9.4 (*Example 9.3 continued*). In Examples 9.1 and 9.2, we change $(\boldsymbol{\alpha}_1, T_1)$ to

$$m_1 = 2, \quad \boldsymbol{\alpha}_1 = (0.005, 0.995), \quad T_1 = \begin{pmatrix} -0.0299 & 0 \\ 0 & -5.985 \end{pmatrix},$$
$$\mathbf{T}_1^0 = \begin{pmatrix} 0.0299 \\ 5.985 \end{pmatrix}. \tag{43}$$

The service rates and standard deviations of the three servers are given by $\mu_1 = 3$, $\mu_2 = 1.75$, $\mu_3 = 0.83333$, and $\sigma_1 = 3.3334$, $\sigma_2 = 0.6061$, and $\sigma_3 = 1.5362$, respectively. Performance measures of the tollbooth tandem queues are given in Table 5.

Table 5 indicates that, if the service time of server 1 has a small mean but it is too variable, the mean queue length can be significantly larger than that of the other cases. Thus, similar to the service rates, the variances of service times also have significant impact on performance measures. Based on numerical experiments and intuition, we have the following observations on the order of servers.

(i) To achieve a smaller waiting (sojourn) time, servers with a smaller mean service time (i.e., a greater service rate) should be used as downstream servers.
(ii) To achieve a smaller waiting (sojourn) time, servers with a smaller standard deviation of the service time should be used as downstream servers.

Example 9.4 indicates that the order of servers to achieve a smaller waiting (sojourn) time depends on both the mean service times and the standard deviations of service time. The next example demonstrates that the minimums and maximums of performance measures can be achieved at different arrangements of service times.

Example 9.5 (*The tollbooth tandem queue with K heterogeneous exponential server*). Consider a tollbooth tandem queue with a Poisson arrival process and K heterogeneous exponential servers. Assume that $\lambda = 1$, $K = 5$, $\mu_1 = 0.54$, $\mu_2 = 0.52$, $\mu_3 = 0.50$, $\mu_4 = 0.48$, and $\mu_5 = 0.46$. For this example, $\rho = 0.918$ and there are 120 permutations to assign the five service times to servers, which correspond to 120 tollbooth tandem queues. Performance measures for the best and worst cases are presented in Table 6.

Numerical results demonstrate that, if faster servers are used as downstream servers, i.e., $(\mu_1, \mu_2, \mu_3, \mu_4, \mu_5)$, $E[q_s]$ and $E[W_d]$ are minimized; performance measures $E[q_{tot}]$ and $E[W_s]$ are minimized at $(\mu_1, \mu_2, \mu_4, \mu_5, \mu_3)$; and performance measures $E[q_w]$ and $E[W_q]$ are minimized at $(\mu_3, \mu_5, \mu_4, \mu_2, \mu_1)$. On the other hand, $E[q_w]$ and $E[W_q]$ are maximized at $(\mu_3, \mu_1, \mu_2, \mu_4, \mu_5)$; $E[q_{tot}]$ and $E[W_s]$ are minimized at $(\mu_5, \mu_3, \mu_2, \mu_1, \mu_4)$; and $E[q_s]$ and $E[W_d]$ are maximized at $(\mu_5, \mu_4, \mu_3, \mu_2, \mu_1)$. Note that $E[q_w]$ and $E[W_q]$ (and $E[q_{tot}]$ and $E[W_s]$) are minimized or maximized at the same arrangement is due to the well-known Little's law.

This numerical example shows that the intuitive result of scheduling the faster servers at the most downstream stations does not necessary minimize the total time a customer spends in system nor the total number of customers in system. Based on all tested examples with $K = 2$, 3, 4, and 5, we have the following observations on the optimal arrangements of the servers.

Observation 9.1. Assume that $\mu_1 \geqslant \mu_2 \geqslant \cdots \geqslant \mu_{K-1} \geqslant \mu_K$. Then

(i) $E[q_s]$ and $E[W_d]$ are minimized at $(\mu_1, \mu_2, \ldots, \mu_{K-1}, \mu_K)$.
(ii) $E[q_s]$ and $E[W_d]$ are maximized at $(\mu_K, \mu_{K-1}, \ldots, \mu_3, \mu_2, \mu_1)$.
(iii) For the case with $K = 2$, all performance measures are minimized at (μ_1, μ_2), and maximized at (μ_2, μ_1).

We conjecture that some of these observations may be true in general, e.g., (iii) with two exponential servers, though we are unable to prove it theoretically. Numerical examples indicates that $E[q_w]$, $E[W_q]$, $E[q_{tot}]$, and $E[W_s]$ can be minimized or maximized at various arrangements of service times. Numerical examples also indicate that where a performance measure is minimized or maximized has much to do with the traffic intensity, i.e., ρ. For instance, if ρ is small, then all performance measures in Table 6 are minimized at $(\mu_1, \mu_2, \mu_3, \ldots, \mu_{K-1}, \mu_K)$; and if ρ is close to one, the performance measures in Table 6 can be minimized or maximized at very different arrangements of $\{\mu_1, \mu_2, \mu_3, \ldots, \mu_{K-1}, \mu_K\}$.

10. Conclusion

In this paper, computational methods are developed for computing the distributions of queue lengths, waiting time, sojourn time, and departure delay in a tollbooth tandem queue with any number of heterogeneous servers. The paper provides an algorithmic approach to the design and performance analysis of real tollbooth tandem systems and gas stations. In the design of tollbooth tandem queues, the order of servers is an important issue and the conjectures and observations obtained in Examples 9.2, 9.3, 9.4, 9.5 can be used to address the issue. Further studies are challenging and interesting.

Acknowledgements

The authors are grateful to the anonymous referees for their detailed comments and suggestions that have helped to significantly improve the exposition of this paper. The research of Xiuli Chao is supported in by part NSF under CMMI-0927631 and CMMI-1131249.

References

Asmussen, S., & Bladt, M. (1996). Renewal theory and queueing algorithms for matrix-exponential distributions. In A. S. Alfa & S. Chakravarthy (Eds.), *Proceedings of the first international conference on matrix analytic methods in stochastic models* (pp. 313–341). New York: Marcel Dekker.

Bini, D. A., Meini, B., Steffe, S., & Van Houdt, B. (2006). Structured Markov chain solver: The algorithms. In *Proceedings of the SMCTOOLS workshop, Pisa.*

Brewer, J. W., & Henwood, K. (1980). A dynamic model of the purchase of gasoline. *Applied Mathematical Modelling, 4*(3), 205–211.

Chaudhry, M. L., & Templeton, J. G. C. (1983). *First course in bulk queues*. New York: John Wiley and Sons.

Daskin, M. S., Shladover, S., & Sobel, K. (1976). An analysis of service station queues under gasoline shortage conditions. *Computers and Operations Research, 3*(1), 83–93.

Disney, R. L., & Mitchell, W. E. (1971). A solution for queues with instantaneous jockeying and other customer selection rules. *Naval Research Logistics, 17*, 315–325.

Grassmann, W. K., & Tavakoli, J. (2007). A Bayesian approach to find random-time probabilities from embedded Markov chain probabilities. *Probability in the Engineering and Informational Sciences, 21*, 551–556.

Gu, W., & Li, Y. (2011). Checkpoint congestion mitigation via tandem service booths: Capacity and delay analysis. In *The 90th annual meeting of transportation research board, Washington DC, USA.*

Gu, W., Cassidy, M. J., & Li, Y. (2012). On the capacity of highway checkpoints: Models for unconventional configurations. *Transportation Research Part B, 46*, 1308–1321.

Hall, R. W., & Daganzo, C. F. (1983). Tandem tollbooths for the Golden Gate Bridge. *Transportation Research Record, 905*, 7–14.

Hong, Y. C., Kim, D. K., Kho, S. Y., Kim, S. W., & Yang, H. (2009). Modeling and simulation of tandem tollbooth operations with max-algebra approach. In *FGIT '09 Proceedings of the first international conference on future generation information technology* (pp. 138–150). Berlin, Heidelberg, Germany: Springer-Verlag.

Horn, R. A., & Johnson, C. R. (1991). *Topics in matrix analysis*. Cambridge, England: Cambridge University Press.

Latouche, G., & Ramaswami, V. (1999). *Introduction to matrix analytic methods in stochastic modeling*. Philadelphia, USA: ASA & SIAM.

Lucantoni, D. M. (1991). New results on the single server queue with a batch Markovian arrival process. *Stochastic Models, 7*, 1–46.

Manitz, M. (2008). Queueing-model based analysis of assembly lines with finite buffers and general service times. *Computers & Operations Research, 35*(8), 2520–2536.

Melamed, B. (1986). A note on the reversibility and duality of some tandem blocking queueing systems. *Management Science, 32*, 128–130.

Neuts, M. F. (1975). Probability distributions of phase type. In *Liber Amicorum Prof. Emeritus H. Florin* (pp. 173-206). University of Louvain.
Neuts, M. F. (1979). A versatile Markovian point process. *Journal of Applied Probability, 16*, 764–779.
Neuts, M. F. (1981). *Matrix-geometric solutions in stochastic models – An algorithmic approach*. Baltimore: The Johns Hopkins University Press.
Ozawa, T. (2006). Sojourn time distributions in the queue defined by a general QBD process. *Queueing Systems, 53*, 203–211.
Papadopoulos, H. T., & O'Kelly, M. E. J. (1993). Exact analysis of production lines with no intermediate buffers. *European Journal of Operational Research, 65*(1), 118–137.
Rubenstein, L. D. (1983). Tandem toll collection systems. *Transportation Research Record, 905*, 14–17.
Sengupta, B. (1989). Markov processes whose steady state distribution is matrix-exponential with an application to the GI/PH/1 queue. *Advances in Applied Probability, 21*, 159–180.
Teimoury, E., Yazdi, M. M., Haddadi, M., & Fathi, M. (2011). Modelling and improvement of non-standard queuing systems: A gas station case study. *International Journal of Applied Decision Sciences, 4*(4), 324–340.
Van Houdt, B., & Alfa, A. S. (2005). The response time in a discrete time tandem queue with blocking, Markovian arrivals and phase-type services. *Operations Research Letters, 33*, 373–381.
Zhao, Y. Q., & Grassmann, W. K. (1995). Queueing analysis of a jockeying model. *Operations Research, 43*, 520–529.