Performance Evaluation 93 (2015) 17–26

Contents lists available at ScienceDirect

Performance Evaluation

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

Construction of Markov chains for discrete time *MAP/PH/K* queues

Qi-Ming He [a,*], Attahiru Sule Alfa [b,c]

[a] Department of Management Sciences, University of Waterloo, 200 University Avenue West, Waterloo, Ontario, Canada N2L 3G1
[b] Department of Electrical & Computer Engineering, University of Manitoba, Winnipeg, Manitoba, Canada R3T 5V6
[c] Department of Electrical, Electronic and Computer Engineering, University of Pretoria, Pretoria, South Africa

ARTICLE INFO

Article history:
Received 8 March 2015
Accepted 6 August 2015
Available online 14 August 2015

Keywords:
Markov chain
Queueing system
Markovian arrival process
Phase-type distribution

ABSTRACT

Two *GI/M/*1 type Markov chains associated with the queue length are often used in analyzing the discrete time *MAP/PH/K* queue. The first Markov chain is introduced by tracking service phases for servers; a method we call TPFS. The transition probability matrix of the Markov chain can be constructed in a straightforward manner. The second Markov chain is introduced by counting servers for phases; which we call CSFP. An algorithm is developed for the construction of the transition probability matrix of the second Markov chain, which is the main contribution of this paper. Whereas the construction of the matrices for the case of continuous time is available in the literature, it is not available for the discrete time case. The effort in constructing the matrices for the discrete time case is extensively more involved than for the continuous time case. Some basic properties of the constructed transition blocks are shown. We demonstrate that for queueing systems with a large number of servers and many service phases, there is a considerable saving in the matrix sizes. For example, when those values are 30 and 2, respectively, the block size for TPFS is more than 3×10^7 times that of CSFP; a major saving.

1. Introduction

Queueing systems are very pervasive in life and analysts are continuously developing mathematical tools for analyzing them. Surprisingly enough, despite the fact that queueing problems have been with us for a very long time and documentation of formal mathematical tools for analyzing them have been known for more than 105 years, some queueing systems are still difficult to analyze. Even though queueing systems could be in the form of a network, usually a decomposition of the system into a set of connected single nodes queues seems a very popular and reasonably approximating approach for analyzing them. When considering single node systems, however, the number of parallel servers involved in the system is usually more than one, especially in telecommunication systems where queueing models are receiving more significant attention these days. Multiserver queues are a major type of queueing models encountered in real life situations especially in telecommunications. For example in wireless communications we are usually dealing with multiple channels, hence multiserver systems. Unless the service time of each channel follows the exponential distribution or geometric distribution, in the discrete time case, analyzing such multichannel systems is usually quite involved especially because the associated transition block matrices, when using matrix analytical methods (MAM), could be huge in size and complicated to generate.

* Corresponding author.
E-mail addresses: q7he@uwaterloo.ca (Q.-M. He), attahiru.alfa@umanitoba.ca (A.S. Alfa).

http://dx.doi.org/10.1016/j.peva.2015.08.001
0166-5316/

As pointed out by Yue and Matsumoto [1], the modeling of discrete-time, multimedia communication systems are more complex than that of continuous-time systems because multiple state changes can occur from one time-unit to the next. This challenge has limited telecommunication analysts to using geometric distributions instead of actual distributions that more properly represent the service systems when dealing with discrete time systems. In fact it is very common for researchers to use continuous time models as alternative to discrete-time ones in order to avoid the challenges. The goal of this paper is to come up with very efficient methods for analyzing multiserver queues without restricting the service time distributions to the geometric distribution.

Telecommunication systems these days are studied in discrete time more than in continuous time [2]. This is mainly because the systems are now more digital than analog. However, with this more realistic system representation comes some additional price, that of computational aspects. As such, systems that are more appropriately modeled as the *MAP/PH/K* queue are approximated by *MAP/Geo/K* system, or even as *MAP/D/K* system, in order to cut down the computational efforts required. In this paper we study the *MAP/PH/K* system and show how to get around one of the challenges involved in its analysis. Analyzing this system using the MAM approach leads to a *GI/M/*1 structure with very huge block matrices [3,4]. For example, if the *MAP* is of order m_a, the *PH* of order m_s, then we could have block matrices of size $m_a m_s^K$, if we record the phases of each server that is busy; a method we call Track-Phase-for-Server (TPFS). Rather we develop a procedure that we call Count-Server-for-Phase (CSFP), which involves keeping the count of the number of busy servers in each phase. This reduces the block sizes to dimension $L = m_a(K + m_s - 1)!/(K!(m_s - 1)!)$. For large m_s and K, L is much smaller than $m_a m_s^K$. However constructing transition blocks for this case is very involved and that is the contribution of this paper. Surprisingly many researchers [5,6] have mentioned it in their papers as a way to get around the size issue. Ramaswami [7] did present an algorithm for constructing the block matrices of the generator matrix for the case of continuous time (also see [8]). However, there is no documentation until now on how to construct the block matrices for the discrete time case. While the construction of the transition blocks for the continuous time case is not straightforward, the construction process for the discrete time case is even more involved due to the fact that several events can occur simultaneously in discrete time, as pointed out earlier. The process for constructing the block matrices for CSFP for the discrete time is quite involved. Given that discrete time models are more relevant these days when it comes to applications to telecommunications, the contribution of this paper is on how to construct the transition blocks.

The remainder of the paper is organized as follows. In Section 2, we define the parameters for the discrete time *MAP/PH/K* queue. In Sections 3 and 4, we develop algorithms for constructing transition probability matrices of the discrete time *MAP/PH/K* queue for the two types of scenarios, respectively. The main contribution of this paper is the algorithm developed in Section 4. Section 5 presents a numerical example to compare the two approaches. Section 6 concludes the paper.

2. Discrete time MAP/PH/K queue

The queueing model under consideration has a single queue and K identical servers. Customers arrive according to a discrete time Markovian arrival process. All customers join a single queue upon arrival. The service discipline is work-conserving (e.g., first-come-first-served, last-come-first-served and non-preemption, random order, etc.) The service times have the same phase-type distribution. The arrival process and service times are defined specifically as follows.

(i) Customers arrive according to discrete time Markovian arrival process (D_0, D_1), where D_0 and D_1 are square matrices of order m_a. Matrices D_0 and D_1 are nonnegative. Let $D = D_0 + D_1$, which is a stochastic matrix (i.e., $D\mathbf{e} = \mathbf{e}$). We assume that D is irreducible. Then D defines an irreducible discrete time Markov chain. Let $I_a(t)$ be the state (phase) of the discrete time Markov chain associated with D, at time t. Then $\{I_a(t), t = 0, 1, 2, \ldots\}$ is an irreducible Markov chain, called the underlying Markov chain. Let $\boldsymbol{\theta}_a$ be the stationary distribution of $\{I_a(t), t = 0, 1, 2, \ldots\}$. Then $\boldsymbol{\theta}_a$ is the unique solution to linear system $\boldsymbol{\theta}_a D = \boldsymbol{\theta}_a$ 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 [9,10].

(ii) All customers join a single queue waiting for service. There are K identical servers. When a server becomes available, a customer in the waiting queue (if there is any) is selected, according to the service discipline, to enter the server for service. If an arriving customer finds an idle server, the customer enters the server for service upon arrival.

(iii) The service time of each customer has a discrete time phase-type distribution with *PH*-representation $(\boldsymbol{\beta}, S)$ of order m_s. We assume that $\boldsymbol{\beta}\mathbf{e} = 1$, i.e., the workload of a customer is always positive. Let $\mathbf{S}^0 = \mathbf{e} - S\mathbf{e}$. We assume that $S + \mathbf{S}^0\boldsymbol{\beta}$ is irreducible, i.e., the *PH*-representation $(\boldsymbol{\beta}, S)$ is *PH*-irreducible. Let $\boldsymbol{\theta}_s = (\theta_{s,1}, \theta_{s,2}, \ldots, \theta_{s,m_s})$ be the row vector satisfying $\boldsymbol{\theta}_s(S + \mathbf{S}^0\boldsymbol{\beta}) = \boldsymbol{\theta}_s$ and $\boldsymbol{\theta}_s\mathbf{e} = 1$. Since the *PH*-representation is irreducible, $\boldsymbol{\theta}_s$ is the unique solution to the linear system. The mean work-load is given by $\boldsymbol{\beta}(I - S)^{-1}\mathbf{e}$. It is well-known that $\boldsymbol{\theta}_s\mathbf{S}^0 = 1/(\boldsymbol{\beta}(I - S)^{-1}\mathbf{e})$, which is called the service rate and is denoted as μ. See [3] for more about phase-type distributions.

In Sections 3 and 4, we introduce two *GI/M/*1 type Markov chains associated with the number of customers in the queueing system. Then we develop methods for constructing transition probability matrices for the Markov chains, respectively. The track-phase-for-server approach in Section 3 is straightforward. However, sizes of the transition blocks increase exponentially in K and m_s. Sizes of the transition blocks obtained in Section 4 using the count-server-for-phase approach are much smaller than that of the blocks in Section 3.

3. The track-phase-for-server (TPFS) approach

We define the following random variables to represent the queueing system. Define

- $q(t)$: the number of customers in the queueing system at time t.
- $I_{s,k}(t)$: the phase of the underlying Markov chain of the service process of server k at time t, if the server is working; otherwise, $I_{s,k}(t) = 0$, for $k = 1, 2, \ldots, K$. We note that the kth server is not referred to a specific physical server. Since all K servers are identical, assigning a different server to be the kth server at a different time does not change the results.

We define a process $\{(q(t), I_a(t), I_{s,1}(t), \ldots, I_{s,\min\{q(t),K\}}(t)),\ t \geq 0\}$. It is easy to see that $\{(q(t), I_a(t), I_{s,1}(t), \ldots, I_{s,\min\{q(t),K\}}(t)),\ t \geq 0\}$ is a discrete time Markov chain. We call $q(t)$ the *level variable* and $(I_a(t), I_{s,1}(t), \ldots, I_{s,\min\{q(t),K\}}(t))$ the *phase variable*. It is also easy to see that $q(t)$ increases at most by one or decreases at most by K at transitions and vector $(I_a(t), I_{s,1}(t), \ldots, I_{s,\min\{q(t),K\}}(t))$ takes a finite number of values. Then $\{(q(t), I_a(t), I_{s,1}(t), \ldots, I_{s,\min\{q(t),K\}}(t)), t \geq 0\}$ is a $GI/M/1$ type Markov chain. The objective of this section is to construct the transition blocks in the transition probability matrix of the Markov chain, which has the following structure:

$$P_{\mathrm{TPFS}} = \begin{pmatrix}
A_{0,0} & A_{0,1} & & & & & & & & & & \\
A_{1,0} & A_{1,1} & A_{1,2} & & & & & & & & & \\
\vdots & \ddots & \ddots & \ddots & & & & & & & & \\
A_{K-1,0} & \cdots & A_{K-1,K-2} & A_{K-1,K-1} & A_{K-1,K} & & & & & & & \\
A_{K,0} & A_{K,1} & \cdots & A_{K,K-1} & A_{K,K} & A_0 & & & & & & \\
 & A_{K+1,1} & A_{K+1,2} & \ddots & A_{K+1,K} & A_1 & A_0 & & & & & \\
 & & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & & & & \\
 & & & A_{2K-1,K-1} & A_{2K-1,K} & \ddots & A_2 & A_1 & A_0 & & & \\
 & & & & A_{2K,K} & A_K & \cdots & A_2 & A_1 & A_0 & & \\
 & & & & & A_{K+1} & A_K & \cdots & A_2 & A_1 & A_0 & \\
 & & & & & & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots
\end{pmatrix}, \tag{1}$$

where $\{A_{k,j},\ k = 0, 1, \ldots, j = 0, 1, \ldots, k+1\}$ and $\{A_0, A_1, \ldots, A_{K+1}\}$ are transition blocks between levels. Next, we develop an algorithm for constructing those matrices. By utilizing the Kronecker-product operation of matrices, the transition blocks can be obtained as follows.

Proposition 3.1. *For the discrete time MAP/PH/K queue, the transition blocks in P_{TPFS} are*

$$\begin{aligned}
&A_{k,0} = D_0 \otimes S_{k,0}, \quad \textit{for } k = 0, 1, 2, \ldots, K;\\
&A_{k,j} = D_0 \otimes S_{k,j} + D_1 \otimes S_{k,j-1} \otimes \boldsymbol{\beta}, \quad \textit{for } k = 1, \ldots, K-1, j = 1, 2, \ldots, k;\\
&A_{k,k+1} = D_1 \otimes S_{k,k} \otimes \boldsymbol{\beta}, \quad \textit{for } k = 0, 1, \ldots, K-1;\\
&A_{k,k-K} = D_0 \otimes S_{k,0} \otimes \boldsymbol{\beta}^{(\otimes(k-K))}, \quad \textit{for } k = K, K+1, \ldots;\\
&A_{k,j} = D_0 \otimes S_{K,K-k+j} \otimes \boldsymbol{\beta}^{(\otimes(k-\max\{j,K\}))} + D_1 \otimes S_{K,K-k+j-1} \otimes \boldsymbol{\beta}^{(\otimes(k-\max\{j,K\}+1))},\\
&\quad \textit{for } k = K, K+1, \ldots, j = k-K+1, k-K+2, \ldots, k;\\
&A_{k,k+1} = D_1 \otimes S_{K,K}, \quad \textit{for } k = K, K+1, \ldots,
\end{aligned} \tag{2}$$

where $S_{k,j}$ is the one step transition matrix beginning with k servers in service and ending with j servers still in service,

$$\begin{aligned}
&S_{0,0} = 1, \qquad S_{k,0} = S_{k-1,0} \otimes \mathbf{S}^0, \quad \textit{for } k = 1, 2, \ldots;\\
&S_{k,j} = S_{k-1,j} \otimes \mathbf{S}^0 + S_{k-1,j-1} \otimes S, \quad \textit{for } k = 1, 2, \ldots, j = 1, 2, \ldots, k-1;\\
&S_{k,k} = S_{k-1,k-1} \otimes S, \quad \textit{for } k = 1,\ 2, \ldots;\\
&(\boldsymbol{\beta})^{(\otimes 0)} = 1, \qquad (\boldsymbol{\beta})^{(\otimes k)} = (\boldsymbol{\beta})^{(\otimes(k-1))} \otimes \boldsymbol{\beta}, \quad \textit{for } k = 1,\ 2, \ldots;
\end{aligned} \tag{3}$$

and

$$\begin{aligned}
&A_0 = D_1 \otimes \hat{S}_{K,K};\\
&A_k = D_0 \otimes \hat{S}_{K,K+1-k} + D_1 \otimes \hat{S}_{K,K-k}, \quad \textit{for } k = 1,\ 2, \ldots,\ K;\\
&A_{K+1} = D_0 \otimes \hat{S}_{K,0},
\end{aligned} \tag{4}$$

where $\hat{S}_{k,j}$ is the one step transition matrix that, beginning with k server in service, (i) $k-j$ *servers complete and restart service immediately, and* (ii) j *servers continue their current service,*

$$\begin{aligned}
&\hat{S}_{0,0}=1, \qquad \hat{S}_{k,0}=\hat{S}_{k-1,0}\otimes(\mathbf{S}^0\boldsymbol{\beta}), \quad \text{for } k=1,\ 2,\ \ldots,\ K;\\
&\hat{S}_{k,j}=\hat{S}_{k-1,j}\otimes(\mathbf{S}^0\boldsymbol{\beta})+\hat{S}_{k-1,j-1}\otimes S, \quad \text{for } k=1,2,\ldots,K,\ j=1,\ 2,\ \ldots,\ k-1;\\
&\hat{S}_{k,k}=\hat{S}_{k-1,k-1}\otimes S, \quad \text{for } k=1,2,\ldots,\ K.
\end{aligned} \tag{5}$$

In addition, the states in level q are given by $\{(q, i_a, i_{s,1}, \ldots, i_{s,\min\{q,K\}}) : i_a = 1, 2, \ldots, m_a, i_{s,k} = 1, 2, \ldots, m_s, k = 1, 2, \ldots, \min\{q, K\}\}$, *for* $q \geq 0$. *The size of matrices* $\{A_0, A_1, \ldots, A_{K+1}\}$ *is* $m_a m_s^K$. □

Note 3.1. Matrices $\{A_0, A_1, \ldots, A_{K+1}\}$ can be defined as $A_0 = A_{k,k+1}, A_1 = A_{k,k}, \ldots, A_{K+1} = A_{k,k-K}$, for any $k \geq 2K$. The interpretations of the elements of the matrices are different, but the analysis of the Markov chain and the queueing model is similar. We use the construction given in Eqs. (4) and (5) for convenience.

Let $A = A_0 + A_1 + \cdots + A_{K+1}$.

Proposition 3.2. *The matrix A is an irreducible transition probability matrix, i.e.,* $A\mathbf{e} = \mathbf{e}$, *and its stationary distribution is given by* $\boldsymbol{\theta}_a \otimes (\boldsymbol{\theta}_s)^{(\otimes K)}$, *where* $(\boldsymbol{\theta}_s)^{(\otimes K)}$ *is the Kronecker product of K vector* $\boldsymbol{\theta}_s$. □

The property is useful for verifying computation programs. For the *GI/M/*1 type Markov chain, its stationary distribution has a matrix-geometric solution (see [3]). In Section 5, we shall use the Markov chain P_{TPFS} to compute the mean queue length, which is useful in verifying the Markov chain to be constructed in Section 4.

Proposition 3.3. *The Markov chain* P_{TPFS} *is ergodic if and only if* $\lambda < K\mu$.

Proof. From queueing point of view, the condition is intuitive. We present a technical proof. By [3], the irreducible *GI/M/*1 type Markov chain is positive recurrent if and only if

$$\sum_{k=1}^{K+1}(k-1)\boldsymbol{\theta}_a\otimes(\boldsymbol{\theta}_s)^{(\otimes K)}A_k\mathbf{e} > \boldsymbol{\theta}_a\otimes(\boldsymbol{\theta}_s)^{(\otimes K)}A_0\mathbf{e}, \tag{6}$$

which is equivalent to $\sum_{k=0}^{K+1} k\boldsymbol{\theta}_a \otimes (\boldsymbol{\theta}_s)^{(\otimes K)}A_k\mathbf{e} > 1$. By routine calculations, we obtain

$$\begin{aligned}
\sum_{k=1}^{K+1}k\boldsymbol{\theta}_a\otimes(\boldsymbol{\theta}_s)^{(\otimes K)}A_k\mathbf{e} &= \boldsymbol{\theta}_a\otimes(\boldsymbol{\theta}_s)^{(\otimes K)}\left(D_0\otimes\sum_{k=0}^{K}\hat{S}_{K,k}+D\otimes\sum_{k=0}^{K}(K-k)\hat{S}_{K,k}\right)\mathbf{e}\\
&= (\boldsymbol{\theta}_a D_0\mathbf{e})\left((\boldsymbol{\theta}_s)^{(\otimes K)}\left(\sum_{k=0}^{K}\hat{S}_{K,k}\right)\mathbf{e}\right)+(\boldsymbol{\theta}_a D\mathbf{e})\left(\sum_{k=0}^{K}(K-k)\left((\boldsymbol{\theta}_s)^{(\otimes K)}\hat{S}_{K,k}\mathbf{e}\right)\right)\\
&= 1-\lambda+K\mu\left(\sum_{k=0}^{K-1}\binom{K-1}{k}(\boldsymbol{\theta}_s S\mathbf{e})^k(\boldsymbol{\theta}_s\mathbf{S}^0)^{K-1-k}\right)\\
&= 1-\lambda+K\mu\left(\boldsymbol{\theta}_s S\mathbf{e}+\boldsymbol{\theta}_s\mathbf{S}^0\right)^{K-1}=1-\lambda+K\mu,
\end{aligned} \tag{7}$$

where we have used $\lambda = \boldsymbol{\theta}_a D_1\mathbf{e} = 1-\boldsymbol{\theta}_a D_0\mathbf{e}$, $(\boldsymbol{\theta}_s)^{(\otimes K)}\hat{S}_{K,k}\mathbf{e} = \binom{K}{k}(\boldsymbol{\theta}_s S\mathbf{e})^k(\boldsymbol{\theta}_s\mathbf{S}^0)^{K-k}$, and $\mu = \boldsymbol{\theta}_s\mathbf{S}^0$. The proof is completed. □

Define $A^*(z) = A_0 + zA_1 + \cdots + z^{K+1}A_{K+1}$, $D^*(z) = D_1 + zD_0$, and $S^*(z) = S + zS^0\boldsymbol{\beta}$. For $z > 0$, let $\rho_A(z)$, $\rho_D(z)$, and $\rho_S(z)$ be the Perron–Frobenius eigenvalue of $A^*(z)$, $D^*(z)$, and $S^*(z)$ (i.e., the eigenvalue with the largest real part), respectively.

Proposition 3.4. $A^*(z) = D^*(z)\otimes(S^*(z))^{(\otimes K)}$ *and* $\rho_A(z) = \rho_D(z)(\rho_S(z))^K$.

Proof. The first result is obtained as follows:

$$\begin{aligned}
A^*(z) &= D_1\otimes\hat{S}_{K,K}+z(D_0\otimes\hat{S}_{K,K}+D_1\otimes\hat{S}_{K,K-1})+\cdots+z^K D_0\otimes\hat{S}_{K,0}\\
&= (D_1+zD_0)\otimes\left(\hat{S}_{K,K}+z\hat{S}_{K,K-1}+z^2\hat{S}_{K,K-2}+\cdots+z^K\hat{S}_{K,0}\right)\\
&= (D_1+zD_0)\otimes\left(\hat{S}_{K-1,K-1}+z\hat{S}_{K-1,K-2}+\cdots+z^{K-1}\hat{S}_{K-1,0}\right)\otimes(S+z\mathbf{S}^0\boldsymbol{\beta})\\
&= (D_1+zD_0)\otimes\left(S+z\mathbf{S}^0\boldsymbol{\beta}\right)^{(\otimes K)}.
\end{aligned} \tag{8}$$

The second result is obtained from the first one. □

4. The count-server-for-phase (CSFP) approach

In Section 3, the service process is represented by vector $(I_{s,1}(t), \ldots, I_{s,\min\{q(t),K\}}(t))$. In this section, we define another random vector to represent the service process. Define

- $n_i(t)$: the number of servers whose service phase is i at time t, for $i = 1, 2, \ldots, m_s$.

Since the service processes of servers can be represented by the same underlying discrete time Markov chain, vector $(n_1(t), \ldots, n_{m_s}(t))$ provides all information about the service process. Thus, the queueing process can be represented by the *GI/M/*1 type Markov chain $\{(q(t), I_a(t), n_1(t), \ldots, n_{m_s}(t)), t \geq 0\}$. Next, we (i) characterize the state space of the Markov chain; (ii) find its transition probability matrix; and (iii) find some stationary distribution related to the Markov chain.

Let $\Omega(q, m_s)$ be the set of states of $(n_1(t), \ldots, n_{m_s}(t))$, given that $q(t) = q \geq 0$, which is called the level q. It is easy to see that

$$\Omega(q, m_s) = \left\{ (n_1, \ldots, n_{m_s}) : \ n_i \geq 0, \ i = 1, \ 2, \ \ldots, \ m_s, \ \sum_{i=1}^{m_s} n_i = \min\{q, K\} \right\}. \tag{9}$$

Based on the number of servers whose service phase is m_s, which can be 0, 1, ..., and q, the states in $\Omega(q, m_s)$ are arranged into $q + 1$ subsets as follows:

$$\Omega(q, m_s) = \{\Omega(q, m_s - 1) \times \{0\}\} \cup \{\Omega(q - 1, m_s - 1) \times \{1\}\} \cup \cdots \cup \{\Omega(0, m_s - 1) \times \{q\}\}. \tag{10}$$

Then the state space of $\{(q(t), I_a(t), n_1(t), \ldots, n_{m_s}(t)), t \geq 0\}$ can be obtained as

$$\bigcup_{q=0}^{\infty} (\{q\} \times \{1, \ 2, \ \ldots, \ m_a\} \times \Omega(q, m_s)). \tag{11}$$

The number of states in level $q \geq K$ is

$$m_a \binom{K + m_s - 1}{m_s - 1} = m_a \frac{(K + m_s - 1)!}{K!(m_s - 1)!}. \tag{12}$$

The transition probability matrix of $\{(q(t), I_a(t), n_1(t), \ldots, n_{m_s}(t)), t \geq 0\}$, denoted as P_{CSFP}, has exactly the same structure as that of P_{TPFS} (see Eq. (1)) with transition blocks given as follows. Unlike the continuous time case, phase transitions can occur simultaneously for the discrete time case. Thus, the construction of P_{CSFP} is more involved than that of the continuous time case. We begin our construction process with some observations on the state transition process. (Note: *phase* is for $m = 1, 2, \ldots, m_s$; *state* is for $\{(q(t), I_a(t), n_1(t), \ldots, n_{m_s}(t)), t \geq 0\}$).

- **Observation 1.** The one-step phase transitions of services in individual servers are independent.
- **Observation 2.** The one-step phase transitions of service completions in individual servers are independent.
- **Observation 3.** The one-step phase transitions of the service processes and the arrival process are independent.

We defined the following matrices, for $q, m, j, k \geq 0$,

- $P_{\mathbf{u},\mathbf{v}}\{q, j, m|k\} = P_{\mathbf{u},\mathbf{v}}\{\Omega(q, m) : \Omega(j, m)|k\}$: The one-step transition matrix from the set $\Omega(q, m)$ to $\Omega(j, m)$, given that there are exactly k service completions, the transitions within the m phases are governed by $S_{[1:m,1:m]}$, and the initial phases of the new services are determined by probabilities in vector $\mathbf{u}$ of size m or larger, and service completion is determined by probabilities in vector $\mathbf{v}$ of size m or larger. (Note that the number of new services is $j + k - q$.)

First, we construct $A_{k,j}$ in P_{CSFP} from $\{P_{\boldsymbol{\beta},\mathbf{S}^0}\{q, j, m_s|k\}, D_0, D_1\}$.

Proposition 4.1. *The transition probability blocks in P_{CSFP} can be obtained as*

$$\begin{aligned}
&(1)\ A_{k,k+1} = D_1 \otimes P_{\boldsymbol{\beta},\mathbf{S}^0}\{k, k+1, m_s|0\}, \quad \text{for } k \leq K-1;\\
&(2)\ A_{k,k+1} = A_0 = D_1 \otimes P_{\boldsymbol{\beta},\mathbf{S}^0}\{K, K, m_s|0\}, \quad \text{for } k \geq K;\\
&(3)\ A_{k,0} = D_0 \otimes P_{\boldsymbol{\beta},\mathbf{S}^0}\{k, 0, m_s|k\}, \quad \text{for } k \leq K;\\
&(4)\ A_{k,k-K} = D_0 \otimes P_{\beta,\mathbf{S}^0}\{K, k-K, m_s|K\}, \quad \text{for } K+1 \leq k \leq 2K-1;\\
&(5)\ A_{k,k-K} = A_{K+1} = D_0 \otimes P_{\boldsymbol{\beta},\mathbf{S}^0}\{K, K, m_s|K\}, \quad \text{for } k \geq 2K.\\
&(6)\ A_{k,j} = D_0 \otimes P_{\boldsymbol{\beta},\mathbf{S}^0}\{k, j, m_s|k-j\} + D_1 \otimes P_{\boldsymbol{\beta},\mathbf{S}^0}\{k, j, m_s|k-j+1\}, \quad \text{for } k \leq K, \ 1 \leq j \leq k;\\
&(7)\ A_{k,j} = D_0 \otimes P_{\boldsymbol{\beta},\mathbf{S}^0}\{K, \min\{j, K\}, m_s|k-j\} + D_1 \otimes P_{\boldsymbol{\beta},\mathbf{S}^0}\{k, \min\{j, K\}, m_s|k-j+1\},\\
&\quad \text{for } K+1 \leq k \leq 2K-1, \ k-K+1 \leq j \leq k;\\
&(8)\ A_{k,j} = A_{k-j+1} = D_0 \otimes P_{\boldsymbol{\beta},\mathbf{S}^0}\{K, K, m_s|k-j\} + D_1 \otimes P_{\boldsymbol{\beta},\mathbf{S}^0}\{K, K, m_s|k-j+1\},\\
&\quad \text{for } 2K \leq k, \ k-K+1 \leq j \leq k.
\end{aligned} \tag{13}$$

Proof. All expressions are obtained easily by definitions. □

To compute matrix $P_{\boldsymbol{\beta},\mathbf{S}^0}\{q, j, m_s|k\}$, based on Observations 1–3, we decompose changes of states into three categories: (i) the service phase of a server entering the set $\{1, 2, \ldots, m\}$ or a new service is initialized; (ii) phase transitions within $\{1, 2, \ldots, m\}$ (or no new service initialization and no service completion); and (iii) a service phase leaving the set $\{1, 2, \ldots, m\}$ or a service completion. For the three types of transitions, we define the following matrices:

- $L^+_{\mathbf{u}}\{q,\ q+j,\ m\} = L^+_{\mathbf{u}}\{\Omega(q, m) :\ \Omega(q+j,\ m)\}$: The one-step transition matrix from the set $\Omega(q, m)$ to $\Omega(q+j, m)$ only due to the initialization of the service of j customers in phases $\{1, 2, \ldots, m\}$, given that the initial phase of the j new customers are determined by probabilities in row vector $\mathbf{u}$ of size m or larger.
- $P\{q, m\} = P\{\Omega(q, m) :\ \Omega(q, m)\}$: The one-step transition matrix from the set $\Omega(q, m)$ to $\Omega(q, m)$, given that the transitions within the m phases are governed by $S_{[1:m,\,1:m]}$. (*Note*: that there is no transition into or going out of $\Omega(q, m)$. Only phase changes within $\{1, 2, \ldots, m\}$.)
- $L^-_{\mathbf{v}}\{q+j,\ q,\ m\} = L^-_{\mathbf{v}}\{\Omega(q+j, m) :\ \Omega(q, m)\}$: The one-step transition matrix from the set $\Omega(q+j, m)$ to $\Omega(q, m)$ only due to the transitions of the service phases of j customers out of phases $\{1, 2, \ldots, m\}$, given that the out-going probabilities of j customers are determined by probabilities in column vector $\mathbf{v}$ of size m or larger (*Note*: that no other type of phase change is considered.)

Each of matrices $L^+_{\mathbf{u}}\{q, q+j, m\}, P\{q, m\}$, and $L^-_{\mathbf{v}}\{q+j,\ q,\ m\}$ is defined specifically for one type of transitions. Thus, their components may not be transition probabilities. Nonetheless, by putting them together properly, the one-step transition matrix $P_{\boldsymbol{\beta},\mathbf{S}^0}\{q, j, m_s|k\}$ is obtained from $\{L^+_{\mathbf{u}}\{q,\ q+j,\ m\}, P\{q, m_s\}, L^-_{\mathbf{v}}\{q+j,\ q,\ m\}\}$. Before we present the results, we have a look at an example.

Consider a binomial distribution with parameters $\{n, a\}$. Suppose that a is the probability to leave the set $\{1, 2, \ldots, m\}$ in one transition. Then the probability that k customers leave the set $\{1, 2, \ldots, m\}$ (and $n-k$ stay within) is given by $a^k(1-a)^{n-k}n!/(k!(n-k)!)$, which can be written as the product of $\{na, (n-1)a, \ldots, (n-k+1)a\}$ and $\{1/k!, (1-a)^{n-k}/(n-k)!\}$. Intuitively, the decomposition can be explained by associating $\{na, (n-1)a, \ldots, (n-k+1)a\}$ with the one step transitions of k out of n customers leaving the set $\{1, 2, \ldots, m\}$, and $(1-a)^{n-k}/(n-k)!$ with all one step transitions of the other $n-k$ customers remaining in the set $\{1, 2, \ldots, m\}$.

Proposition 4.2. *For given* $\mathbf{u}$ *and* $\mathbf{v}$*, the following relationships hold among the matrices defined above:*

$$
\begin{aligned}
&(1)\ P_{\mathbf{u},\mathbf{v}}\{q, q, m|0\} = P\{q, m\}, \quad \text{for } q = 1,\ 2,\ \ldots,\ K;\\
&(2)\ P_{\mathbf{u},\mathbf{v}}\{q, j, m|k\} = L^-_{\mathbf{v}}\{q, q-k, m\}P\{q-k, m\}L^+_{\mathbf{u}}\{q-k, j, m\}, \quad \text{for } k \le q \le k+j;\\
&(3)\ L^-_{\mathbf{v}}\{q+k, q, m\} = \frac{1}{k!}\prod_{j=q+k}^{q+1} L^-_{\mathbf{v}}\{j, j-1, m\}, \quad \text{for } k,\ q \ge 0;\\
&(4)\ L^+_{\mathbf{u}}\{q, q+k, m\} = \prod_{j=q}^{q+k-1} L^+_{\mathbf{u}}\{j, j+1, m\}, \quad \text{for } k,\ q \ge 0.
\end{aligned}
\tag{14}
$$

Proof. Parts (1), (2), and (4) are obtained by definitions. Part (3) is also obtained by definition, plus the fact that the k leaving customers (i.e., leaving the set $\{1, 2, \ldots, m\}$) are selected from $q+k$ customers. Since the order of the k leaving customers does not affect the probabilities, we must have the factor $1/k!$ in part (3). (Note: For part (4), the k new customers are not selected from any set. Therefore, the factor $1/k!$ does not appear in part (4).) □

Next, we construct $\{L^+_{\mathbf{u}}\{q,\ q+1,\ m\}, L^-_{\mathbf{v}}\{q+1,\ q,\ m\}\}$ from parameters $\{\mathbf{u}, \mathbf{v}\}$.

Proposition 4.3. *For given* $\mathbf{u} = (u_1, \ldots, u_m)$ *and* $\mathbf{v} = (v_1, \ldots, v_m)'$*, the matrix* $L^+_{\mathbf{u}}\{\Omega(k, m) :\ \Omega(k+1, m)\}$ *and* $L^-_{\mathbf{v}}\{\Omega(k+1, m) :\ \Omega(k, m)\}$ *can be obtained as*

$$
L^+_{\mathbf{u}}\{k, k+1, m\} = \begin{array}{c} \\ \Omega(k, m-1)\times\{0\} \\ \Omega(k-1, m-1)\times\{1\} \\ \vdots \\ \Omega(1, m-1)\times\{k-1\} \\ \Omega(0, m-1)\times\{k\} \end{array}
\begin{array}{c}
\begin{array}{ccc} \Omega(k+1, m-1)\times\{0\} & \cdots & \Omega(0, m-1)\times\{k+1\} \end{array} \\
\begin{pmatrix}
L^+_{\mathbf{u}}\{k, k+1, m-1\}\ u_m I & & \\
\ddots\ u_m I & & \\
\ddots\ \ddots & & \\
\ddots & u_m I & \\
 & L^+_{\mathbf{u}}\{0, 1, m-1\}\ u_m
\end{pmatrix}
\end{array};
\tag{15}
$$

and

$$L_{\mathbf{v}}^{-}\{k,k-1,m\} = \begin{array}{c} \\ \Omega(k,m-1)\times\{0\} \\ \Omega(k-1,m-1)\times\{1\} \\ \vdots \\ \vdots \\ \Omega(1,m-1)\times\{k-1\} \\ \Omega(0,m-1)\times\{k\} \end{array} \begin{array}{c} \begin{array}{cccc} \Omega(k-1,m-1)\times\{0\} & \cdots & & \Omega(0,m-1)\times\{k-1\} \end{array} \\ \begin{pmatrix} L_{\mathbf{v}}^{-}\{k,k-1,m-1\} & & & \\ v_m I & \ddots & & \\ & \ddots & \ddots & \\ & & \ddots & \ddots \\ & & & (k-1)v_m I \quad L_{\mathbf{v}}^{-}\{1,0,m-1\} \\ & & & \qquad\qquad k v_m \end{pmatrix} \end{array}; \tag{16}$$

and

$$\begin{aligned} &L_{\mathbf{u}}^{+}\{0,1,m\} = \mathbf{u}_{[1:m]}; \qquad L_{\mathbf{u}}^{+}\{k,k+1,1\} = u_1; \\ &L_{\mathbf{v}}^{-}\{1,0,m\} = \mathbf{v}_{[1:m]}; \qquad L_{\mathbf{v}}^{-}\{k+1,k,1\} = (k+1)v_1. \end{aligned} \tag{17}$$

Proof. All results are obtained by definition. Note that the size of vector $\mathbf{u}$ or $\mathbf{v}$ can be greater than m. Once m is given, we only need the first m elements of vectors $\mathbf{u}$ and $\mathbf{v}$ (i.e., $\mathbf{u}[1:m]$ and $\mathbf{v}[1:m]$) in the construction of the matrices. □

Finally, we find $P\{k,m\}$ recursively from system parameter S.

Proposition 4.4.

$$\begin{aligned} P\{k,m\} &= P\{k,k,m|0\} \\ &= \begin{array}{c} \\ \\ \Omega(j,m-1)\times\{k-j\} \\ \\ \\ \end{array} \begin{array}{c} \begin{array}{ccc} \cdots & \Omega(q,m-1)\times\{k-q\} & \cdots \end{array} \\ \begin{pmatrix} \vdots & \cdots & \vdots \\ \cdots\ P_{S_{[m,1:m-1]},S_{[1:m-1,m]}}\{\Omega(j,m-1)\times\{k-j\} : \Omega(q,m-1)\times\{k-q\}\} & \cdots \\ \vdots & \cdots & \vdots \end{pmatrix} \end{array} \end{aligned} \tag{18}$$

where $1 \le j, q \le k$,

$$\begin{aligned} &P_{S_{[m,1:m-1]},S_{[1:m-1,m]}}\{\Omega(j,m-1)\times\{k-j\} : \Omega(q,m-1)\times\{k-q\}\} \\ &= \sum_{l=\max\{0,j-q\}}^{\min\{j,k-q\}} P_{S_{[m,1:m-1]},S_{[1:m-1,m]}}\{j,q,m-1|l\} \binom{k-j}{k-q-l} (s_{m,m})^{k-q-l} \\ &= \begin{cases} \sum_{l=0}^{\min\{j,k-q\}} P_{S_{[m,1:m-1]},S_{[1:m-1,m]}}\{j,q,m-1|l\} \binom{k-j}{k-q-l} (s_{m,m})^{k-q-l}, & \text{if } j \le q; \\ \sum_{l=j-q}^{\min\{j,k-q\}} P_{S_{[m,1:m-1]},S_{[1:m-1,m]}}\{j,q,m-1|l\} \binom{k-j}{k-q-l} (s_{m,m})^{k-q-l}, & \text{if } j > q \end{cases} \end{aligned} \tag{19}$$

and

$$\begin{aligned} &P\{0,\ m\} = P\{\Omega(0,m) : \Omega(0,m)\} = 1; \\ &P\{1,\ m\} = P\{\Omega(1,m) : \Omega(1,m)\} = S_{[1:m,1:m]}; \\ &P\{k,\ 1\} = P\{\Omega(k,1) : \Omega(k,1)\} = s_{1,1}^{k}. \end{aligned} \tag{20}$$

Note: Elements of S are denoted as $s_{i,j}$, i.e., $S = (s_{i,j})$. Matrix $S_{[m,1:m-1]}$ consists of elements in the mth row and 1 to $m-1$ columns of S. Matrices $S_{[1:m-1,\,m]}$ and $S_{[1:m,1:m]}$ are defined similarly.

Proof. Although there is no customer entering or leaving the set $\{1,2,\ldots,m\}$, there can be transitions between the phases themselves. We first consider the transitions between $\{1,2,\ldots,m-1\}$ and $\{m\}$. In this manner, the problem becomes solving problems within the subset $\{1,2,\ldots,m-1\}$, which leads to the recursive formulas. In this case, among $k-j$ customers who are originally in phase m, $k-q-l$ customers, transit to phases in $\{1,2,\ldots,m-1\}$. The number of selections of the $k-q-l$ customers is $k-q-l$ out of $k-j$. The rest of the proof is straightforward. □

Recall that $A = A_0 + A_1 + \cdots + A_{K+1}$. Next, we find the stationary distribution of A. Define vector $\boldsymbol{\phi}$ as follows:

$$\phi(\mathbf{n}) = \frac{K!}{n_1! \cdots n_{m_s}!} \prod_{j=1}^{m_s} \theta_{s,j}^{n_j}, \quad \text{for } \mathbf{n} = (n_1, \ \ldots, \ n_{m_s}) \in \Omega(K, m_s). \tag{21}$$

The vector $\boldsymbol{\phi}$ is a probability vector (i.e., $\boldsymbol{\phi} \geq 0$ and $\boldsymbol{\phi}\mathbf{e} = 1$) since it can be considered as the probability mass function of a multinomial distribution, i.e.,

$$\sum_{\mathbf{n} \in \Omega(K, m_s)} \phi(\mathbf{n}) = \sum_{\mathbf{n} \in \Omega(K, m_s)} \frac{K!}{n_1! \cdots n_{m_s}!} \prod_{j=1}^{m_s} \theta_{s,j}^{n_j} = \left(\sum_{j=1}^{m_s} \theta_{s,j} \right)^K = 1. \tag{22}$$

The vector $\boldsymbol{\phi}$ is can be constructed as follows:

(i) $\phi(0, m) = 1$, for $m = 1, 2, \ldots, m_s$, and $\phi(k, 1) = \theta_{s,1}^k / k!$, for $k = 0, 1, 2, \ldots, K$;
(ii) $\phi(k, m) = (\phi(k, m-1), \phi(k-1, m-1)\theta_{s,m}, \phi(k-2, m-1)\theta_{s,m}^2/2!, \ldots, \phi(0, m-1)\theta_{s,m}^k/k!)$, for $m = 1, 2, \ldots, m_s$, for $k = 1, 2, \ldots, K$; and
(iii) $\boldsymbol{\phi} = K!\phi(K, m_s)$.

In computation, we use $\omega(k, m) = k!\phi(k, m)$ in the above procedure to improve accuracy.

Proposition 4.5. *Matrix* $A = D \otimes \left(\sum_{k=0}^{K} P_{\boldsymbol{\beta}, \mathbf{S}^0}\{K, K, m_s | k\} \right)$, *which is an irreducible transition probability matrix (stochastic matrix), i.e.,* $A\mathbf{e} = \mathbf{e}$, *and its stationary distribution is given by* $\boldsymbol{\theta}_a \otimes \boldsymbol{\phi}$.

Proof. Since the *PH*-representation of the service workload is irreducible, the transition probability matrix is also irreducible. In steady state, since the probability that the service phase of a server is j is $\theta_{s,j}$, the probability that the service state is $\mathbf{n} \in \Omega(K, m_s)$ is given by $\phi(\mathbf{n})$. Thus, $\boldsymbol{\phi}$ is the stationary distribution of the Markov chain associated with service process of the K servers, assuming the servers are working all the time. Then it is clear that $\boldsymbol{\theta}_a \otimes \boldsymbol{\phi}$ is the stationary distribution of A. □

Proposition 4.6. *The Markov chain* P_{CSFP} *is ergodic if and only if* $\lambda < K\mu$.

Proof. Similar to the proof of Proposition 3.3, we obtain

$$\sum_{k=1}^{K+1} k\boldsymbol{\theta}_a \otimes (\boldsymbol{\theta}_s)^{(\otimes K)} A_k \mathbf{e} = 1 - \lambda + \boldsymbol{\phi} \sum_{k=1}^{K} k P_{\boldsymbol{\beta}, \mathbf{S}^0}\{K, K, m_s | k\}\mathbf{e}. \tag{23}$$

The last part in Eq. (23) is the mean number of customers served by the K servers per unit time, which is $K\mu$. □

5. Numerical examples and discussion

While the steps for the computation of transition blocks in P_{TPFS} is straightforward (see Proposition 3.1), they are more involved for P_{CSFP}. We outline the steps for P_{CSFP} as follows.

1. Based on Propositions 4.2–4.4, construct $P\{k, m\}$.
   1.1 Start from Proposition 4.4.
   1.2 For each pair ($\mathbf{u} = S_{[m,1:m-1]}, \mathbf{v} = S_{[1:m,1:m]}$), use Proposition 4.3 to construct $L_{\mathbf{u}}^+\{k, \ k+1, \ m\}$ and $L_{\mathbf{v}}^-\{k, \ k-1, \ m\}$;
   1.3 For each pair ($\mathbf{u} = S_{[m,1:m-1]}, \mathbf{v} = S_{[1:m,1:m]}$), use Proposition 4.2 and transition blocks obtained in step (1.2) to construct $P_{\mathbf{u},\mathbf{v}}\{q, j, m \mid k\}$, $L_{\mathbf{u}}^+\{q, \ q+k, \ m\}$ and $L_{\mathbf{v}}^-\{q+k, \ q, \ m\}$;
   1.4 Go back to Proposition 4.4 to complete $P\{k, m\}$.
2. Based on Proposition 4.1, construct transition block $A_{k,j}$.
   2.1 Start from Proposition 4.1. Choose $\mathbf{u} = \boldsymbol{\beta}$ and $\mathbf{v} = S^0$.
   2.2 Use Proposition 4.3 to construct $L_{\mathbf{u}}^+\{k, \ k+1, \ m\}$ and $L_{\mathbf{v}}^-\{k, \ k-1, \ m\}$;
   2.3 Use Proposition 4.2 to construct $P_{\mathbf{u},\mathbf{v}}\{q, j, m \mid k\}$, $L_{\mathbf{u}}^+\{q, \ q+k, \ m\}$ and $L_{\mathbf{v}}^-\{q+k, \ q, \ m\}$; Note that $P_{\mathbf{u},\mathbf{v}}\{q, q, m \mid 0\} = P\{q, m\}$ has been obtained from Step (1).
   2.4 Go back to Proposition 4.1 to complete $A_{k,j}$.

We consider an *MAP/PH/K* queue with following parameters:

$$\begin{aligned} &m_a = 2, \qquad D_0 = \begin{pmatrix} 0.2 & 0.3 \\ 0.1 & 0.4 \end{pmatrix}, \qquad D_1 = \begin{pmatrix} 0.4 & 0.1 \\ 0.5 & 0.0 \end{pmatrix}; \\ &m_s = 2, \qquad \boldsymbol{\beta} = (0.1, \ 0.9), \qquad S = \begin{pmatrix} 0.2 & 0.2 \\ 0.1 & 0.7 \end{pmatrix}. \end{aligned} \tag{24}$$

Table 1
Size of transition blocks $A_0, A_1, \ldots, A_{K+1}$ and the mean queue length.

K	2	4	6	8	10	15	20	30
TPFS	8	32	128	512	2048	65,536	2097,152	2147,483,648
CSFP	6	10	14	18	22	32	42	62
$E[q]$	26.757	1.9968	1.9550	1.9545	1.9545	1.9545	1.9545	1.9545

Using the algorithms developed in Sections 3 and 4, we can construct the transition probability matrices P_{TPFS} and P_{CSFP}. The size of transition blocks $\{A_0, A_1, \ldots, A_{K+1}\}$, as a function of the number of servers K, is given in Table 1 (see rows 2 and 3 in Table 1). The distribution of the queue length can be found by using the matrix-geometric solution for the stationary distribution of the $GI/M/1$ type Markov chains P_{TPFS} and P_{CSFP} (see [3]). We also present the mean queue length $E[q]$, as a function of K, in Table 1.

It is clear that the CSFP approach is significantly better than the TPFS approach with respect to the size of transition blocks. Therefore, the extra effort in the construction of P_{CSFP} makes it possible to analyze the queueing system by matrix-analytic methods even if K is not small.

The mean queue length $E[q]$ can be obtained from the matrix-geometric solution of the $GI/M/1$ type Markov chains. First, we find the rate matrix R that is the minimal nonnegative solution to

$$R = \sum_{k=0}^{K+1} R^k A_k. \tag{25}$$

Then, iteratively, we compute $\{R_K, R_{K-1}, \ldots, R_1\}$ as follows: Let $R_{K+k} = R$, for $k = 1, 2, \ldots, K$, and

$$R_k = A_{k-1,k}\left(I - A_{k,k} - \sum_{j=1}^{K}\left(\prod_{t=1}^{j} R_{k+t}\right)A_{k+j,k}\right)^{-1}, \quad \text{for } k = K,\ K-1,\ \ldots,\ 2,\ 1. \tag{26}$$

Denote by $\boldsymbol{\pi} = (\boldsymbol{\pi}_0, \boldsymbol{\pi}_1, \boldsymbol{\pi}_2, \ldots)$ the stationary distribution of P_{TPFS} or P_{CSFP}, which is partitioned according to the level variable $q(t)$. It is well-known that $\boldsymbol{\pi}$ has matrix-geometric solution [3]: $\boldsymbol{\pi}_0, \boldsymbol{\pi}_1 = \boldsymbol{\pi}_0 R_1, \ldots, \boldsymbol{\pi}_K = \boldsymbol{\pi}_{K-1} R_K$, and $\boldsymbol{\pi}_n = \boldsymbol{\pi}_K R^{n-K}$, for $n = K+1, \ldots$, where $\boldsymbol{\pi}_0$ can be obtained by solving the following linear system:

$$\begin{aligned} &\boldsymbol{\pi}_0 = \boldsymbol{\pi}_0\left(A_{0,0} + \sum_{j=1}^{K}\left(\prod_{t=1}^{j} R_t\right)A_{j,0}\right); \\ &\boldsymbol{\pi}_0\left(\mathbf{e} + \sum_{j=1}^{K-1}\left(\prod_{t=1}^{K} R_t\right)\mathbf{e} + \left(\prod_{t=1}^{j} R_t\right)(I-R)^{-1}\mathbf{e}\right) = 1. \end{aligned} \tag{27}$$

Then the mean queue length can be found as

$$E[q] = \boldsymbol{\pi}_0\left(\sum_{j=1}^{K-1} j\left(\prod_{t=1}^{j} R_t\right)\right)\mathbf{e} + \boldsymbol{\pi}_0\left(\prod_{t=1}^{K} R_t\right)(K(I-R)+R)(I-R)^{-2}\mathbf{e}. \tag{28}$$

6. Conclusions

It is clear that using CSFP is more efficient than using TPFS when it comes to computing the matrices R or G, the queue length and all associated measures, especially when the dimensions of the *PH* distributions for service, dimension for the *MAP*, and the number of servers are not small. When those numbers are small the difference in size of the matrices and hence computational efficiency is not a major concern. However one will also notice that the block matrices may be much easier to construct for the TPFS. But not only that, when it comes to computing the decay rate for studying the tail distribution of queue length or waiting time, there is still some advantage to using the TPFS approach. This is because the decay rate is simply the Perron–Frobenius eigenvalue of matrix R. By Eq. (25) and Proposition 3.4, it can be shown that the Perron–Frobenius eigenvalue of R is the unique solution of $z = \rho_D(z)(\rho_S(z))^K$ in $(0, 1)$, if $\lambda < K\mu$. On the other hand, the CSFP approach does not lead to such an explicit and simple equation for the decay rate.

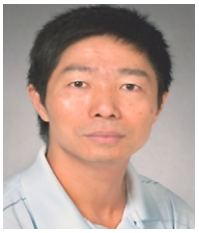
Acknowledgments

We would like to thank three reviewers for their valuable comments and suggestions on the paper. This study was funded by NSERC (RGPIN-203546-2012), and partially by NSERC (RGPIN-6584-2013) and NSERC (RGPAS-446147-2013).

References

[1] W. Yue, Y. Matsumoto, Performance Analysis of Multichannel and Multi-Traffic on Wireless Communication Networks, Kluwer Academic Pub., 2002.
[2] A.S. Alfa, Discrete Time Queues for Telecommunications: A Single Node Case, Springer, New York, 2010.

[3] M.F. Neuts, Matrix-Geometric Solution in Stochastic Model: An Algorithmic Application, The Johns Hopkins University Press, Baltimore, MD, 1981.
[4] G. Latouche, V. Ramaswami, Introduction to Matrix Analytic Methods in Stochastic Modeling, in: ASA-SIAM Series on Statistic and Applied Probability., SIAM, Philadelphia, PA, 1999.
[5] V. Ramaswami, D.M. Lucantoni, Algorithms for the multi-server queue with phase type service, Stoch. Models 1 (1985) 393–417.
[6] S. Asmussen, C.A. O'Cinneide, Representations for matrix-geometric and matrix-exponential steady-state distributions with applications to many-server queues, Stoch. Models 14 (1998) 369–387.
[7] V. Ramaswami, Independent Markov process in parallel, Stoch. Models 1 (1985) 419–432.
[8] Q.M. He, H. Zhang, Q.Q. Ye, An *M*/*PH*/*K* queue with constant impatient time, 2014, submitted for publication.
[9] M.F. Neuts, A versatile Markovian point process, J. Appl. Probab. 16 (1979) 764–779.
[10] D.M. Lucantoni, New results on the single server queue with a batch Markovian arrival process, Stoch. Models 7 (1991) 1–46.



Qi-Ming He is currently a professor in the Department of Management Sciences at the University of Waterloo. He received a Ph.D from the Institute of Applied Mathematics, Chinese Academy of Sciences in 1989 and a Ph.D from the Department of Management Science at the University of Waterloo in 1996. His main research areas are algorithmic methods in applied probability, queueing theory, inventory control, and production management. In investigating various stochastic models, his favourite methods are matrix analytic methods. Recently, he is working on queueing systems with multiple types of customers, inventory systems with multiple types of demands, and representations of phase-type distributions and their applications.

Attahiru S. Alfa is a professor of telecommunication systems at the University of Manitoba, Department of Electrical and Computer Engineering and also a SARChI Chair professor at the University of Pretoria, Department of Electrical, Electronic and Computer Engineering. His research covers, but not limited to, the following areas: performance analysis and resource allocation in telecommunication systems, modeling of communication networks, queueing theory, optimization, analysis of cognitive radio networks, modeling and analysis of wireless sensor networks, developing efficient decoding algorithms for LDPC codes, channel modeling, traffic estimation for the Internet, and cross layer analysis. Dr. Alfa also works in the application of queueing theory to other areas such as transportation systems, manufacturing systems and healthcare systems. He was NSERC Chair for tele-traffic from 2004 to 2012. He has carried out applied research for Nortel Networks, Bell-Northern Research, TRLabs (now TRTech), Bell Canada, Winnipeg Regional Health Authority, Motor-coach Industries, and several other industries. He has authored a book, "Queueing Theory for Telecommunications: Discrete Time Modelling of a Single Node System", published by Springer in 2010, and a forthcoming book, "Applied Discrete Time Queueing Theory" to be published by Springer in 2015.