Math Meth Oper Res (2018) 87:139–168
https://doi.org/10.1007/s00186-017-0612-2

ORIGINAL ARTICLE

An *M/PH/K* queue with constant impatient time

Qi-Ming He[1] · **Hao Zhang**[2] · **Qingqing Ye**[3]

Received: 26 September 2016 / Published online: 5 October 2017

Abstract This paper is concerned with an *M/PH/K* queue with customer abandonment, constant impatient time, and many servers. By combining the method developed in Choi et al. (Math Oper Res 29:309–325, 2004) and Kim and Kim (Perform Eval 83–84:1–15, 2015) and the state space reduction method introduced in Ramaswami (Stoch Models 1:393–417, 1985), the paper develops an efficient algorithm for computing performance measures for the queueing system of interest. The paper shows a number of properties associated with matrices used in the development of the algorithm, which make it possible for the algorithm, under certain conditions, to handle systems with up to one hundred servers. The paper also obtains analytical properties of performance measures that are useful in gaining insight into the queueing system of interest.

Keywords Queueing systems · Markov process · Matrix-analytic methods · Impatient customers

Mathematics Subject Classification Primary: 60K25 · Secondary: 90-08

We thank NSERC.

✉ Qi-Ming He
q7he@uwaterloo.ca

1 Department of Management Sciences, University of Waterloo, 200 University Avenue West, Waterloo, ON N2L 3G1, Canada

2 Department of Management Engineering, Wuhan University, Wuhan 430072, China

3 Department of Statistics and Financial Mathematics, Nanjing University of Science and Technology, Nanjing 210094, China

1 Introduction

Due to customer impatience, system congestion control, or other factors, customer abandonment can be seen in many queueing systems. For some cases (e.g., supermarket checkout systems), the impact of customer abandonment on the system is negligible. For other cases (e.g., call centers), the impact can be significant. For the call center case, ignoring customer abandonment may lead to over or under staffing and hence huge financial loss (see Mandelbaum and Zeltyn 2013). Thus, the study of queues with customer abandonment or queues with impatient customers has attracted the attention of both researchers and practitioners for decades.

The study of queues with customer abandonment traces back to as early as 1950's. Early works include Barrer (1957a, b), Finch (1960), Daley (1965), Gnedenko and Kovalenko (1968), and Jurkevic (1970, 1971), in which the *M/M/c* type abandonment queues are investigated. Extensions to *M/G/*1 and *GI/M/*1 type abandonment queues are carried out in later years (e.g., Baccelli et al. 1984; Kok and Tijms 1985; Stanford 1990; Boots and Tijms 1999; Xiong et al. 2008). While some of existing works consider models with random impatient times, others consider models with a constant impatient time (e.g., Daley 1965; Boots and Tijms 1999). More recently, customer impatience is considered in queueing models with variable arrival rate, multi-types of customers, and service priority (e.g., Movaghar 1998; Brandt and Brandt 1999a, b; Choi et al. 2001, and van Houdt 2012). Approximation methods have been utilized in analyzing abandonment queueing systems with a large number of servers (e.g., Brandt and Brandt 2002; Garnett et al. 2002; Dai and Tezcan 2008; Dai and He 2010, 2011; Dai et al. 2010).

While classical queueing analysis focuses on analytic results or approximation results, recent works explore algorithmic methods for computing performance measures. In Choi et al. (2004), an *MAP/M/K* queue with constant impatient time is investigated. In Kim and Kim (2015), an *M/PH/*1 queue with constant impatience time and variable service rate is investigated. To take the advantage of matrix-analytic methods, both papers develop computational methods for computing performance measures related to customer waiting time and queue length.

This paper combines the solution approach developed in Choi et al. (2004) and Kim and Kim (2015) and the state space reduction method developed in Ramaswami (1985) to investigate the *M/PH/K* queue with constant impatient time, Markov modulated service rate, and many servers. Compared to Choi et al. (2004), this paper considers a model with a more general service time distribution through an environment process for service. Compared to Kim and Kim (2015), this paper considers a model with multiple servers. Similar to Kim and Kim (2015), a Markov process associated with the age of the head-of-queue customer is introduced (also see He 2005). Using the Markov process, the paper obtains performance measures related to waiting time and queue length. To deal with the state space dimensionality issue, this paper uses an approach developed in Ramaswami (1985) (also see Ramaswami and Lucantoni 1985; Asmussen and O'Cinneide 1998) to reduce the state space so that the algorithm developed in this paper can handle systems with up to one hundred servers. A number of properties associated with matrices used in the algorithm are proved.

Kawanishi and Takine (2016) also investigated the *M/PH/K* queue with constant impatience time. Their analysis is based on approach introduced in Choi et al. (2004) as well. An algorithm for performance analysis, which is similar to ours, is developed in that paper. Compared to Kawanishi and Takine (2016), our paper: (i) provides details on the construction of the associated Markov process using the space reduction method introduced in Ramaswami (1985), (ii) finds some quantities explicitly, and (iii) shows a number of properties of matrices involved in computation. Consequently, the computational procedure developed in our paper is more convenient to use and more efficient numerically. Since typical applications of queues with customer abandonment are in the design and/or analysis of call centers and telecommunications systems that have many servers, this paper increases the applicability of the theory developed in Choi et al. (2004), Kim and Kim (2015), and Kawanishi and Takine (2016) significantly.

For queues with customer abandonment and a large number of servers, diffusion approximation works well for performance analysis (see Dai and He 2011). If the system has only one server, exact results are obtained for many such queueing systems. The method developed in this paper works efficiently for the abandonment queue with many servers, and the analysis is exact. Thus, the results in this paper fill the gap in the existing literature.

Markov modulated fluid queue (MMFQ) has been proven to be an effective tool in analyzing queueing models (e.g., see Dzial et al. (2005); Houdt (2012), and Meini 2013). The basic idea of the approach is to introduce a Markov modulated fluid flow process associated with the workload process of the queueing system of interest. If the stationary distribution of the fluid flow process can be found, then some queueing quantities can be obtained. In the literature, the MMFQ method has been used in analyzing single-server queues successfully. For our $M/PH/K$ queue with constant impatient time, however, the associated state space can become too large with the increase of the number of servers (as shown in Sect. 2) for numerical computation. In this case, the MMFQ method would encounter the same dimensionality issue with the approach taken in this paper. Thus, the advantage of MMFQ is not immediately clear in the study of the $M/PH/K$ queue with constant impatient time. In this paper, we choose to follow the lead by Choi et al. (2004) and Kim and Kim (2015), and leave the MMFQ approach for future research.

The main contributions of this paper include: (i) combining two existing approaches to develop computational procedures for computing distributions and moments of any order of waiting times and queue lengths for queueing systems with multiple servers; and (ii) showing properties to improve the efficiency of algorithms, which make it possible to analyze queues with a large number of servers, and to gain insight into the queueing system of interest.

The rest of the paper is organized as follows. In Sect. 2, we introduce the queueing model of interest and a Markov process associated with the age of the customer at the head of the queue. The stationary distribution of the Markov process is obtained in Sect. 3. Some technical details for the Markov process and its stationary distribution are collected in Sect. 4 and "Appendixes A and B". In Sect. 5, computational procedures are developed for two auxiliary matrices. A number of performance measures are obtained in Sect. 6. In Sect. 7, we present a few numerical examples and discuss some computational issues when the number of servers is big. Section 8 concludes the paper.

2 Queueing model and Markov process associated with customer age

We consider a multi-server queueing model with impatient customers. Upon arrival, all customers join a single queue and are served on a first-come-first-served basis. There are K identical servers. The service rate of servers changes according to a Markov process. When the waiting time of a customer reaches constant time τ, the customer leaves the system without service. The queueing model is defined explicitly as follows.

(i) Customers arrive according to a Poisson process with parameter λ.
(ii) All customers join a single queue waiting for service and are served on a first-come-first-served basis. If a customer's waiting time reaches constant time τ, the customer leaves the system immediately without service.
(iii) There are K identical servers. When a server becomes available, the customer at the head of the queue (if there is any) enters the server for service. If an arriving customer finds an idle server, the customer enters the server for service upon arrival.
(iv) The workload associated with each customer has a phase-type distribution with PH-representation $(\boldsymbol{\beta}, T)$ of order m_s. We assume that $\boldsymbol{\beta}\mathbf{e} = 1$, i.e., the workload of a customer is always positive. Let $\mathbf{T}^0 = -T\mathbf{e}$, where $\mathbf{e}$ is the column vector of ones. We assume that $T + \mathbf{T}^0\boldsymbol{\beta}$ is irreducible, i.e., the PH-representation $(\boldsymbol{\beta}, T)$ is PH-irreducible. Let $\boldsymbol{\varphi} = (\varphi_1, \varphi_2, \ldots, \varphi_{m_s})$ be the row vector satisfying $\boldsymbol{\varphi}(T + \mathbf{T}^0\boldsymbol{\beta}) = \mathbf{0}$ and $\boldsymbol{\varphi}\mathbf{e} = 1$. Note that $\mathbf{0}$ is a vector of zeros. Since the PH-representation is irreducible, $\boldsymbol{\varphi}$ is the unique solution to the linear system. Note that $\boldsymbol{\varphi}$ is the stationary distribution of the underlying Markov chain associated with the PH-renewal process whose interarrival times have PH-distribution $(\boldsymbol{\beta}, T)$. The mean workload is given by $-\boldsymbol{\beta}T^{-1}\mathbf{e}$. It is well-known that $\boldsymbol{\varphi}\mathbf{T}^0 = 1/(-\boldsymbol{\beta}T^{-1}\mathbf{e})$. See Neuts (1981) for more about phase-type distributions.
(v) The service rate of servers is modulated by a continuous time Markov chain with infinitesimal generator Q and m_e states (to be called the environment process). We assume that the process is irreducible. Let $\boldsymbol{\pi}$ be the stationary distribution of Q. Then $\boldsymbol{\pi}$ is the unique solution to linear system $\boldsymbol{\pi}Q = \mathbf{0}$ and $\boldsymbol{\pi}\mathbf{e} = 1$.
(vi) The service rate in state j of the environment process is μ_j, for $j = 1, 2, \ldots, m_e$. Thus, if the state of the environment process changes, the service rate may change in the middle of a service. Let M be an $m_e \times m_e$ matrix with $\{\mu_j, j = 1, 2, \ldots, m_e\}$ on its diagonal and all other elements zero. Let $\mu = \boldsymbol{\pi}M\mathbf{e}$, which is the mean service rate.
(vii) Define $\rho = \lambda\boldsymbol{\beta}(-T^{-1})\mathbf{e}/(K\mu)$, which can be considered to be the offered load to the system.

To obtain performance measures for the queueing model, we utilize a Markov process associated with the age of the customer at the head of the queue. We introduce the Markov process in this section and develop computational methods for performance measures in Sects. 3, 4, 5, and 6.

The *age* of a customer is defined as the time elapsed since the customer enters the system. Since customers arrive according to a Poisson process, tracking the age of the customer in the head of the queue and the service processes of individual servers, together with information on the environment, provides enough information

to describe the dynamics of the queueing system. Consequently, system performance can be analyzed. Define

- $a(t)$: the age of the first customer waiting in the queue at time t, if the (waiting) queue is not empty; otherwise, $a(t) = -\infty$.
- $I_e(t)$: the state of the environment process at time t.
- $n_i(t)$: the number of servers whose service phase is i at time t, for $i = 1, 2, \ldots, m_s$.

It is easy to see that $\{(a(t), I_e(t), n_1(t), \ldots, n_{m_s}(t)),\ t \geq 0\}$ is a continuous time Markov process. Based on the total number of working servers, the state space of $(n_1(t), \ldots, n_{m_s}(t))$ can be organized as $\Omega(0) \cup \Omega(1) \cup \ldots \cup \Omega(K)$, where, for $k = 0, 1, 2, \ldots, K$,

$$\Omega(k) = \left\{ (n_1, \ldots, n_{m_s}) :\ n_i \geq 0, n_i \text{ integer},\ i = 1, \ldots,\ m_s,\ \sum_{i=1}^{m_s} n_i = k \right\}. \quad (1)$$

The set $\Omega(k)$ consists of all states such that there are exactly k customers in service (or k working servers), for $k = 0, 1, \ldots, K$. The number of states in $\Omega(k)$ is given by $(k + m_s - 1)!/(k!(m_s - 1)!)$. Then the state space of the Markov process can be given as

$$\left\{ \{-\infty\} \times \{1, \ldots, m_e\} \times \{\cup_{k=0}^{K} \Omega(k)\} \right\} \cup \{[0, \tau) \times \{1, \ldots, m_e\} \times \Omega(K)\}. \quad (2)$$

Note Instead of using $\{n_1(t), \ldots, n_{m_s}(t)\}$, a more straightforward and simple way to model the service process is to keep track of the workload process for each server. However, the number of states required by that approach to track the service status of the K servers is m_s^K, which is significantly greater than that of the current approach.

To analyze the Markov process, we need to identify the transitions between states explicitly. For that purpose, we first construct the transitions within $\Omega(0) \cup \Omega(1) \cup \ldots \cup \Omega(K)$. Note that the number of customers in service can change at most one when the system state changes. Thus, the matrix for transition rates within $\Omega(0) \cup \Omega(1) \cup \ldots \cup \Omega(K)$ can be written as

$$-\lambda I + \begin{array}{c} \\ \Omega(0) \\ \Omega(1) \\ \vdots \\ \Omega(K-1) \\ \Omega(K) \end{array} \begin{array}{c} \begin{array}{cccccc} \Omega(0) & \Omega(1) & & \cdots & \Omega(K-1) & \Omega(K) \end{array} \\ \begin{pmatrix} S(0,m_s) & S^+(0,m_s) & & & & \\ S^-(1,m_s) & S(1,m_s) & S^+(1,m_s) & & & \\ & \ddots & \ddots & \ddots & & \\ & & S^-(K-1,m_s) & S(K-1,m_s) & S^+(K-1,m_s) & \\ & & & S^-(K,m_s) & S(K,m_s) \end{pmatrix} \end{array}, \quad (3)$$

where I is the identity matrix (whose order depends on the context) and the three sets of transition blocks $\{S^+(k, m_s), k = 0, 1, \ldots, K-1\}$, $\{S^-(k, m_s), k = 1, 2, \ldots, K\}$, and $\{S(k, m_s), k = 0, 1, 2, \ldots, K\}$ can be constructed explicitly using an algorithm developed in Ramaswami (1985). Those transition blocks (or matrices) are associated with the arrivals, departures, and workloads of customers, respectively. In this paper, we use a slightly different iterative method to construct those matrices. The method

used in this paper is more amenable for showing properties (e.g., Propositions 2.1 and 2.2) that are used in developing the main algorithm. For completeness, we present the iterative method in "Appendix A". The following properties of the transition blocks are used in the paper.

Proposition 2.1 *For the transition blocks in expression (3), we have (i)* $S^+(k, m_s)\boldsymbol{e} = \lambda \boldsymbol{e}$, *for* $k = 0, 1, 2, \ldots, K-1$; *(ii)* $S(0, m_s) = 0$; *and (iii)* $S^-(k, m_s)\boldsymbol{e} + S(k, m_s)\boldsymbol{e} = \mathbf{0}$, *for* $k = 1, 2, \ldots, K$.

Proof We first point out that this proof does not depend on the details of the construction process of the matrices. Since the customer arrival rate is λ, it is easy to see that $S^+(k, m_s)\mathbf{e} = \lambda \mathbf{e}$. Since the service process is governed only by individual (underlying) Markov processes associated with individual servers, we must have $S^-(k, m_s)\mathbf{e} + S(k, m_s)\mathbf{e} = \mathbf{0}$. This completes the proof of Proposition 2.1.

Next, we focus on transitions related to states in $\Omega(K)$. Recall that $\boldsymbol{\varphi} = (\varphi_1, \varphi_2, \ldots, \varphi_{m_s})$ is the stationary distribution associated with the PH-renewal process. Define row vector $\boldsymbol{\phi} = (\phi(\mathbf{n}) : \mathbf{n} \in \Omega(K))$, where

$$\phi(\mathbf{n}) = \frac{K!}{n_1! \cdots n_{m_s}!} \prod_{j=1}^{m_s} \varphi_j^{n_j}, \quad \text{for } \mathbf{n} = (n_1, \ldots, n_{m_s}) \in \Omega(K), \tag{4}$$

which is the probability mass function of the well-known multi-nomial distribution and we have $\boldsymbol{\phi}\mathbf{e} = 1$, i.e.

$$\sum_{\mathbf{n} \in \Omega(K)} \phi(\mathbf{n}) = \sum_{\mathbf{n} \in \Omega(K)} \frac{K!}{n_1! \cdots n_{m_s}!} \prod_{j=1}^{m_s} \varphi_j^{n_j} = \left(\sum_{j=1}^{m_s} \varphi_j \right)^K = 1, \tag{5}$$

since $\boldsymbol{\varphi}\mathbf{e}=1$. Elements in vector $\boldsymbol{\phi}$ are organized in the same order as the states in $\Omega(K)$, which are arranged lexicographically (see Eq. (50)). The vector $\boldsymbol{\phi}$ can be constructed recursively as follows. Let $\varpi(0, m) = 1$, for $m = 1, 2, \ldots, m_s$, and $\varpi(k, 1) = \varphi_1^k$, for $k = 0, 1, 2, \ldots, K$;

$$\begin{aligned} &\boldsymbol{\varpi}(k, m) \\ &= \left(\boldsymbol{\varpi}(k, m-1), \frac{\boldsymbol{\varpi}(k-1, m-1)k!\varphi_m}{(k-1)!1!}, \ldots, \frac{\varpi(0, m-1)k!\varphi_m^k}{0!k!} \right) \\ &= \left(\boldsymbol{\varpi}(k-j, m-1) \frac{k!\varphi_m^j}{(k-j)!j!} : \quad j = 0, 1, \ldots, k \right), \end{aligned} \tag{6}$$

for $m = 1, 2, \ldots, m_s$, and $k = 1, 2, \ldots, K$. In Eq. (6), the expression $k!\varphi_m^j/((k-j)!j!)$ can be evaluated numerically by using, for $j = 1, 2, \ldots, k$,

$$\frac{k!\varphi_m^j}{(k-j)!j!} = \left(\frac{k}{j}\varphi_m \right) \left(\frac{k-1}{j-1}\varphi_m \right) \cdots \left(\frac{k-(j-1)}{j-(j-1)}\varphi_m \right), \tag{7}$$

which avoids the computation of $k!$. Then we obtain $\boldsymbol{\phi} = \boldsymbol{\varpi}(K, m_s)$. □

Proposition 2.2 *That* $S(K, m_s) + S^-(K, m_s)S^+(K-1, m_s)/\lambda$ *is an irreducible infinitesimal generator of a continuous time Markov chain and its corresponding stationary distribution is given by* $\boldsymbol{\phi}$*, i.e,* $\boldsymbol{\phi}(S(K, m_s) + S^-(K, m_s)S^+(K-1, m_s)/\lambda) = \mathbf{0}$ *and* $\boldsymbol{\phi}\mathbf{e} = 1$*. In addition, we have* $\boldsymbol{\phi}S^-(K, m_s)\boldsymbol{e} = K/(-\boldsymbol{\beta}S^{-1}\boldsymbol{e})$*.*

Proof Note that $S^+(K-1, m_s)/\lambda$ gives the transition probabilities from $\Omega(K–1)$ to $\Omega(K)$, given that a customer has just arrived. It is then clear that $S(K, m_s) + S^-(K, m_s)S^+(K-1, m_s)/\lambda$ gives transition rates for transitions between states in $\Omega(K)$, given that there are always waiting customers. Thus, the matrix has to be an infinitesimal generator. Since the *PH*-representation of the service workload is irreducible, the infinitesimal generator is also irreducible. In steady state, since φ_j is the probability that the service phase of a server is j, the probability that the service state is $\mathbf{n} \in \Omega(K)$ is given by $\phi(\mathbf{n})$. Thus, $\boldsymbol{\phi}$ is the stationary distribution of the Markov chain associated with the infinitesimal generator.

For state $\mathbf{n} \in \Omega(K)$, the total service completion rate is $\sum_{j=1}^{m_s} n_j t_j^0$, where t_j^0 is the j-th element of $\mathbf{T}^0$ (i.e., $\mathbf{T}^0 = \left(t_j^0\right)_{m_s \times 1}$). Then we have

$$
\begin{aligned}
\boldsymbol{\phi}S^-(K, m_s)\mathbf{e} &= \sum_{\mathbf{n}\in\Omega(K)} \phi(\mathbf{n}) \left(\sum_{j=1}^{m_s} n_j t_j^0 \right) = \sum_{j=1}^{m_s} \left(\sum_{\mathbf{n}\in\Omega(K)} \phi(\mathbf{n}) n_j \right) t_j^0 \\
&= \sum_{j=1}^{m_s} \left(\sum_{\mathbf{n}\in\Omega(K)} \frac{K!}{n_1! \cdots n_{m_s}!} \prod_{i=1}^{m_s} \varphi_i^{n_i} n_j \right) t_j^0 \\
&= K \sum_{j=1}^{m_s} \left(\sum_{\mathbf{n}\in\Omega(K):\, n_j \geq 1} \frac{(K-1)! \varphi_j^{n_j - 1}}{n_1! \cdots (n_j - 1)! \cdots n_{m_s}!} \prod_{i=1:\, i\neq j}^{m_s} \varphi_i^{n_i} \right) \varphi_j t_j^0 \\
&= K \sum_{j=1}^{m_s} \left(\sum_{\mathbf{n}\in\Omega(K-1)} \frac{(K-1)!}{n_1! \cdots n_{m_s}!} \prod_{i=1}^{m_s} \varphi_i^{n_i} \right) \varphi_j t_j^0 \\
&= K \sum_{j=1}^{m_s} \left(\sum_{i=1}^{m_s} \varphi_i \right)^{K-1} \varphi_j t_j^0 = K \sum_{j=1}^{m_s} \varphi_j t_j^0 = \frac{K}{\boldsymbol{\beta}(-S)^{-1}\mathbf{e}}.
\end{aligned} \tag{8}
$$

This completes the proof of Proposition 2.2. □

3 Stationary distribution

We are interested in the stationary distribution of $\{(a(t), I_e(t), n_1(t), \ldots, n_{m_s}(t)), t \geq 0\}$. If $\tau < \infty$, the Markov process is Harris recurrent (Harris 1956) and its stationary distribution exists, which can be proved similar to that in Kim and Kim (2015). In this section, we introduce a set of equations for the stationary distribution and present its solution. In steady state, we define, for an empty system,

- $p_0(i) = P\{a(t) = -\infty, I_e(t) = i_e, n_1(t) = \cdots = n_{m_s}(t) = 0\}$, for $i_e = 1, 2, \ldots, m_e$, and
- $\mathbf{p}_0 = (p_0(1), p_0(2), \ldots, p_0(m_e))$;

for a system with only k customers (in service), for $k = 1, 2, \ldots, K$,

- $p_k(i_e, \mathbf{n}) = P\{a(t) = -\infty, I_e(t) = i_e, n_1(t) = n_1, \ldots, n_{m_s}(t) = n_{m_s}\}$, for $i_e = 1, 2, \ldots, m_e$, and $\mathbf{n} \in \Omega(k)$,
- $\mathbf{p}_k(i_e) = (p_k(i_e, \mathbf{n}), \mathbf{n} \in \Omega(k))$, for $i_e = 1, 2, \ldots, m_e$, and
- $\mathbf{p}_k = (\mathbf{p}_k(1), \mathbf{p}_k(2), \ldots, \mathbf{p}_k(m_e))$;

and, for a system with at least one waiting customer,

$$p_{K+1}(x, i_e, \mathbf{n}) = \frac{\mathrm{d}}{\mathrm{d}x} P\{a(t) < x,\ I_e(t) = i_e, n_i(t) = n_i,\ i = 1,\ \ldots,\ m_s\}, \quad (9)$$

for $0 \le x < \tau$, $i_e = 1, 2, \ldots, m_e$, and $\mathbf{n} \in \Omega(K)$, where τ is constant and the maximum waiting time of each customer. The vector $\mathbf{p}_{K+1}(x)$ is defined in a way similar to $\mathbf{p}_K$.

For better understanding of the analysis, we would like to point out that elements in vectors $\{\mathbf{p}_0, \mathbf{p}_1, \ldots, \mathbf{p}_K\}$ are probabilities and the elements in $\mathbf{p}_{K+1}(x)$ are probability density functions (we shall call them transition rates throughout the paper as they associate with transitions of the age of the customer at the head of the queue). Similar to Choi et al. (2004) and Kim and Kim (2015), the following fundamental equations can be established for $\{\mathbf{p}_0, \mathbf{p}_1, \ldots, \mathbf{p}_K, \mathbf{p}_{K+1}(x)\}$:

$$\begin{aligned}
\mathbf{0} &= \mathbf{p}_0(-\lambda I + Q) + \mathbf{p}_1(M \otimes S^-(1, m_s));\\
\mathbf{0} &= \mathbf{p}_{k-1}(I \otimes S^+(k-1, m_s)) + \mathbf{p}_k(-\lambda I + Q \otimes I + M \otimes S(k, m_s))\\
&\quad + \mathbf{p}_{k+1}\left(M \otimes S^-(k+1, m_s)\right), \quad \text{for } k = 1,\ 2, \ldots,\ K-1;\\
\mathbf{0} &= \mathbf{p}_{K-1}(I \otimes S^+(K-1, m_s)) + \mathbf{p}_K(-\lambda I + Q \otimes I + M \otimes S(K, m_s))\\
&\quad + \frac{1}{\lambda}\int_0^\tau \mathbf{p}_{K+1}(y) e^{-\lambda y} \mathrm{d}y \left(M \otimes \left(S^-(K, m_s) S^+(K-1, m_s)\right)\right)\\
&\quad + \mathbf{p}_{K+1}(\tau) e^{-\lambda\tau};\\
\frac{\mathrm{d}}{\mathrm{d}x}\mathbf{p}_{K+1}(x) &= \mathbf{p}_{K+1}(x)(Q \otimes I + M \otimes S(K, m_s)) + \mathbf{p}_{K+1}(\tau)\lambda e^{-\lambda(\tau - x)}\\
&\quad + \int_x^\tau \mathbf{p}_{K+1}(y) e^{-\lambda(y-x)} \mathrm{d}y \left(M \otimes \left(S^-(K, m_s) S^+(K-1, m_s)\right)\right),\\
&\quad \text{for } 0 \le x < \tau;\\
\mathbf{p}_{K+1}(0) &= \lambda \mathbf{p}_K,
\end{aligned} \quad (10)$$

where "$\otimes$" represents Kronecker product of matrices. The derivation of the above equations is routine but tedious. Intuitively, the first term in the second, third, and fourth equations is related to arrivals, the middle term is related to phase changes of the environment and workload, and the last term is related to departures. Note that the Q matrix is used to modulate the phase of the environment since the service rate can change if the environment changes. We omit the details.

We present expressions of $\{\mathbf{p}_0, \mathbf{p}_1, \ldots, \mathbf{p}_K, \mathbf{p}_{K+1}(x)\}$ in the rest of this section. First, we find a relationship between $\{\mathbf{p}_0, \mathbf{p}_1, \ldots, \mathbf{p}_K\}$. To that end, we define the following auxiliary matrices that appear in the expressions of those vectors:

$$\begin{aligned} D_1 &= (M \otimes S^-(1, m_s))(\lambda I - Q)^{-1}; \\ D_{k+1} &= \left(M \otimes S^-(k+1, m_s)\right) \\ &\quad \times \left(\lambda I - Q \otimes I - M \otimes S(k, m_s) - D_k \left(I \otimes S^+(k-1, m_s)\right)\right)^{-1}, \end{aligned} \tag{11}$$

for $k = 1, 2, \ldots, K-1$.

Proposition 3.1 *The matrices* $\{D_k, k = 1, 2, \ldots, K\}$ *exist and are nonnegative.*

Proof Since all eigenvalues of Q have a negative real part, except the eigenvalue zero, it is easy to see that λ is not an eigenvalue of Q. Then $\lambda I - Q$ is invertible. Let $\delta = \max\{|Q_{j,j}|\}$. Then $\lambda I - Q = (\lambda + \delta)I - (Q + \delta I)$, and all eigenvalues of matrix $(Q + \delta I)/(\lambda + \delta)$ are within the unit disk. Then

$$(\lambda I - Q)^{-1} = \frac{1}{\lambda + \delta} \sum_{n=0}^{\infty} \left(\frac{\delta I + Q}{\lambda + \delta}\right)^n > 0, \tag{12}$$

elementwise. Consequently, D_1 exists and is nonnegative. Since $(\lambda I - Q)\mathbf{e} = \lambda \mathbf{e}$, we have $\lambda(\lambda I - Q)^{-1}\mathbf{e} = \mathbf{e}$ and $D_1\mathbf{e} = (M \otimes S^-(1, m_s))\mathbf{e}/\lambda$. By induction and Proposition 2.1, it can be shown that, for $k = 1, 2, \ldots, K$,

$$\begin{aligned} D_k\mathbf{e} &= \left(M \otimes S^-(k, m_s)\right)\mathbf{e}/\lambda; \\ &\quad \left(Q \otimes I + M \otimes S(k, m_s) + D_k \left(I \otimes S^+(k-1, m_s)\right)\right)\mathbf{e} = \mathbf{0}. \end{aligned} \tag{13}$$

In the proof, we use the facts that matrix $Q \otimes I + M \otimes S(k, m_s) + D_k \left(I \otimes S^+(k-1, m_s)\right)$ is an infinitesimal generator and $\lambda I - (Q \otimes I + M \otimes S(k, m_s) + D_k(I \otimes S^+(k-1, m_s)))$ is invertible and its inverse is nonnegative, by a similar argument used for Eq. (12). This completes the proof of Proposition 3.1.

By Eq. (10) and Proposition 3.1, we obtain

$$\mathbf{p}_k = \mathbf{p}_K D_K \cdots D_{k+1}, \quad \text{for } k = 0,\ 1,\ 2, \ldots,\ K-1. \tag{14}$$

Next, we find $\mathbf{p}_{K+1}(x)$, which is used to find $\mathbf{p}_K$. By differentiating with respect to x the 4-th equation in the system of Eq. (10), we obtain, for $0 \le x < \tau$,

$$\begin{aligned} 0 = &\frac{\mathrm{d}^2\mathbf{p}_{K+1}(x)}{\mathrm{d}x^2} - \frac{\mathrm{d}\mathbf{p}_{K+1}(x)}{\mathrm{d}x}(\lambda I + Q \otimes I + M \otimes S(K, m_s) \\ &+\lambda \mathbf{p}_{K+1}(x)(Q \otimes I) \\ &+\lambda \mathbf{p}_{K+1}(x)\left(M \otimes \left(S(K, m_s) + \frac{S^-(K, m_s)S^+(K-1, m_s)}{\lambda}\right)\right). \end{aligned} \tag{15}$$

Following a similar approach in Choi et al. (2004) or Kim and Kim (2015), a closed form solution of $\mathbf{p}_{K+1}(x)$ can be obtained explicitly (see "Appendix B" for more details). If $\rho \neq 1$, we obtain, for $0 \leq x < \tau$,

$$\begin{aligned}\mathbf{p}_{K+1}(x) = &\mathbf{u}_1 \exp\{\lambda(R-I)(\tau-x)\} \\ &+\mathbf{u}_2 \exp\{(\lambda G + Q \otimes I + M \otimes S(K, m_s))\, x\},\end{aligned} \tag{16}$$

where R and G are the minimal nonnegative solutions to matrix equations

$$\begin{aligned}R^2 A_0 + R A_1 + A_2 &= \mathbf{0}; \\ A_0 G^2 + A_1 G + A_2 &= \mathbf{0},\end{aligned} \tag{17}$$

respectively,

$$\begin{aligned}A_0 &= \lambda I, \\ A_1 &= Q \otimes I + M \otimes S(K, m_s) - \lambda I, \\ A_2 &= \frac{1}{\lambda} M \otimes \left(S^-(K, m_s) S^+(K-1, m_s)\right);\end{aligned} \tag{18}$$

and $\{\mathbf{u}_1, \mathbf{u}_2\}$ is the unique solution to linear system,

$$\begin{aligned}\mathbf{0} = &\mathbf{u}_1 e^{\lambda(R-I)\tau} \left(D_K(I \otimes S^+(K-1, m_s))\frac{1}{\lambda} - R\right) \\ &+ \mathbf{u}_2 \left(G + \frac{1}{\lambda}\left(A_1 + D_K(I \otimes S^+(K-1, m_s))\right)\right); \\ \mathbf{0} = &\mathbf{u}_1 (A_1 + \lambda I + \lambda R) + \mathbf{u}_2 \lambda e^{(\lambda G + A_1 + \lambda I)\tau} (I - G),\end{aligned} \tag{19}$$

with normalization condition $1 = \sum_{k=0}^{K} \mathbf{p}_k \mathbf{e} + \int_0^\tau \mathbf{p}_{K+1}(x)\mathbf{e}dx$. If $\rho < 1$,

$$\begin{aligned}1 = &\frac{1}{\lambda}\mathbf{u}_1 (R - I + \boldsymbol{\xi}(\boldsymbol{\pi} \otimes \boldsymbol{\phi}))^{-1} \left(e^{\lambda(R-I)\tau} - I + \lambda\tau\boldsymbol{\xi}(\boldsymbol{\pi} \otimes \boldsymbol{\phi})\right)\mathbf{e} \\ &+\mathbf{u}_2 (\lambda(G+I) + A_1)^{-1} \left(e^{(\lambda(G+I)+A_1)\tau} - I\right)\mathbf{e} \\ &+\frac{1}{\lambda}\left(\mathbf{u}_1 e^{\lambda(R-I)\tau} + \mathbf{u}_2\right)\left(\mathbf{e} + \sum_{k=0}^{K-1}\left(\prod_{j=0}^{K-(k+1)} D_{K-j}\right)\mathbf{e}\right),\end{aligned} \tag{20}$$

and, if $\rho > 1$,

$$\begin{aligned}1 = &\frac{1}{\lambda}\mathbf{u}_1 (R - I)^{-1} \left(e^{\lambda(R-I)\tau} - I\right)\mathbf{e} \\ &+\mathbf{u}_2 (\lambda(G+I) + A_1 + \boldsymbol{\zeta}(\boldsymbol{\pi} \otimes \boldsymbol{\phi}))^{-1} \\ &\times \left(e^{(\lambda(G+I)+A_1)\tau} - I + \tau\boldsymbol{\zeta}(\boldsymbol{\pi} \otimes \boldsymbol{\phi})\right)\mathbf{e} \\ &+\frac{1}{\lambda}\left(\mathbf{u}_1 e^{\lambda(R-I)\tau} + \mathbf{u}_2\right)\left(\mathbf{e} + \sum_{k=0}^{K-1}\left(\prod_{j=0}^{K-(k+1)} D_{K-j}\right)\mathbf{e}\right),\end{aligned} \tag{21}$$

vector $\boldsymbol{\xi}$, is the right maximal eigenvector of R satisfying $(\boldsymbol{\pi}\otimes\boldsymbol{\phi})\boldsymbol{\xi}$ =1 (i.e., $R\boldsymbol{\xi} = \boldsymbol{\xi}$ and $(\boldsymbol{\pi}\otimes\boldsymbol{\phi})\boldsymbol{\xi} = 1$), and vector $\boldsymbol{\zeta}$ is the right maximal eigenvector of $\lambda(G + I) + A_1$ satisfying $(\boldsymbol{\pi} \otimes \boldsymbol{\phi})\boldsymbol{\zeta}$ =1 (i.e., $(\lambda(G + I) + A_1)\boldsymbol{\zeta} = 0$ and $(\boldsymbol{\pi}\otimes\boldsymbol{\phi})\boldsymbol{\zeta} = 1$).

Details on properties related to $\{R, G, \mathbf{u}_1, \mathbf{u}_2\}$ and justification of the above solutions can be found in Sect. 4 and "Appendix B".

By Eqs. (10) and (16), we obtain

$$\mathbf{p}_K = \frac{1}{\lambda}\mathbf{p}_{K+1}(0) = \frac{1}{\lambda}\left(\mathbf{u}_1 \exp\{\lambda(R - I)\tau\} + \mathbf{u}_2\right). \tag{22}$$

In summary, $\{\mathbf{p}_0, \mathbf{p}_1, \ldots, \mathbf{p}_K, \mathbf{p}_{K+1}(x)\}$ can be obtained in the following steps, if $\rho \neq 1$.

(i) Computing $\{R, G\}$ Eq. (17).
(ii) Use Eq. (4) to find $\boldsymbol{\phi}$, and compute $\boldsymbol{\xi}$ or $\boldsymbol{\zeta}$.
(iii) Computing $\{\mathbf{u}_1, \mathbf{u}_2\}$ based on Eqs. (19)–(20) or (19) and (21).
(iv) Compute $\mathbf{p}_{K+1}(x)$ by Eq. (16).
(v) Compute $\mathbf{p}_K$ by Eq. (22).
(vi) Compute $\{\mathrm{D}_k, k = 1, 2, \ldots, K\}$ by Eq. (11).
(vii) Compute $\{\mathbf{p}_0, \mathbf{p}_1, \ldots, \mathbf{p}_{K-1}\}$ by Eq. (14).

Theorem 3.1 *If $\rho \neq 1$, Eq. (10) has a unique solution given by Eqs. (14), (16), and (22).*

Proof The solution is presented in terms of $\{\mathbf{u}_1, \mathbf{u}_2, R, G\}$. Justification of the solution, e.g., the invertibility of some matrices, shall be given in Sect. 4 and "Appendix B". The uniqueness of the solution can be addressed in a way similar to that in Choi et al. (2004) and Kim and Kim (2015). Details are omitted.

Note We would like to point out that solutions for the case with $\rho = 1$ can be obtained but are tedious. Thus, we present solutions for that case in "Appendix B".

4 Properties of interest

In this section, we present some properties that can be used for the justification of the results presented in Sect. 3. To understand matrices R and G, it is helpful to consider a fictitious quasi birth-and-death (QBD) process defined by $\{A_0, A_1, A_2\}$ (and some matrices for boundary transitions). The theory on QBD processes can be applied to show properties related to R and G (see Neuts 1981; Latouche and Ramaswami 1999; He 2014).

Proposition 4.1 *Vector $\boldsymbol{\pi} \otimes \boldsymbol{\phi}$ is the stationary distribution of infinitesimal generator $A_0+A_1+A_2$. Further, we have $\rho = (\boldsymbol{\pi} \otimes \boldsymbol{\phi})A_0\boldsymbol{e}/(\boldsymbol{\pi} \otimes \boldsymbol{\phi})A_2\boldsymbol{e}$).*

Proof By the definition of $\boldsymbol{\pi}$ and Proposition 2.2, we have

$$\begin{aligned}
&(\boldsymbol{\pi} \otimes \boldsymbol{\phi})\,(A_0 + A_1 + A_2) \\
&= (\boldsymbol{\pi} \otimes \boldsymbol{\phi}) \left(Q \otimes I + M \otimes S(K, m_s) + \frac{1}{\lambda} M \otimes \left(S^-(K, m_s) S^+(K-1, m_s) \right) \right) \\
&= \boldsymbol{\pi} Q \otimes \boldsymbol{\phi} + \boldsymbol{\pi} M \otimes \boldsymbol{\phi} \left(S(K, m_s) + \frac{1}{\lambda} \left(S^-(K, m_s) S^+(K-1, m_s) \right) \right) = \mathbf{0}.
\end{aligned} \tag{23}$$

The second part of the proposition also comes from Proposition 2.2. This completes the proof of Proposition 4.1.

Proposition 4.2 *(a) The matrices R and G are invertible only if K = 1 and m_s = 1; (b) The matrices $\lambda R + A_1$ and $\lambda G + A_1$ are always invertible; and (c) The matrices $\lambda(R + I) + A_1$ and $\lambda(G + I) + A_1$ are invertible, if $\rho < 1$; and non-invertible, if $\rho \geq 1$.*

Proof To show the results, we define matrices $\hat{R}$ and $\hat{G}$ as the minimal nonnegative solutions to

$$\begin{aligned}
A_0 + \hat{R} A_1 + \hat{R}^2 A_2 &= \mathbf{0}; \\
A_0 + A_1 \hat{G} + A_2 \hat{G}^2 &= \mathbf{0},
\end{aligned} \tag{24}$$

respectively. Since $A_0 = \lambda \mathrm{I}$ is invertible, both $\hat{R}$ and $\hat{G}$ are invertible. Let sp(.) the maximal absolute value of all eigenvalues of a matrix. It is well-known that, if $\rho < 1$, $\mathrm{sp}(\hat{R}) < 1$ and $\mathrm{sp}(\hat{G}) < 1$; otherwise, $\mathrm{sp}(\hat{R}) = 1$ and $\mathrm{sp}(\hat{G}) = 1$.

(a) By Latouche (1987), we have $\lambda R = A_2 \hat{G}$ and $\lambda G = \hat{R} A_2$. Since A_2 is invertible only if $K = 1$ and $m_s = 1$, R and G are invertible only if $K = 1$ and $m_s = 1$.
(b) Since $\lambda G + A_1 = -\lambda \hat{R}^{-1}$, $\lambda G + A_1$ is always invertible. Since $\lambda R + A_1 = -\lambda \hat{G}^{-1}$, $\lambda R + A_1$ is always invertible.
(c) Since $\lambda(G+I)+A_1 = \lambda(I - \hat{R}^{-1})$, $\lambda(G + I) + A_1$ is invertible, if $\rho < 1$; and non-invertible, if $\rho \geq 1$. Since $\lambda(R + I) + A_1 = \lambda(I - \hat{G}^{-1})$, $\lambda(R + I) + A_1$ is invertible, if $\rho < 1$; and non-invertible, if $\rho \geq 1$.

This completes the proof of Proposition 4.2. □

Proposition 4.3 *The infinitesimal generator $A_0+A_1+A_2$ is irreducible.*

(i) If $\rho \leq 1$, sp(R) = 1; otherwise, sp(R) < 1.
(ii) Vector $\boldsymbol{\pi} \otimes \boldsymbol{\phi}$ is a left maximal eigenvector of R. If $\rho \leq 1$, $(\boldsymbol{\pi} \otimes \boldsymbol{\phi})R = \boldsymbol{\pi} \otimes \boldsymbol{\phi}$; otherwise, $(\boldsymbol{\pi} \otimes \boldsymbol{\phi})R \leq \boldsymbol{\pi} \otimes \boldsymbol{\phi}$ and $(\boldsymbol{\pi} \otimes \boldsymbol{\phi})R \neq \boldsymbol{\pi} \otimes \boldsymbol{\phi}$. (Note: The comparison between vectors is element-wise.)
(iii) If $\rho \leq 1$, G is stochastic and sp(G) = 1; otherwise, G is substochastic and sp(G) < 1.
(iv) Vector $\boldsymbol{\pi} \otimes \boldsymbol{\phi}$ is a left maximal eigenvector of $\lambda(G + I) + A_1$. If $\rho < 1$, $(\boldsymbol{\pi} \otimes \boldsymbol{\phi})(\lambda(G + I) + A_1) \leq \mathbf{0}$ and $(\boldsymbol{\pi} \otimes \boldsymbol{\phi})(\lambda(G + I) + A_1) \neq \mathbf{0}$. otherwise, $(\boldsymbol{\pi} \otimes \boldsymbol{\phi})(\lambda(G + I) + A_1) = \mathbf{0}$.

Proof Parts (i), (ii), and (iii) are known results for a QBD process with transition blocks $\{\lambda I,\ Q \otimes I + M \otimes S(K, m_s) - \lambda I,\ M \otimes (S^-(K, m_s)S^+(K-1, m_s))/\lambda\}$.

Next, we prove part (iv). By Eq. (24), we have $-\lambda \hat{R}^{-1} = A_1 + \hat{R}A_2$. Since $(\boldsymbol{\pi} \otimes \boldsymbol{\phi})\hat{R} \leq \boldsymbol{\pi} \otimes \boldsymbol{\phi}$, we obtain $\lambda(\boldsymbol{\pi} \otimes \boldsymbol{\phi})\left(I - \hat{R}^{-1}\right) \leq \mathbf{0}$. If $\rho < 1$, $\mathrm{sp}\left(\hat{R}\right) < 1$ and, consequently, $\lambda(\boldsymbol{\pi} \otimes \boldsymbol{\phi})\left(I - \hat{R}^{-1}\right) \neq 0$; otherwise, $\lambda(\boldsymbol{\pi} \otimes \boldsymbol{\phi})\left(I - \hat{R}^{-1}\right) = \mathbf{0}$. By Proposition 4.2, we have $\lambda(G + I) + A_1 = \lambda(I - \hat{R}^{-1})$. Part (iv) is proved. This completes the proof of Proposition 4.3. □

Recall vectors $\boldsymbol{\xi}$ and $\boldsymbol{\zeta}$ defined and used in Sect. 3. Proposition 4.3 implies the existence of $\boldsymbol{\xi}$, if $\rho < 1$, $\boldsymbol{\zeta}$, if $\rho > 1$, and both $\boldsymbol{\xi}$ and $\boldsymbol{\zeta}$, if $\rho = 1$. Based on Proposition 4.3, the following results can be proved routinely.

Proposition 4.4 *If $\rho < 1$, the matrix $R - I$ is non-invertible and the matrix $R - I + \boldsymbol{\xi}(\boldsymbol{\pi} \otimes \boldsymbol{\phi})$ is invertible. If $\rho > 1$, the matrix $R - I$ is invertible and the matrix $\lambda(G + I) + A_1 + \boldsymbol{\zeta}(\boldsymbol{\pi} \otimes \boldsymbol{\phi})$ is invertible. If$\rho = 1$, the matrix $R - I$ is non-invertible, the matrix $R - I + \boldsymbol{\xi}(\boldsymbol{\pi} \otimes \boldsymbol{\phi})$ is invertible, and the matrix $\lambda(G + I) + A_1 + \boldsymbol{\zeta}(\boldsymbol{\pi} \otimes \boldsymbol{\phi})$ is invertible.*

The inverses of $R - I + \boldsymbol{\xi}(\boldsymbol{\pi} \otimes \boldsymbol{\phi})$ and $\lambda(G + I) + A_1 + \boldsymbol{\xi}(\boldsymbol{\pi} \otimes \boldsymbol{\phi})$ play a key role in computing various quantities and performance measures (e.g., Eq. (20) and (27)). If $\rho \leq 1$, the inverses exist, and $\boldsymbol{\xi}$ and $(\boldsymbol{\pi} \otimes \boldsymbol{\phi})$ are their right and left eigenvectors corresponding to eigenvalue one, respectively.

5 Computation of two integrals

In this section, we develop computational procedures for two integrals that are used for computing the distributions and moments of the waiting times and queue lengths in Sect. 6. We begin with the following integral, which is the key for computing the moments of waiting time,

$$\mathbf{h}_n = \int_0^\tau x^n \mathbf{p}_{K+1}(x)\mathrm{d}x, \quad \text{for } n = 0,\ 1,\ 2,\ \ldots \tag{25}$$

Note that $\mathbf{h}_0 = \int_0^\tau \mathbf{p}_{K+1}(x)\mathrm{d}x$, which is used in the normalization condition for $\mathbf{u}_1$ and $\mathbf{u}_2$ (see (38)), and the computations of the distributions of waiting time and queue length (see (44)). Define two auxiliary matrices, for $n = 0, 1, 2, \ldots$,

$$\begin{aligned} H_{R,n} &= \int_0^\tau x^n \exp\{\lambda(R - I)(\tau - x)\}\mathrm{d}x; \\ H_{G,n} &= \int_0^\tau x^n \exp\left\{(\lambda G + Q \otimes I + M \otimes S(K, m_s))\, x\right\}\mathrm{d}x. \end{aligned} \tag{26}$$

By Proposition 4.3, if $\rho \leq 1$, matrices $R - I + \boldsymbol{\xi}(\boldsymbol{\pi} \otimes \boldsymbol{\phi})$ and $\lambda G + A_1 + \lambda I$ are invertible, and, if $\rho > 1$, matrices $R - I$ and $\lambda G + A_1 + \lambda I + \boldsymbol{\zeta}(\boldsymbol{\pi} \otimes \boldsymbol{\phi})$ are invertible. By routine calculations, we obtain

$$H_{R,0} = \begin{cases} \frac{1}{\lambda}\left(e^{\{\lambda(R-I)\tau\}} - I + \lambda\tau\boldsymbol{\xi}(\boldsymbol{\pi}\otimes\boldsymbol{\phi})\right) \\ \qquad \times (R - I + \boldsymbol{\xi}(\boldsymbol{\pi}\otimes\boldsymbol{\phi}))^{-1}, & \text{if } \rho \le 1; \\ \frac{1}{\lambda}\left(\exp\{\lambda(R-I)\tau\} - I\right)(R-I)^{-1}, & \text{if } \rho > 1, \end{cases}$$

$$H_{R,n} = \begin{cases} \frac{1}{\lambda}\left(\frac{\lambda\tau^{n+1}}{n+1}\boldsymbol{\xi}(\boldsymbol{\pi}\otimes\boldsymbol{\phi}) + nH_{R,n-1} - \tau^n I\right) \\ \qquad \times (R - I + \boldsymbol{\xi}(\boldsymbol{\pi}\otimes\boldsymbol{\phi}))^{-1}, \quad \text{for } n = 1,\ 2,\ \ldots, \text{ if } \rho \le 1; \\ \frac{1}{\lambda}\left(nH_{R,n-1} - \tau^n I\right)(R-I)^{-1}, \quad \text{for } n = 1,\ 2,\ \ldots, \text{ if } \rho > 1. \end{cases} \tag{27}$$

and

$$H_{G,0} = \begin{cases} \left(e^{(\lambda(G+I)+A_1)\tau} - I\right)(\lambda(G+I)+A_1)^{-1}, & \text{if } \rho \le 1; \\ \left(e^{(\lambda(G+I)+A_1)\tau} - I + \tau\boldsymbol{\zeta}(\boldsymbol{\pi}\otimes\boldsymbol{\phi})\right) \\ \qquad \times (\lambda(G+I)+A_1+\boldsymbol{\zeta}(\boldsymbol{\pi}\otimes\boldsymbol{\phi}))^{-1}, & \text{if } \rho > 1; \end{cases}$$

$$H_{G,n} = \begin{cases} \left(\tau^n e^{(\lambda(G+I)+A_1)\tau} - nH_{G,n-1}\right) \\ \qquad \times (\lambda(G+I)+A_1)^{-1}, \quad \text{for } n = 1,\ 2,\ \ldots, \text{ if } \rho \le 1; \\ \left(\tau^n e^{(\lambda(G+I)+A_1)\tau} + \frac{\tau^{n+1}}{n+1}\boldsymbol{\zeta}(\boldsymbol{\pi}\otimes\boldsymbol{\phi}) - nH_{G,n-1}\right) \\ \times (\lambda(G+I)+A_1+\boldsymbol{\zeta}(\boldsymbol{\pi}\otimes\boldsymbol{\phi}))^{-1}, \text{for } n = 1, 2, \ldots, \text{if } \rho > 1; \end{cases} \tag{28}$$

Consequently, by Eq. (16), we obtain

Proposition 5.1 *The integral* $\mathbf{h}_n$ *can be computed by*

$$\mathbf{h}_n = \mathbf{u}_1 H_{R,n} + \mathbf{u}_2 H_{G,n}, \quad \textit{for } n = 0,\ 1,\ 2,\ \ldots \tag{29}$$

Next, for computing the mean queue length (see equations (41) and (42) in Sect. 6), we consider auxiliary vectors, for l = 0, 1, 2, …,

$$\Gamma(l,0) = \sum_{n=K+1}^{\infty}\int_0^{\tau} n^l \mathbf{p}_{K+1}(x)\left(\frac{(\lambda x)^{n-K-1}}{(n-K-1)!}e^{-\lambda x}\right)dx. \tag{30}$$

In general, we define, for $l = 0, 1, 2, \ldots$, and $j = 0, 1, 2, \ldots$,

$$\Gamma(l,j) = \lambda^j \sum_{n=K+1+j}^{\infty}\int_0^{\tau} x^j n^l \mathbf{p}_{K+1}(x)\left(\frac{(\lambda x)^{n-K-1-j}}{(n-K-1-j)!}e^{-\lambda x}\right)\mathrm{d}x. \tag{31}$$

By rewriting n as $n = n - K - 1 - j + K + 1 + j$, we obtain

$$\Gamma(l,j) = \Gamma(l-1, j+1) + (K+1+j)\Gamma(l-1,j). \tag{32}$$

Note that $\Gamma(0, j)$ can be obtained as, for $j = 0, 1, 2, \ldots$,

$$\begin{aligned}\Gamma(0, j) &= \lambda^j \sum_{n=K+1+j}^{\infty} \int_0^{\tau} x^j \mathbf{p}_{K+1}(x) \left(\frac{(\lambda x)^{n-K-1-j}}{(n-K-1-j)!} e^{-\lambda x} \right) \mathrm{d}x \\ &= \lambda^j \int_0^{\tau} x^j \mathbf{p}_{K+1}(x) \mathrm{d}x = \lambda^j \left(\mathbf{u}_1 H_{R,j} + \mathbf{u}_2 H_{G,j} \right).\end{aligned} \tag{33}$$

Consequently, we obtain

Proposition 5.2 *The sequence* $\{\Gamma(l, 0), l = 0, 1, 2, \ldots\}$ *can be computed recursively by using Eqs. (32) and (33).*

6 Performance measures

A number of performance measures can be obtained from $\{\mathbf{p}_0, \mathbf{p}_1, \ldots, \mathbf{p}_K, \mathbf{p}_{K+1}(x)\}$ and the two quantities obtained in Sect. 5. The most natural one is the stationary distribution of the age of the customer at the head of the queue, which is given by

$$\lim_{t\to\infty} P\{a(t) < x\} = \begin{cases} \sum_{k=0}^{K} \mathbf{p}_k \mathbf{e}, & \text{for } x < 0; \\ \sum_{k=0}^{K} \mathbf{p}_k \mathbf{e} + \int_0^x \mathbf{p}_{K+1}(y) \mathbf{e} \mathrm{d}y, & \text{for } 0 \le x < \tau; \\ 1, & \text{for } x \ge \tau. \end{cases} \tag{34}$$

The integral $\int_0^x \mathbf{p}_{K+1}(y)\mathrm{d}y$ can be computed using (29) with τ replaced by x in (27) and (28). Letting $x = \tau$, the above equation leads to $1 = \sum_{k=0}^{K} \mathbf{p}_k \mathbf{e} + \mathbf{h}_0 \mathbf{e}$, which is the normalization condition used in (20) and (21) and can be used to check computation accuracy.

In the rest of this section, we focus on the customer loss probability, waiting times, and queue lengths.

Note that if the number of customers in the system is less than K, a new arrival can enter a server for service. Therefore, the loss of customers can happen only if all servers are occupied. The ratio of customer loss rate and the customer arrival rate gives the percentage of customers lost per unit time, which is also the probability that a customer is lost.

Proposition 6.1 *The customer loss probability is given by*

$$\begin{aligned} p_{loss} &= \frac{1}{\lambda} \mathbf{p}_{K+1}(\tau-) \mathbf{e} = \frac{1}{\lambda} \left(\mathbf{u}_1 \mathbf{e} + \mathbf{u}_2 \exp\{(\lambda(G+I) + A_1)\tau\} \mathbf{e} \right); \\ \lim_{\tau\to\infty} p_{loss} &= \max\left\{0,\ 1 - \frac{1}{\rho}\right\}. \end{aligned} \tag{35}$$

Proof When the waiting time of a customer reaches τ, the customer is lost. The probability of a customer being lost (or the probability a customer leaves the system without service) can be obtain as the ratio of customer loss rate $\mathbf{p}_{K+1}(\tau-)\mathbf{e}$ at an arbitrary time and the total arrival rate λ. The expression of p_{loss} can be obtained from Eq. (16) directly.

For the limit of p_{loss}, it is easy to see the result for $\rho \leq 1$. For $\rho > 1$, we must have $\mathbf{p}_K$ approaches zero as τ goes to infinity. By Eq. (22), we must have $\mathbf{u}_2 = 0$ if $\tau \to \infty$ (recall that $\text{sp}(R) < 1$ for this case). By the expression of p_{loss}, we obtain $p_{loss} = \mathbf{u}_1\mathbf{e}/\lambda$ if$\tau \to \infty$. Next, we find $\mathbf{u}_1\mathbf{e}/\lambda$ if $\tau = \infty$. By equations (19) and (21), if $\tau \to \infty$, we obtain

$$\begin{aligned} \mathbf{0} &= \mathbf{u}_1 \left(\lambda R + \lambda I + A_1\right); \\ 1 &= \frac{\mathbf{u}_1}{\lambda}(I - R)^{-1}\mathbf{e}. \end{aligned} \tag{36}$$

The first equation in (36) leads to $\mathbf{u}_1 = \mathbf{u}_1\hat{G}$. Thus, $\mathbf{u}_1$ is unique up to a constant multiplier.

To show the results, we establish a relationship between vector $\mathbf{u}_1$ and the probabilities that the system is empty in a *PH/M/*1 queue. We consider a *PH*/*M*/1 queue for which the service times are exponentially distributed with parameter λ and the arrival process is determined by the superposition of K identical *PH*-renewal process (see Neuts 1981) with *PH*-representation $(\boldsymbol{\beta}, T)$ modularized by the Markov process Q. For this *PH/M/*1 queue, a QBD process can be constructed for its queue length process, which has the same R matrix. It can then be verified that $\boldsymbol{\pi}_0 = \mathbf{u}_1/\lambda$, where $\boldsymbol{\pi}_0$ contains the conditional probability that the queueing system is empty. By Neuts (1981), we obtain $p_{loss} = \mathbf{u}_1\mathbf{e}/\lambda = \boldsymbol{\pi}_0\mathbf{e}$. It is obvious that, for the *PH/M/*1 queue, the service rate is λ and the arrival rate is $(\boldsymbol{\pi} \otimes \boldsymbol{\phi})A_2\mathbf{e} = K\mu/(\boldsymbol{\beta}(-T)^{-1}\mathbf{e})$. Thus, we have $p_{loss} = \boldsymbol{\pi}_0\mathbf{e} = 1 - K\mu/(\boldsymbol{\beta}(-T)^{-1}\mathbf{e}\lambda) = 1 - 1/\rho$. This completes the proof of Proposition 6.1.

Let W_a be the waiting time of an arbitrary customer, and W_q the waiting time of a customer who receives service. By the conditional PASTA (König and Schmidt 1990), the distribution of the waiting time of an arbitrary customer equals the distribution of the waiting time of the "fictitious" customer leaving the waiting queue at an arbitrary time (leaving the system or entering service).

Proposition 6.2 *The density function of the waiting time W_a of an arbitrary customer is given by*

$$\begin{aligned} P\{W_a = 0\} &= \sum_{k=0}^{K-1} \mathbf{p}_k\mathbf{e}; \\ \frac{d}{dx}P\{W_a \leq x\} &= \frac{1}{\lambda}\mathbf{p}_{K+1}(x)\left(M \otimes S^-(K, m_s)\right)\mathbf{e}, \quad \textit{for } 0 < x < \tau; \\ P\{W_a = \tau\} &= \frac{1}{\lambda}\mathbf{p}_{K+1}(\tau-)\mathbf{e}. \end{aligned} \tag{37}$$

Proof A customer enters service with zero waiting time if and only if there is at least one available server when the customer arrives. Then $P\{W_a = 0\}$ is obtained. Note that the total rate for customers leaving the waiting queue is λ. A customer with waiting time $0 < x < \tau$ enters a server if and only if there is a service completion at that epoch. This gives the density at $W_a = x$. A customer leaves the system without service if and only if its age reaches τ. This leads to $P\{W_a = \tau\}$. By König and Schmidt (1990), the proposition is proved. □

Note that the total average departure (including both customers with or without service) is λ. By the law of total probability, we must have

$$\begin{aligned} 1 &= P\{W_a \le \tau\} \\ &= \sum_{k=0}^{K-1} \mathbf{p}_k \mathbf{e} + \frac{1}{\lambda} \int_0^\tau \mathbf{p}_{K+1}(y)\mathrm{d}y \left(M \otimes S^-(K, m_s)\right) \mathbf{e} + \frac{\mathbf{p}_{K+1}(\tau)\mathbf{e}}{\lambda} \\ &= \sum_{k=0}^{K-1} \mathbf{p}_k \mathbf{e} + \frac{1}{\lambda} \mathbf{h}_0 \left(M \otimes S^-(K, m_s)\right) \mathbf{e} \\ &\quad + \frac{1}{\lambda} \left(\mathbf{u}_1 + \mathbf{u}_2 \exp((\lambda(G + I) + A_1)\tau\right) \mathbf{e}, \end{aligned} \tag{38}$$

which is useful for checking computation accuracy. Explicit expression for integral $\mathbf{h}_0 = \int_0^\tau \mathbf{p}_{K+1}(y)\mathrm{d}y$ is given in (29). The moments of waiting times can be obtained in terms of $\{H_{R,n}, H_{G,n}, n = 0, 1, 2, \ldots\}$ as follows, for $n = 1, 2, 3, \ldots$,

$$\begin{aligned} E[W_a^n] &= \frac{1}{\lambda} \int_0^\tau x^n \mathbf{p}_{K+1}(x)\mathrm{d}x \left(M \otimes S^-(K, m_s)\right) \mathbf{e} + \tau^n \frac{1}{\lambda} \mathbf{p}_{K+1}(\tau-)\mathbf{e} \\ &= \frac{1}{\lambda} \left(\mathbf{u}_1 H_{R,n} + \mathbf{u}_2 H_{G,n}\right) \left(M \otimes S^-(K, m_s)\right) \mathbf{e} + \tau^n \frac{1}{\lambda} \mathbf{p}_{K+1}(\tau-)\mathbf{e}. \end{aligned} \tag{39}$$

Using conditional probabilities, the density function of the waiting time W_q of arbitrary customer who received service can be obtained as

$$\begin{aligned} P\{W_q = 0\} &= \frac{1}{1 - p_{loss}} \sum_{k=0}^{K-1} \mathbf{p}_k \mathbf{e}; \\ \frac{\mathrm{d}P\{W_q \le x\}}{\mathrm{d}x} &= \frac{\mathbf{p}_{K+1}(x) \left(M \otimes S^-(K, m_s)\right) \mathbf{e}}{(1 - p_{loss})\lambda}, \quad \text{for } 0 < x < \tau; \\ P\{W_q \ge \tau\} &= 0. \end{aligned} \tag{40}$$

Let N_{all} be the total number of customers in the system at an arbitrary time, which is the sum of the number of customers in service and the number of customers waiting in queue. By the well-known PASTA property, N_{all} has the same probability distribution as the number of customers seen by an arriving customer. Customers waiting in queue are the customer at the head of the queue and the customers who arrive during the waiting period of the head-of-queue customer.

Proposition 6.3 *The distribution of N_{all} is given by*

$$P\{N_{all} = n\} = \begin{cases} \mathbf{p}_n\mathbf{e}, & \text{for } n = 0,\ 1,\ \ldots,\ K; \\ \displaystyle\int_0^\tau \mathbf{p}_{K+1}(x)\mathbf{e}\left(\frac{(\lambda x)^{n-K-1}e^{-\lambda x}}{(n-K-1)!}\right)dx, & \text{for } n \geq K+1. \end{cases} \tag{41}$$

Proof If $N_{all} = n > K$, there are K customers in service and $n - K$ customers are waiting for service. Among the $n - K$ customers, one is waiting at the head of the queue and the other $n - K - 1$ customers are all arrived when the first customer is waiting. By conditioning on the age of the customer at the head of the queue, we can calculate the probability that there are $n - K - 1$ customers wait behind that customer. Note that the number of arrivals in $[0, x]$ has a Poisson distribution with parameter λ. Then the proposition is proved by conditioning on the age of the head-of-queue customer.

The moments of N_{all} can be obtained as, for $l = 1, 2, \ldots$,

$$\begin{aligned} E[N_{all}^l] &= \sum_{n=0}^{K} n^l \mathbf{p}_n\mathbf{e} + \sum_{n=K+1}^{\infty}\int_0^\tau n^l \mathbf{p}_{K+1}(x)\mathbf{e}\left(\frac{(\lambda x)^{n-K-1}}{(n-K-1)!}e^{-\lambda x}\right)\mathrm{d}x \\ &= \sum_{n=0}^{K} n^l \mathbf{p}_n\mathbf{e} + \Gamma(l,0)\mathbf{e}. \end{aligned} \tag{42}$$

Let N_q be the number of waiting customers in queue. Then the distribution of N_q is given by

$$P\{N_q = n\} = \begin{cases} \displaystyle\sum_{k=0}^{K} \mathbf{p}_k\mathbf{e}, & \text{for } n = 0; \\ \displaystyle\int_0^\tau \mathbf{p}_{K+1}(x)\mathbf{e}\left(\frac{(\lambda x)^{n-1}}{(n-1)!}e^{-\lambda x}\right)\mathrm{d}x, & \text{for } n = 1,\ 2,\ \ldots \end{cases} \tag{43}$$

Let N_{ws} be the number of working servers. Then the distribution of N_{ws} is given by

$$P\{N_{ws} = k\} = \begin{cases} \mathbf{p}_k\mathbf{e}, & \text{for } k = 0,\ 1,\ 2,\ \ldots,\ K-1; \\ \mathbf{p}_K\mathbf{e} + \displaystyle\int_0^\tau \mathbf{p}_{K+1}(x)\mathbf{e}\mathrm{d}x = \mathbf{p}_K\mathbf{e} + \mathbf{h}_0\mathbf{e}, & \text{for } k = K. \end{cases} \tag{44}$$

Note that $E[N_{all}]$, $E[N_q]$, $E[N_{ws}]$, $E[W_a]$, and $E[W_q]$ are finite if τ is finite. Therefore, Little's law holds for our queueing system and we must have

$$
\begin{aligned}
E[N_{all}] &= \lambda E[W_a] + E[N_{ws}] \\
&= \lambda E[W_a] + \sum_{k=0}^{K} k\mathbf{p}_k\mathbf{e} + K\int_0^{\tau} \mathbf{p}_{K+1}(x)\mathbf{e}dx \\
&= \lambda\left(p_{loss}\tau + (1-p_{loss})E[W_q]\right) + \sum_{k=0}^{K} k\mathbf{p}_k\mathbf{e} + K\mathbf{h}_0\mathbf{e}. \qquad (45)
\end{aligned}
$$

This relationship is useful for checking computation accuracy.

7 Numerical examples

We present three examples to demonstrate the feasibility of the computation procedure developed in Sects. 3, 4, 5 and 6 and "Appendixes A and B", and to explore the relationship between system parameters $\{\lambda, \tau, K\}$ and system performance measures such as the loss probability, mean waiting time and mean queue length. Before presenting our examples, we briefly discuss a few computational issues and, for some of them, provide hints on how the issue can be addressed.

(i) Computation results, especially inverse matrices, become inaccurate if $|\Omega(K)|$ is big (e.g., $|\Omega(K)| \geq 1000$). Thus, avoiding the use of those inverse matrices could improve computation efficiency. In our computation, we evaluate expressions with an inverse matrix by transforming the problem into a linear system. For example, to calculate $X = BA^{-1}$, we solve the linear system $XA = B$ to find X, which avoids the use of the inverse matrix of A. The computation of all quantities in Sect. 5 and all performance measures in Sect. 6 can be done in this manner.

(ii) Computation results may becomes inaccurate if ρ is close to one. For such a case, the source of inaccuracy comes mainly from matrices R and G. More iterations are needed in the computation of the two matrices if ρ is one or close to one.

(iii) The normalization conditions (20) and (21) can be handled by solving linear systems to avoid computing inverse matrices.

(iv) Use Propositions 2.1 and 2.2 to check the correctness/accuracy of the transition blocks and vector $\boldsymbol{\phi}$. Use Proposition 4.3 to check the correctness/accuracy of matrices R and G. Use Eqs. (38) and (45) to check the correctness/accuracy of performance measures. Use limits to check correctness of computation: (a) If τ goes to infinity, the queueing system becomes the classical $M/PH/K$ queue; and (b) If K goes to infinity, the queueing system becomes the classical $M/PH/\infty$ queue.

Example 7.1 We consider the example in Kawanishi and Takine (2016), for which $\lambda = 4.8$, $\tau = 1$, $K = 20$, $m_e = 1$, $Q = 0$ (since there is only one environment state), $\mu = \mu_1 = 1$,

$$
m_s = 2, \quad \boldsymbol{\beta} = (1, 0), \quad T = \begin{pmatrix} -0.25 & 0.25 \\ 0 & -1 \end{pmatrix}; \qquad (46)
$$

Table 1 The moments of waiting times and queue length

n	1	2	3	4	5	6	7	8
$E[W_a^n]$	0.519	0.425	0.374	0.3418	0.3194	0.3031	0.2905	0.2806
$E[N_{all}^n]$	21.81	487.5	1.1e+4	2.5e+5	6.1e+6	1.4e+8	3.6e+9	9.0e+10

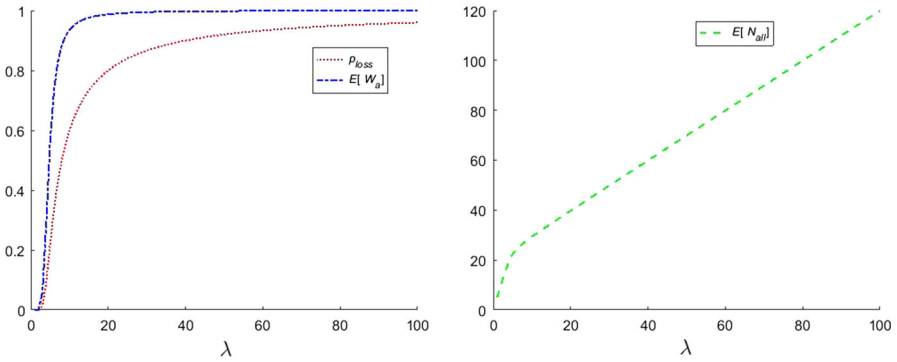

Fig. 1 Performance measures for λ in [0, 100]

For this queue, $\rho = 1.2$, $p_{loss} = 0.1950$, $E[N_q] = 2.49$, $E[N_{ws}] = 19.32$, and the moments of waiting time W_a and queue lengths N_{all} are given in Table 1.

Next, by letting parameters $\{\lambda, \tau, K\}$ vary, we have a look at their impact on performance measures $\{p_{loss}, E[W_a], E[N_{all}]\}$.

(i) First, we allow the customer arrival rate λ to go from 1 to 100, while all other system parameters remain the same. We plot $\{p_{loss}, E[W_a], E[N_{all}]\}$ as functions of λ in Fig. 1.

Figure 1 shows that p_{loss} approaches 1 and $E[W_a]$ approaches 1. The mean waiting time $E[W_a]$ approaches 1 since $\tau = 1$. The mean queue length $E[N_{all}]$ approaches infinity as λ increases to infinity.

(ii) Second, we let the impatient time τ to go from 1 to 100. We plot $\{p_{loss}$,$E[W_a]$, $E[N_{all}]\}$ as functions of τ in Figure 2.

For this case, the traffic intensity is $\rho = 1.2$, which is greater than 1. As τ goes to infinity, the system becomes the classical $M/PH/K$ queue, and the loss probability converges to $1 - 1/\rho$ and both the mean waiting time and mean queue length go to infinity.

(iii) Lastly, we change K from 1 to 300. We plot $\{p_{loss}, E[W_a], E[N_{all}]\}$ as functions of K in Fig. 3.

For this case, the performance measures converge to that of the classical $M/PH/\infty$ queue. For example, the mean queue length converges to 24, which is the mean queue length of the corresponding classical $M/PH/\infty$ queue.

In Example 7.1, since $m_s = 2$, we have $|\Omega(K)| = K + 1$. Consequently, the size of all matrix blocks (to be called *block size*) involved in the computation of performance

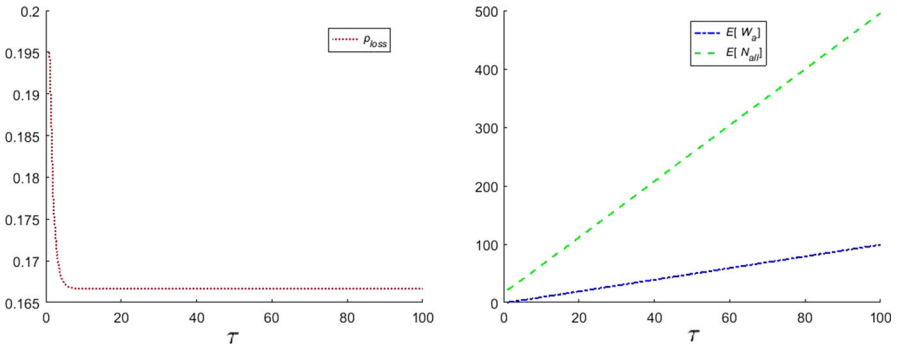

Fig. 2 Performance measures for τ in [1, 100]

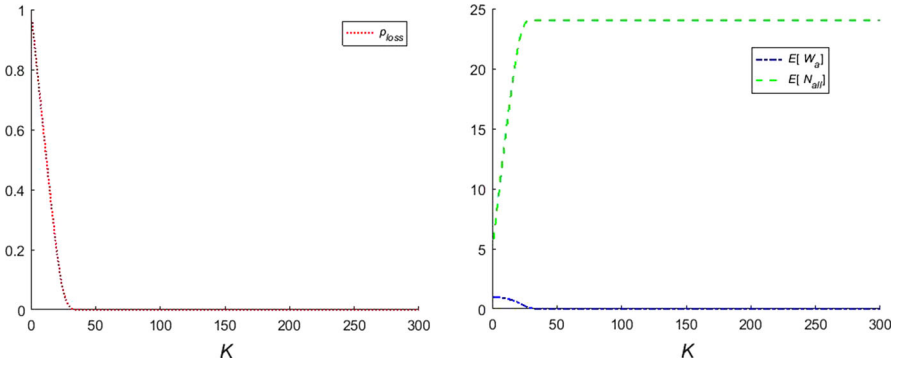

Fig. 3 Performance measures for K from 1 to 300

Table 2 Values of $|\Omega(K)|$ for (m_s, K)

m_s/K	10	16	20	30	50	100	200
2	11	17	21	31	51	101	201
3	66	153	231	496	1326	5151	20301
4	286	969	1771	5456	23426	–	–
5	1001	4845	10626	46376	−−	–	–

"–" means that the number is more than one hundred thousand.

measures is smaller or equal to $m_e(K+1)$, which is linearly increasing in K. The computation of performance measures can be done for large K effectively. Recall that $|\Omega(K)| = (K + m_s - 1)!/((m_s - 1)!K!)$. If $m_s > 2$, the sizes of matrix blocks such as $\{A_0, A_1, A_2\}$ increase much faster as K increases. We show the values of $|\Omega(K)|$ (i.e., the size of the matrices $\{A_0, A_1, A_2\}$) in Table 2 for a few pairs of (m_s, K).

- Suppose that we limit $|\Omega(K)|$ to be less than or approximately equal to 5000. If $m_s = 3$, K can go up to 100. If $m_s = 4$, K can go up to 30. If $m_s = 5$, K can go up to 16. If $m_e = 1$, all those cases can be handled by an average computer with

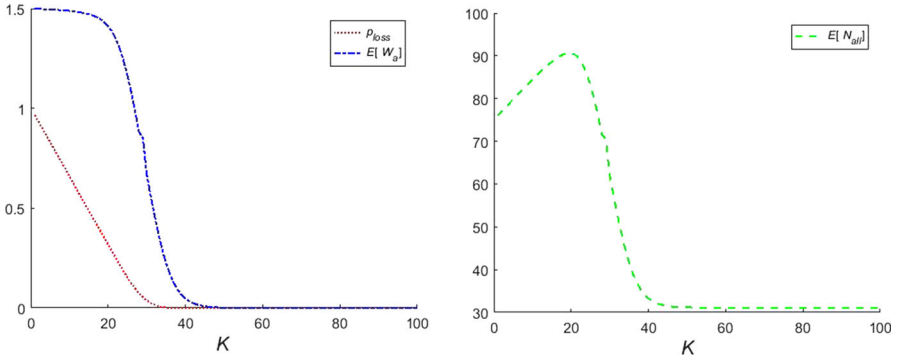

Fig. 4 Performance measures for K from 1 to 100

eight gigabytes of memory. If m_e is two or bigger, the computer memory required for the implementation of the algorithm increases at the rate m_e^2.

- According to the literature on PH-distributions (see Neuts 1981; He 2014), the probability distributions that can be generated using PH-representations of order $m_s = 2, 3, 4$, and 5 are quite versatile. Thus, the algorithms developed in this paper does have the potential to be used in practice.

Example 7.2 We consider a system with $\lambda = 50$, $\tau = 1.5$, $\mu_1 = 1$, $\mu_2 = 0.5$,

$$m_s = 3, \quad \boldsymbol{\beta} = (0.6,\ 0.2,\ 0.2), \quad T = \begin{pmatrix} -4 & 0.2 & 0.5 \\ 1 & -3 & 0.5 \\ 0.1 & 1 & -3.5 \end{pmatrix};$$
$$m_e = 2, \quad Q = \begin{pmatrix} -1 & 1 \\ 0.5 & -0.5 \end{pmatrix}. \tag{47}$$

For this example, we have $m_s = 3$ and $m_e = 2$. We compute performance quantities for systems with K up to 100. At $K = 100$, the largest block size is nearly 10, 000. We plot $\{p_{loss}, E[W_a], E[N_{all}]\}$ as functions of K in Fig. 4. It is interesting to see that the mean queue length goes up to 90 and then down to 31.1164, which is the mean queue length of the corresponding $M/PH/\infty$ queue. The reason for that is that the total queue consists of two parts: customers waiting for service and customers in service. When K is small, more customers are waiting and abandon the queue. When K increases, waiting time becomes shorter and the waiting queue increases, and the queue of customers in service also increases. When K is sufficiently large, due to fixed customer arrival rate, the queue of waiting customers disappears. Consequently, the total queue decreases and converges to that of the $M/PH/\infty$ queue.

Example 7.3 We consider a system with $\lambda = 20$, $\tau = 2$, $m_e = 1$, $Q = 0$, $\mu_1 = 1$,

$$m_s = 4, \quad \boldsymbol{\beta} = (0.2, 0.2,\ 0.3,\ 0.3), \quad T = \begin{pmatrix} -5 & 0.5 & 0 & 1 \\ 0.5 & -4 & 0.5 & 0 \\ 0 & 1 & -3 & 1 \\ 1 & 0 & 1 & 2 \end{pmatrix}. \tag{48}$$

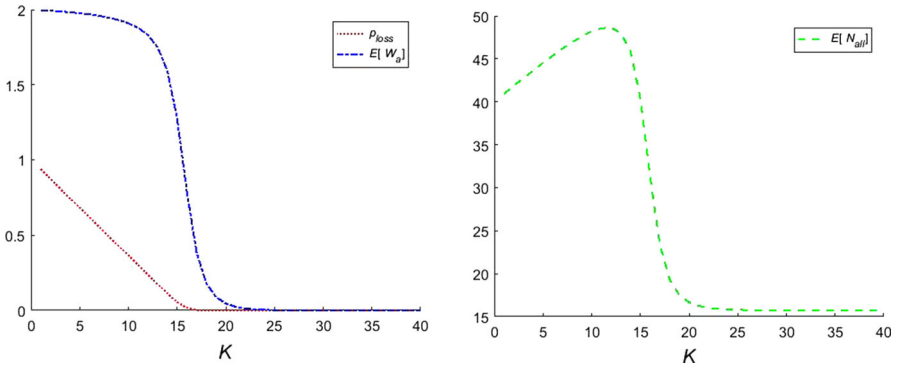

Fig. 5 Performance measures for K from 1 to 40

For this example, $m_s = 4$ and $m_e = 1$. We plot performance measures for K from 1 to 40 in Fig. 5. The queueing behaviour is quite similar to that of Example 7.2.

8 Discussion

To end this paper, we would like to discuss several issues.

- First, we would like to point out that the solution to (15) can be constructed by using different sets of $\{R, G\}$. For instance, matrices $\{\hat{R}^{-1}, \hat{G}^{-1}\}$, defined in the proof of Proposition 4.2, also satisfy Eq. (17) if $\{\hat{R}, \hat{G}$ are the minimal nonnegative solutions to Eq. (24). Replacing $\{R, G\}$ with $\{\hat{R}^{-1}, \hat{G}^{-1}\}$ in all formulas in Sect. 3, we obtain the same solution $\{\mathbf{p}_0, \mathbf{p}_1, \ldots, \mathbf{p}_K, \mathbf{p}_{K+1}(x)\}$. Different from the solution from Eq. (17), both $\hat{R}^{-1}$ and $\hat{G}^{-1}$ are invertible. However, for this case, the matrix $\lambda(\hat{G}^{-1}+I)+A_1$ is always non-invertible (see Proposition 4.2). Thus, with respect to numerical computation, we don't see immediately the advantage of $\{\hat{R}^{-1}, \hat{G}^{-1}\}$. The use of $\{\hat{R}^{-1}, \hat{G}^{-1}\}$ is an interesting issue for further study, though.
- Second, recall that the number of states in $\Omega(K)$ is $m_e(K+m_s-1)!/(K!(m_s-1)!)$. If K is big (e.g., $K = 100$), a small increase in m_s leads to a huge increase in the number of states in $\Omega(K)$. Consequently, as indicated by Table 2, the algorithm developed in this paper works well only for small m_s. Alternative approach is required to analyze the queueing model with a moderate or big m_s.
- Third, an immediate extension of the *M/PH/K* queue with customer abandonment is the *MAP/PH/K* model. Technically, the analysis of the *MAP/PH/K* model requires the commutability of some matrices involved in a set of equations similar to Eq. (10). Thus, a new approach has to be introduced to analyze the *MAP/PH/K* extension.
- Finally, the customer impatient time is assumed to be constant in this paper. Needless to say that the generalization to models with a non-constant impatient time is interesting and important. Again, a new approach has to be introduced to analyze such queueing models. Research in those directions is undergoing.

Acknowledgements The authors would like to thank three anonymous reviewers and the associate editor for their insightful comments and suggestions on this paper. The authors would also like to thank Dr. Stan Dimitrov for sharing computing resource with us.

Appendix A: transition blocks for the count-server-for-phase approach

In this appendix, we construct the transition blocks in (3) explicitly. First, we need to specify how states in $\Omega(0) \cup \Omega(1) \cup \ldots \cup \Omega(K)$ are organized. In general, we define, for $k = 0, 1, \ldots, K$, and $m = 1, 2, \ldots, K$,

$$\Omega(k, m) = \left\{ (n_1, \ldots, n_m) : \quad \text{integer } n_i \geq 0, \ i = 1, 2, \ldots, m, \ \sum_{i=1}^{m} n_i = k \right\}. \quad (49)$$

Note that $\Omega(k) = \Omega(k, m_s)$, for $k = 0, 1, 2, \ldots, K$. We organize the states in $\Omega(k, m)$ lexicographically. Then we have

$$\Omega(k, m) = \cup_{i=0}^{k} (\Omega(k - i, m - 1) \times \{i\}). \quad (50)$$

It is easy to see that, for $m = 1$, we have $\Omega(k, 1) = \{k\}$, and for $k = 0$, $\Omega(0, m) = \{(0, \ldots, 0)\}$.

We begin with $\{S^+(k, m_s), k = 0, 1, \ldots, K–1\}$. The basic components to construct those matrices are $\{\lambda, \boldsymbol{\beta} = (\beta_1, \ldots, \beta_{m_s})\}$, since the corresponding transitions are triggered by the arrival of a customer. The vector $\boldsymbol{\beta}$ has to be utilized to specify the service phase of the arriving customer. An effective way to construct those matrices is to generate them iteratively. To that end, we need to construct matrices $\{S^+(k, m), k = 0, 1, \ldots, K–1, m = 1, 2, \ldots, m_s\}$ for transitions from $\Omega(k, m)$ to $\Omega(k+1, m)$. We further decompose the transitions into transitions from $\{\Omega(k, m–1)\times\{0\}, \Omega(k–1, m–1)\times\{1\}, \ldots, \Omega(0, m–1)\times\{k\}\}$ to $\{\Omega(k+1, m–1)\times\{0\}, \Omega(k, m–1)\times\{1\}, \ldots, \Omega(0, m–1)\times\{k+1\}\}$, respectively. Specifically, for $S^+(k, m)$, the construction components are $\{\lambda\beta_1, \ldots, \lambda\beta_m\}$, and $S^+(k, m)$ is given by

$$\begin{array}{c} \begin{array}{cccccc} \Omega(k+1,m-1)\times\{0\} & \Omega(k,m-1)\times\{1\} & \cdots & \Omega(1,m-1)\times\{k\} & \Omega(0,m-1)\times\{k+1\} \end{array} \\ \begin{pmatrix} S^+(k,m-1) & \lambda\beta_m I & & & \\ & S^+(k-1,m-1) & \lambda\beta_m I & & \\ & \ddots & \ddots & \ddots & \\ & & & S^+(1,m-1) & \lambda\beta_m I \\ & & & & S^+(0,m-1) \quad \lambda\beta_m \end{pmatrix} \end{array} \quad (51)$$

and $S^+(0, m) = \lambda(\beta_1, \ldots, \beta_m)$, for $m = 1, 2, \ldots, m_s$, and $S^+(k, 1) = \lambda\beta_1$, for $k = 0, 1, \ldots, K–1$.

Now, we construct $\{S^-(k, m), k = 1, 2, \ldots, K, m = 1, 2, \ldots, m_s\}$, which are for transitions from $\Omega(k, m)$ to $\Omega(k–1, m)$. The corresponding transitions are triggered by a service completion. Thus, the construction is based on $\{t_1^0, t_2^0, \ldots, t_{m_s}^0\}$ (Note: Recall that $\mathbf{T}^0 = \left(t_j^0\right)_{m_s \times 1}$), and we obtain $S^-(k, m)$ as

$$
\begin{array}{c}
\begin{matrix}
\Omega(k-1,m-1)\times\{0\} & \Omega(k-2,m-1)\times\{1\} & \cdots & \Omega(1,m-1)\times\{k-2\} & \Omega(0,m-1)\times\{k-1\}
\end{matrix} \\
\begin{pmatrix}
S^-(k,m-1) & & & & \\
t_m^0 I & S^-(k-1,m-1) & & & \\
& \ddots & \ddots & & \\
& & \ddots & \ddots & \\
& & & (k-1)t_m^0 I & S^-(1,m-1) \\
& & & & kt_m^0
\end{pmatrix}
\end{array}
\tag{52}
$$

and $S^-(1, m) = (t_1^0,\ t_2^0, \ldots, t_m^0)\prime$, for $m = 1, 2, \ldots, m_s$, and $S^-(k, 1) = kt_1^0$, for $k = 1, 2, \ldots, K$.

Finally, we construct $\{S(k, m), k = 0, 1, \ldots, K, m = 1, 2, \ldots, m_s\}$, which are for transitions from $\Omega(k, m)$ to $\Omega(k, m)$. We decompose the transitions into three types, based on the decomposition of the states in $\Omega(k, m)$:

(i) Type "+": transitions from $\Omega(j, m–1)\times\{k–j\}$ to $\Omega(j+1, m–1)\times\{k–j–1\}$;
(ii) Type "–": transitions from $\Omega(j, m–1)\times\{k–j\}$ to $\Omega(j–1, m–1)\times\{k–j+1\}$; and
iii) Transitions from $\Omega(j, m–1)\times\{k–j\}$ to $\Omega(j, m–1)\times\{k–j\}$.

Then the transition matrix $S(k, m)$ from $\Omega(k, m)$ to $\Omega(k, m)$ can be written as

$$
\begin{array}{c}
\begin{matrix}
\Omega(k,m-1)\times\{0\} & \Omega(k-1,m-1)\times\{1\} & \cdots & \Omega(1,m-1)\times\{k-1\} & \Omega(0,m-1)\times\{k\}
\end{matrix} \\
\begin{pmatrix}
S(k,m-1) & \hat{S}^-(k,m-1) & & & \\
\hat{S}^+(k-1,m-1) & S(k-1,m-1) & \hat{S}^-(k-1,m-1) & & \\
& \ddots & \ddots & \ddots & \\
& & (k-1)\hat{S}^+(1,m-1) & S(1,m-1) & \hat{S}^-(1,m-1) \\
& & & k\hat{S}^+(0,m-1) & S(0,m-1)
\end{pmatrix} \\
+ \begin{pmatrix}
0 & & & \\
& t_{m,m} I & & \\
& & \ddots & \\
& & & (k-1)t_{m,m} I \\
& & & & kt_{m,m}
\end{pmatrix}.
\end{array}
\tag{53}
$$

If $m = 1$, we have $S(k, 1) = kt_{1,1}$, for $k = 0, 1, 2, \ldots, K$. We need to construct two sets of matrices $\{\hat{S}^+(k, m), k = 1, 2, \ldots, K, m = 1, 2, \ldots, m_s–1\}$ and $\{\hat{S}^-(k, m), k = 0, 1, \ldots, K–1, m = 1, 2, \ldots, m_s–1\}$. Note that $\{\hat{S}^+(k, m), k = 1, 2, \ldots, K, m = 1, 2, \ldots, m_s–1\}$ are for the transitions from phase m to phases $\{1, 2, \ldots, m–1\}$, and $\{\hat{S}^-(k, m), k = 1, 2, \ldots, K, m = 1, 2, \ldots, m_s–1\}$ are for the transitions from phases $\{1, 2, \ldots, m–1\}$ to phase m. We use the construction methods for $\{S^+(k, m), k = 0, 1, \ldots, K–1, m = 1, 2, \ldots, m_s\}$ and $\{S^-(k, m), k = 1, 2, \ldots, K, m = 1, 2, \ldots, m_s\}$ in this construction.

(i) The construction of $\{\hat{S}^+(k, m), k = 1, 2, \ldots, K, m = 1, 2, \ldots, m_s - 1\}$ is similar to that of $\{S^+(k, m), k = 0, 1, \ldots, K–1, m = 2, 3, \ldots, m_s\}$, except that $\{\lambda\beta_1, \ldots, \lambda\beta_m\}$ is replaced with $\{t_{m+1,1}, t_{m+1,2}, \ldots, t_{m+1,m}\}$. In addition, we have $\hat{S}^+(0, m) = (t_{m+1,1}, \cdots, t_{m+1,m})$, for $m = 1, 2, \ldots, m_s–1$, and $\hat{S}^+(k, 1) = t_{2,1}$, for $k = 0, 1, \ldots, K–1$.

(ii) The construction of $\{\hat{S}^-(k,m),\ k = 0, 1, \ldots, K\text{–}1,\ m = 2, \ldots, m_s\text{–}1\}$ is similar to that of $\{S^-(k,m),\ k = 0, 1, \ldots, K\text{–}1,\ m = 2, 3, \ldots, m_s\}$, except that $\{t_1^0,\ t_2^0, \ldots, t_{m_s-1}^0\}$ is replaced with $\{t_{1,m+1}, t_{2,m+1}, \ldots, t_{m,m+1}\}$. In addition, we have $\hat{S}^-(1,m) = (t_{1,m+1},\ \ldots,\ t_{m,m+1})'$, for $m = 1, 2, \ldots, m_s\text{–}1$, and $\hat{S}^-(k,1) = kt_{1,2}$, for $k = 1, 2, \ldots, K$.

Finally, we summarize the above construction methods to outline the steps to construct $\{S^+(k,m_s),\ k = 0, 1, \ldots, K\text{-}1\}$, $\{S^-(k,m_s),\ k = 1, 2, \ldots, K\}$, and $\{S(k,m_s),\ k = 0, 1, 2, \ldots, K\}$.

Algorithm A.1 *Construction of transition blocks for the count-server-for-phase approach*

A.I.1 Compute $\{S^+(k,m_s),\ k = 0, 1, \ldots, K\text{–}1\}$*:*
- *(i)* $S^+(k,1) = \lambda\beta_1$*, for* $k = 0, 1, \ldots, K$*;*
- *(ii) Use Eq. (51) to construct* $\{S^+(k,m)$*, for* $k = 0, 1, \ldots, K\}$*, for* $m = 2, 3, \ldots, m_s$*.*

A.I.2 Compute $\{S^-(k,m_s),\ k = 1, 2, \ldots, K\}$*:*
- *(i)* $S^-(k,1) = kt_1^0$*, for* $k = 1, 2, \ldots, K$*;*
- *(ii) Use equation (52) to construct* $\{S^-(k,m)$*, for* $k = 0, 1, \ldots, K\}$*, for* $m = 2, 3, \ldots, m_s$*.*

A.I.3 Compute $\{S(k,m_s),\ k = 1, 2, \ldots, K\}$*:*
- *If* $m = 1$*, we have* $S(k,1) = kt_{1,1}$*, for* $k = 0, 1, 2, \ldots, K$*.*
- *For* $m = 2, 3, \ldots, m_s$*,*
  - *(i) Construct* $\{\hat{S}^+(k,j), k = 1, 2, \ldots, K, j = 1, 2, \ldots, m\text{–}1\}$ *using Eq. (51) with* $(t_{m+1,1}, t_{m+1,2}, \ldots, t_{m+1,m})$ *in place of* $\lambda(\beta_1, \ldots, \beta_m)$*.*
  - *(ii) Construct* $\{\hat{S}^-(k,m), k = 0, 1, \ldots, K\text{–}1, m = 2, \ldots, m_s - 1\}$ *using Eq. (52) with* $(t_{1,m+1}, t_{2,m+1}, \ldots, t_{m,m+1})'$ *in place of* $(t_1^0,\ t_2^0, \ldots, t_m^0)'$*.*
  - *(iii) Construct* $\{S(k,m), k = 0, 1, \ldots, K\}$*.*

end

Appendix B. Matrices *R* and *G*, and Vectors $\mathbf{u}_1$ and $\mathbf{u}_2$

The following solution approach was introduced in Choi et al. (2004), and used in solving the *M/PH*/1 case in Kim and Kim (2015). The theoretical basis for this solution approach is the following theorem (Theorem 2.15 and 2.16 in Gohberg et al. 1982).

Theorem B.1 *(Gohberg et al. (1982)) Consider second order matrix differential equation*

$$\frac{d^2}{dx^2}\mathbf{u}(x) + \frac{d}{dx}\mathbf{u}(x)B_1 + \mathbf{u}(x)B_2 = \mathbf{0}, \tag{54}$$

where $\boldsymbol{u}(x)$ *is the row vector function to be found, and* B_1 *and* B_2 *are matrices. Suppose that* X_1 *and* X_2 *are matrices that are solutions of the auxiliary equation*

$$X^2 + XB_1 + B_2 = \mathbf{0}. \tag{55}$$

If X_1 and X_2 have no common eigenvalues, then the general solution of Eq. (54) is given by

$$\mathbf{u}(x) = \mathbf{u}_1 \exp\{X_1 x\} + \mathbf{u}_2 \exp\{X_2 x\} \tag{56}$$

where $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$ are two constant vectors.

For our problem (15), we have

$B_1 = -(\lambda \mathrm{I} + Q \otimes I + M \otimes S(K, m_s))$ and

$B_2 = \lambda \mathrm{Q} \otimes I + \lambda \mathrm{M} \otimes (S(K, m_s) + S^-(K, m_s) S^+(K-1, m_s)/\lambda)$.

Let $X_1 = \lambda(I - R)$ and $X_2 = \lambda \mathrm{G} + Q \otimes I + M \otimes S(K, m_s)$, where R and G are defined in Sect. 3. It can be shown that the two matrices are solutions to Eq. (55). Then one can use Eq. (56) to find a solution to Eq. (15), which satisfies all boundary conditions. For that purpose, we need that $\exp\{X_1 x\}$ and $\exp\{X_2 x\}$ provides $2m_e m_s{}^K$ independent solutions to (15). To ensure that, we need the following results.

Proposition B.2 *If $\rho \neq 1$, matrices $\lambda(I - R)$ and $\lambda G + Q \otimes I + M \otimes S(K, m_s)$ have no common eigenvalues. If $\rho = 1$, matrices $\lambda(I - R)$ and $\lambda G + Q \otimes I + M \otimes S(K, m_s)$ have one common eigenvalue zero with algebraic multiplicity one.*

Proof First, we consider the case with $\rho < 1$. Since $\mathrm{sp}(R) = 1$, it is clear that the real parts of all eigenvalues of $\lambda(I - R)$ are nonnegative. By Proposition 4.2, the maximal real part of all eigenvalues of $\lambda \mathrm{G} + Q \otimes I + M \otimes S(K, m_s)$ is negative. Note that the eigenvalue of $\lambda \mathrm{G} + Q \otimes I + M \otimes S(K, m_s)$ with the maximal real part has to be real. Thus, the two matrices have no common eigenvalues.

If $\rho > 1$, we have $\mathrm{sp}(R) = \mathrm{sp}(G) < 1$. Then all eigenvalues of $\lambda(I - R)$ have a positive real part. On the other hand, by Proposition 4.2, all eigenvalues of $\lambda \mathrm{G} + Q \otimes I + M \otimes S(K, m_s)$ have a nonpositive real part. Therefore, the two matrices have no common eigenvalue.

If $\rho = 1$, zero is an eigenvalue of both $\lambda(I - R)$ and $\lambda \mathrm{G} + Q \otimes I + M \otimes S(K, m_s)$. Similar to Choi et al. (2004), it can be shown that the algebraic multiplicity of the two matrices is one. This completes the proof of Proposition B.2. □

By Proposition B.2, if $\rho \neq 1$, it can be shown that Eq. (56) gives $2m_e m_s{}^K$ independent solutions to (15). Then all we need to do is to find $\{\mathbf{u}_1, \mathbf{u}_2\}$ for a solution to (10). To do so, we use the function $\mathrm{d}\mathbf{p}_{K+1}(x)/\mathrm{d}x$, which can be found in two ways: using Eqs. (10) and (16). Equalizing the resulted expressions at $x = 0$ and $x = \tau-$ leads to the following linear system for $\{\mathbf{u}_1, \mathbf{u}_2\}$:

$$\begin{aligned}
\mathbf{0} = {} & \mathbf{u}_1 e^{\lambda(R-I)\tau} \left(D_K (I \otimes S^+(K-1, m_s)) \frac{1}{\lambda} - R \right) \\
& + \mathbf{u}_2 \left(G - I + \frac{1}{\lambda} (Q \otimes I + M \otimes S(K, m_s)) \right) \\
& + \mathbf{u}_2 \frac{1}{\lambda} \left(D_K (I \otimes S^+(K-1, m_s)) \right); \\
\mathbf{0} = {} & \mathbf{u}_1 (Q \otimes I + M \otimes S(K, m_s) + \lambda R) \\
& + \mathbf{u}_2 \lambda e^{(\lambda G + Q \otimes I + M \otimes S(K, m_s))\tau} (I - G).
\end{aligned} \tag{57}$$

Finally, we use the law of total probability to normalize $\{\mathbf{u}_1, \mathbf{u}_2\}$, i.e., $\sum_{k=0}^{K} \mathbf{p}_k \mathbf{e} + \int_0^\tau \mathbf{p}_{K+1}(x)\mathrm{d}x\mathbf{e} = 1$. By routine calculations, we obtain

$$\begin{aligned}\sum_{k=0}^{K} \mathbf{p}_k \mathbf{e} &= \mathbf{p}_K \left(I + \sum_{k=0}^{K-1} \left(\prod_{j=0}^{K-(k+1)} D_{K-j} \right) \right) \mathbf{e} \\ &= (\mathbf{u}_1 \exp\{\lambda(R-I)\tau\} + \mathbf{u}_2) \frac{1}{\lambda} \left(I + \sum_{k=0}^{K-1} \left(\prod_{j=0}^{K-(k+1)} D_{K-j} \right) \right) \mathbf{e}, \end{aligned} \tag{58}$$

and

$$\begin{aligned}&\int_0^\tau \mathbf{p}_{K+1}(x)\mathrm{d}x \\ &= \mathbf{u}_1 \int_0^\tau e^{\{\lambda(R-I)(\tau-x)\}}\mathrm{d}x + \mathbf{u}_2 \int_0^\tau \lambda e^{(\lambda G + Q\otimes I + M \otimes S(K,m_s))x}\mathrm{d}x. \end{aligned} \tag{59}$$

Using properties given in Sect. 4, the integrals in Eq. (59) can be obtained.

Proposition B.3 *We have, for $\rho < 1$,*

$$\begin{aligned}&\int_0^\tau \exp\{\lambda(R-I)(\tau-x)\}dx \\ &= \frac{1}{\lambda}\left(e^{\lambda(R-I)\tau} - I + \lambda\tau\xi(\boldsymbol{\pi}\otimes\boldsymbol{\phi}) \right) (R - I + \boldsymbol{\xi}(\boldsymbol{\pi}\otimes\boldsymbol{\phi}))^{-1}; \end{aligned} \tag{60}$$

$$\begin{aligned}&\int_0^\tau e^{(\lambda G + Q\otimes I + M\otimes S(K,m_s))x}dx \\ &= \left(e^{(\lambda G + Q\otimes I + M\otimes S(K,m_s))x} - I \right) (\lambda G + Q\otimes I + M\otimes S(K,m_s))^{-1}, \end{aligned} \tag{61}$$

for $\rho > 1$,

$$\int_0^\tau \exp\{\lambda(R-I)(\tau-x)\}dx = \frac{1}{\lambda}\left(e^{(R-I)\tau} - I \right) (R-I)^{-1}; \tag{62}$$

$$\begin{aligned}&\int_0^\tau e^{(\lambda G + Q\otimes I + M\otimes S(K,m_s))x}dx \\ &\quad = \left(e^{(\lambda G + Q\otimes I + M\otimes S(K,m_s))\tau} - I + \tau\boldsymbol{\zeta}(\boldsymbol{\pi}\otimes\boldsymbol{\phi}) \right) \\ &\qquad \times (\lambda G + Q\otimes I + M\otimes S(K,m_s+) + \boldsymbol{\zeta}(\boldsymbol{\pi}\otimes\boldsymbol{\phi}))^{-1}, \end{aligned} \tag{63}$$

and, for $\rho = 1$, Eqs. (60) and (63) hold.

Proof The proof is based on Propositions 4.3 and 4.4. Details are omitted.

By Proposition B.3, the linear system with (19), (20), and (21) for $\{\mathbf{u}_1, \mathbf{u}_2\}$ is obtained for the case with $\rho \neq 1$.

If $\rho = 1$, by Proposition B.2, (56) gives $2m_e m_s{}^K - 1$ independent solutions to (15). We need to find one more solution to (15). By Proposition 4.2, it is easy to verify that

$$\mathbf{v}(x) = (\boldsymbol{\pi} \otimes \boldsymbol{\phi}) \left(\lambda x I + (A_0 - A_2) (A_0 + A_1 + A_2 + \mathbf{e} \otimes (\boldsymbol{\pi} \otimes \boldsymbol{\phi}))^{-1} \right) \quad (64)$$

is another independent solution to (15). Then the solution to (15) can be expressed as

$$\mathbf{u}(x) = \mathbf{u}_1 \exp\{X_1 x\} + \mathbf{u}_2 \exp\{X_2 x\} + u_3 \mathbf{v}(x), \quad (65)$$

where u_3 is a constant. Similar to the case with $\rho \neq 1$, a linear system for $\{\mathbf{u}_1, \mathbf{u}_2, u_3\}$ can be established by using two boundary conditions (i.e., conditions at $x = 0$ and $x = \tau-$) and the law of total probability. Once $\{\mathbf{u}_1, \mathbf{u}_2, u_3\}$ is obtained, a solution to (15) can be obtained. Details are omitted.

References

Asmussen S, O'Cinneide CA (1998) Representations for matrix-geometric and matrix-exponential steady-state distributions with applications to many-server queues. Stoch Models 14:369–387

Baccelli F, Boyer P, Hebuterne G (1984) Single server queues with impatient customers. Adv Appl Probab 16:887–905

Barrer DY (1957a) Queuing with impatient customer and indifferent clerks. Oper Res 5:644–649

Barrer DY (1957b) Queuing with impatient customers and ordered service. Oper Res 5:650–656

Boxma OJ, Waal PR (1994) Multisever queue with impatient customers. In: The fundamental role of teletraffic in the evolution of telecommunication network (Proc. ITC14). North-Holland, Amsterdam, pp 743–756

Boots NK, Tijms H (1999) An M/M/c queue with impatient customers. TOP 7:213–220

Brandt A, Brandt M (1999a) On the M(n)/M(n)/s queues with impatient call. Perform Eval 35:1–18

Brandt A, Brandt M (1999b) On a two-queue priority system with impatience and its application to a call center. Methodol Comput Appl Probab 1:191–210

Brandt A, Brandt M (2002) Asymptotic result and a Markovian approximation for the M(n)/M(n)/s + GI system. Queueing Syst 41:73–94

Choi BD, Kim B, Chung J (2001) M/M/1 queue with impatient customers of high priority. Queueing Syst 38:49–66

Choi BD, Kim B, Zhu D (2004) MAP/M/c queue with constant impatient time. Math Oper Res 29:309–325

Dai JG, He S (2010) Customer abandonment in many-server queues. Math Oper Res 35:347–362

Dai JG, He S (2011) Queues in service systems: customer abandonment and diffusion approximation. Tutor Oper Res 2011:36–59

Dai JG, Tezcan T (2008) Optimal control of parallel server system with many servers in heavy traffic. Queueing Syst 59:95–134

Dai JG, He S, Tezcan T (2010) Many-server diffusion limits for G/Ph/n+GI queues. Ann Appl Probab 20:1854–1890

Daley DJ (1965) General customer impatience in the queue GI/G/1. J Appl Probab 2:186–209

de Kok AG, Tijms HC (1985) A queueing system with impatient customers. J Appl Probab 22:688–696

Dzial T, Breuer L, da Silva Soares A, Latouche G, Remiche M (2005) Fluid queues to solve jump processes. Perform Eval 62:132–146

Finch PD (1960) Deterministic customer impatience in the queueing system GI/M/1. Biometrika 47(1,2): 45-52

Garnett O, Mandelbaum A, Reiman M (2002) Designing a call center with impatient customers. Manuf Serv Oper Manag 4(3):208–227

Gohberg I, Lancaster P, Rodman L (1982) Matrix polynomials. Academic Press, New York

Harris TE (1956) The existence of stationary measures for certain Markov processes. In: Proceedings of 3rd Berkeley symposium, Vol II:113-124

He QM (2005) Age process, workload process, sojourn times and waiting time in a discrete time SM[K]/PH[K]/1/FCFS queue. Queueing Syst 49:363–403
He QM (2014) Fundamentals of matrix-analytic methods. Springer, New York
Jurkevic OM (1970) On the investigation of many-server queueing systems with bounded waiting time. Izv. Akad. Nauk SSSR Techniceskaja Kibernetika (Russian) 5:50–58
Jurkevic OM (1971) On many-server systems with stochastic bounds for the waiting time. Izv. Akad. Nauk SSSR Techniceskaja Kibernetika (Russian) 4:39–46
Kawanishi K, Takine T (2016) MAP/M/c and M/PH/c queues with constant impatience times. Queueing Syst 82:381–420
Kim B, Kim J (2015) A single server queue with Markov modulated service rates and impatient customers. Perform Eval 83–84:1–15
König D, Schmidt V (1990) Extended and conditional versions of the PASTA property. Adv Appl Probab 22:510–512
Latouche G (1987) A note on two matrices occurring in the solution of quasi-birth-and-death processes. Stoch Models 3:251–257
Latouche G, Ramaswami V (1999) Introduction to matrix analytic methods in stochastic modeling. ASA-SIAM Series on Statistic and Applied Probability, SIAM, Philadelphia, PA
Mandelbaum A, Zeltyn S (2013) Data-stories about (im) patient customers in tele-queues. Queueing Syst 75(2–4):115–146
Meini B (2013) On the numerical solution of a structured nonsymmetric algebraic Riccati equation. PEVA 70(9):682–690
Movaghar A (1998) On queueing with customer impatience until the beginning of service. Queueing Syst 29:337–350
Neuts MF (1981) Matrix-geometric solution in stochastic model: an algorithmic application. The Johns Hopkins University Press, Baltimore, MD
Ramaswami V (1985) Independent Markov processes in parallel. Stoch Models 1:419–432
Ramaswami V, Lucantoni DM (1985) Algorithms for the multi-server queue with phase type service. Stoch Models 1:393–417
Stanford RE (1990) On queues with impatience. Adv Appl probab 22:768–769
Van Houdt B (2012) Analysis of the adaptive MMAP[K]/PH[K]/1 queue: a multi-type queue with adaptive arrivals and general impatience. Eur J Oper Res 220:695–704
Xiong W, Jagerman D, Altiok T (2008) M/G/1 queue with deterministic reneging times. Perform Eval 65:308–316