Queueing Systems 49, 363–403, 2005

Age Process, Workload Process, Sojourn Times, and Waiting Times in a Discrete Time *SM*[*K*]/*PH*[*K*]/1/FCFS Queue

QI-MING HE Qi-Ming.He@dal.ca
Department of Industrial Engineering, Dalhousie University, Halifax, Nova Scotia, Canada B3J 2X4; School of Economics and Management, Tsinghua University, Beijing, P.R. China

Received 6 August 2004; Revised 30 December 2004

Abstract. In this paper, we study a discrete time queueing system with multiple types of customers and a first-come-first-served (FCFS) service discipline. Customers arrive according to a semi-Markov arrival process and the service times of individual customers have PH-distributions. A *GI*/*M*/1 type Markov chain for a generalized age process of batches of customers is introduced. The steady state distribution of the *GI*/*M*/1 type Markov chain is found explicitly and, consequently, the steady state distributions of the age of the batch in service, the total workload in the system, waiting times, and sojourn times of different batches and different types of customers are obtained. We show that the generalized age process and a generalized total workload process have the same steady state distribution. We prove that the waiting times and sojourn times have PH-distributions and find matrix representations of those PH-distributions. When the arrival process is a Markov arrival process with marked transitions, we construct a QBD process for the age process and the total workload process. The steady state distributions of the waiting times and the sojourn times, both at the batch level and the customer level, are obtained from the steady state distribution of the QBD process. A number of numerical examples are presented to gain insight into the waiting processes of different types of customers.

Keywords: queueing systems, *GI*/*M*/1 type Markov chain, waiting times, workload, matrix analytic methods, semi-Markov chain

AMS subject classification: 60K25, 60J10

1. Introduction

Modern communication networks are required to handle different types of data with drastically different characteristics in burstiness and volume. Modern supply chains of goods are designed to meet customized demands with different features. Motivated by the design and performance analysis of these stochastic systems, in this paper, we introduce and study a class of discrete time queueing models with multiple types of customers. The focus is on the distributions of waiting times and sojourn times of individual types of customers.

Because of various applications, the study of discrete time queueing models was extensive (e.g., Alfa [1], Cinlar [6], Cortizo et al. [8], De Smit [9], Grassmann and

Jian [14], Van Houdt and Blondia [34–37], Yang and Chaudhry [39], and references therein). In the classical literature, most of the studies on discrete time queues were focused on queues with a single type of customers or multiple types of customers with different service priorities. The arrival processes of customers are usually assumed to be (independent) Poisson processes. With the introduction of the Markov arrival process with marked transitions (*MMAP*[*K*]) (Asmussen and Koole [4], HE [15,17], HE and Neuts [21]), complicated queues with multiple types of customers becomes tractable (HE [16], HE and Alfa [20], Takine [31,32], and Takine and Hasegawa [33], etc.) In this paper, we use a semi-Markov arrival process with marked transitions, which is more general than *MMAP*[*K*]. We assume no priority among different types of customers. The queueing model is still tractable mathematically.

Queueing systems with multiple types of customers without priority can be categorized by the service discipline of individual customers. In HE [16], HE and Alfa [20], and Takine and Hasegawa [33], queueing systems with a last-come-first-served service discipline were investigated. By analyzing Markov chains with a tree structure, results were obtained for performance measures such as busy periods, queue strings, and waiting times. For the first-come-first-served case, the distributions of waiting times of individual types of customers can be obtained (HE [15,17] and Takine [31]). But the queue string (length) process for the first-come-first-served case is difficult to analyze since one has to keep track of the types of customers in queue. In Takine [31], the queue length distributions were found by using information about the waiting times. In general, for the queueing system of interest, unlike the analysis of the waiting times in some classical queueing systems, there is no information about the queue lengths to be utilized in the analysis of the waiting times. Fortunately, a Markov chain associated with the age of the batch of customers in service can be constructed so as to analyze the waiting times and sojourn times, which is the approach used in this paper.

Waiting times and sojourn times are among the most important performance measurements of queueing models. The study of waiting times and sojourn times in queues was extensive (e.g., Asmussen and O'Cinneide [3], Cinlar [6], Cohen [7], De Smit [9], Grassmann and Jian [14], Neuts [25], Sengupta [28,30], Takine and Hasegawa [33], etc.) In Grassmann and Jian [14], a method based on Wiener-Hopf factorization was applied to develop an algorithm for computing the waiting time distribution (for a discrete time queueing model with a single class of customers). Sengupta [28–30] showed that the waiting time and sojourn time in a continuous time *GI*/*PH*/1 queue have matrix exponential distributions. Asmussen and O'Cinneide [3] extended Sengupta's results to a continuous time *GI*/*PH*/*c* queue. Recently, Van Houdt and Blondia [35–37] studied the waiting times and sojourn times in a discrete time queue with multiple types of customers (the model studied in Section 5 of this paper). In [37], they constructed a Markov chain similar to the one used in this paper for the age process and obtained the distributions of sojourn times from the steady state distribution of the age process. In this paper, we consider a more general queueing model in discrete time. We construct a generalized age process, which has no boundary at level zero, and derive the distributions of the waiting times and sojourn times from the steady state distribution of the generalized age process. Furthermore, we show that the distributions of waiting times and sojourn times for individual types of customers are discrete time PH-distributions.

Matrix analytic methods are the basic tools used in this paper. We refer readers to Latouche and Ramaswami [23] and Neuts [26,27] for more about the matrix analytic methods. One of the advantages of such methods is the development of efficient algorithms for computing performance measures for various stochastic models, by using the structures of Markov chains of of interest (e.g., the QBD, *M*/*G*/1, or *GI*/*M*/1 structure). In this paper, we shall construct a *GI*/*M*/1 type Markov chain and a QBD process associated with the age of the batch in service to analyze the waiting times and sojourn times. In HE [18], an *MMAP*[*K*]/*SM*[*K*]/1/FCFS qeueueing system is considered and an *M*/*G*/1 type Markov chain associated with the total workload is constructed in analysis. By using existing results of these well structured Markov chains, we are able to obtain detailed results on the distributions of waiting times and sojourn times.

As mentioned above, the basic idea in this paper was to construct a *GI*/*M*/1 type Markov chain, that is associated with the age of customers in service, to study the waiting times and sojourn times of individual types of customers. The idea was used for analyzing waiting times in queues in the past. For continuous time queues, Asmussen and O'Cinneide [3] and Sengupta [28,30] used this idea to find the waiting time distribution for an *SM*/*PH*/1 queue. In [3,28,30], the queueing models have only one type of customers. Their extension to continuous queueing models with multiple types of customers is considered in HE [19]. Van Houdt and Blondia [35–37] used the idea to find the distributions of waiting times of the discrete time *MMAP*[*K*]/*PH*[*K*]/(1,2) queue. Compared to the work of Van Houdt and Blondia [35–37], this paper studies a more general queueing model, provides more details for the steady state distributions of the age process and the sojourn time distributions, and provides a formal introduction of the Markov chain associated with the age of the batch in service. The formal construction of the Markov chain gives insight into the solution approach and sheds light on the extension to queues with multiple servers.

The rest of the paper is organized as follows. In Section 2, the discrete time *SM*[*K*]/*PH*[*K*]/1/FCFS queue is introduced. In Section 3, a generalized age process is introduced and analyzed. Particularly, the steady state distribution of that process is found. Based on that steady state distribution, in Section 4, distributions of the age of the batch in service, the total workload in the system, sojourn times and waiting times of different types of batches and customers are found. It is shown that these distributions are PH-distributions. In Section 5, a special queueing model whose arrival process is Markovian is analyzed. More detailed results are obtained for this case. Numerical examples are presented in Section 6 to gain insight into the methods developed and the performance of queueing systems. Finally, in Section 7, we offer a brief discussion on the extension to queues with multiple servers.

2. The *SM*[*K*]/*PH*[*K*/1/FCFS queue

The queueing system of interest has K types of customers, where K is a positive integer. All customers, regardless of their types, join a single queue and are served by a single server on a first-come-first-served (FCFS) basis. In the rest of this section, we describe the customer arrival process, the queueing process, and the service times in detail.

2.1. The customer arrival process

The customer arrival process is a discrete time semi-Markov process with marked transitions. Customers are distinguished into K types and arrive in batches. To characterize the batches of customers, we define a set of strings of integers:

$$\aleph = \{J_k\colon J_k = j_1j_2\dots j_{n_k},\ 1 \le j_i \le K,\ 1 \le i \le n_k,\ 1 \le k \le N\} \tag{2.1}$$

where N is the total number of different strings in set $\aleph$ and n_k is the number of customers in the k-th batch. We assume that N is finite. For the arrival process, a string $J = j_1j_2\dots j_n \in \aleph$ represents a batch that has n customers. These n customers are of types $j_1, j_2, \dots$, and j_n, respectively. We call J a string representation of that batch. Thus, there are in total N different types of batches.

Consider a semi-Markov chain $\{(\xi_n, \tau_n), n \ge 0\}$ with m_a phases. The variable ξ_n is the phase of a semi-Markov chain right after the n-th transition. The variable τ_n is the number of periods between the $(n-1)$-st transition and the n-th transition (i.e., the inter-transition time). The arrivals of batches of customers are associated with transitions of the semi-Markov process in the following manner. Let J_n be the string representation of the batch associated with the n-th transition. That is: a batch J_n arrives at that transition epoch. Define

$$\begin{aligned} &P\{\xi_n = j, \tau_n = t, J_n = J \mid \xi_{n-1} = i\} = p_{J,i,j}(t), \\ &\quad 1 \le i, j \le m_a, \quad n \ge 1, \quad J \in \aleph, \end{aligned} \tag{2.2}$$

where t is a positive integer. The variable $p_{J,i,j}(t)$ is the probability that a batch J arrives after t periods of time from the arrival of the last batch and the phase of the underlying semi-Markov process becomes j after the arrival, given that the phase was i. Let $D_{a,J}(t)$ be an $m_a \times m_a$ matrix with (i, j)-th element $p_{J,i,j}(t)$. Matrices $\{D_{a,J}(t), t \ge 1, J \in \aleph\}$ provide all information about the semi-Markov arrival process with marked transitions. Define

$$\begin{aligned} &D_a(t) = \sum_{J\in\aleph} D_{a,J}(t), \quad t \ge 1; \quad D_{a,J} = \sum_{t=1}^{\infty} D_{a,J}(t), \quad J \in \aleph; \\ &D_a = \sum_{J\in\aleph} D_{a,J} = \sum_{t=1}^{\infty} D_a(t). \end{aligned} \tag{2.3}$$

The matrix D_a is the transition probability matrix of the embedded Markov chain of the semi-Markov chain $\{(\xi_n, \tau_n), n \geq 0\}$ at transition epochs. We assume that D_a is irreducible. Let $\boldsymbol{\theta}_a$ be the invariant probability vector of the stochastic matrix D_a, i.e., $\boldsymbol{\theta}_a D_a = \boldsymbol{\theta}_a$ and $\boldsymbol{\theta}_a \mathbf{e} = 1$, where $\mathbf{e}$ is a column vector with all elements being one.

In steady state, the inter-transition time of the semi-Markov process (i.e., the interarrival time of batches) can be calculated as follows:

$$E_{\boldsymbol{\theta}_a}(\tau) = \boldsymbol{\theta}_a \left(\sum_{t=1}^{\infty} t D_a(t) \right) \mathbf{e}. \tag{2.4}$$

The arrival rate of batches of customers is given as $\lambda = (E_{\boldsymbol{\theta}_a}(\tau))^{-1}$, i.e., the average number of batches arrived per period. The probability that an arbitrary batch is of the type J is $\boldsymbol{\theta}_a D_{a,J} \mathbf{e}$ for $J \in \aleph$. The arrival rate of type J batches is given by $\lambda_J = \lambda \boldsymbol{\theta}_a D_{a,J} \mathbf{e}$, i.e., the average number of type J batches arrived per period. The arrival rate of type k customers is given by

$$\lambda_{(k)} = \sum_{J \in \aleph} N(J, k) \lambda_J, \qquad 1 \leq k \leq K, \tag{2.5}$$

where $N(J, k)$ is the number of appearances of the integer k in the string J. Note that λ_k is the arrival rate of the batch k, and $\lambda_{(k)}$ is the arrival rate of type k customers. Without loss of generality, throughout this paper, we assume that λ, $\{\lambda_J, J \in \aleph\}$, and $\{\lambda_k, 1 \leq k \leq K\}$ are positive and finite. Results about the arrival rates and equations (2.4) and (2.5) can be shown by using the classical generating function approach. Details are omitted. For more about the construction of Markov arrival processes, see HE [15,17], HE and Neuts [21], Sengupta [30], and Takine [32]. An important special case is the Markov arrival process with marked transitions (*MMAP*[*K*]), which will be introduced and used in Section 5.

2.2. *The queueing process*

After a batch of customers arrived, all customers join a single queue according to the order in the batch. All batches are served by a single server on a FCFS basis. Within each batch, customers are served on their order in the batch. Let $q(t)$ be a string of integers for the queue in period t, which is obtained after possible service completion and arrival in period $t - 1$. If $q(t) = j_1 j_2 \ldots j_n$, then there are n customers in the system at time t, the customer in service is of type j_1, the first customer waiting in queue is of type $j_2, \ldots$, and the last customer waiting in queue is of type j_n. These n customers shall be served in the order $j_1, j_2, \ldots,$ and j_n. If a type J batch arrives next, the queue becomes $q(t) + J$. If the service is completed next, the queue becomes $j_2 j_3 \ldots j_n$ and the server starts serving customer j_2. According to HE [17], the service order of customers within a batch can be specified easily by arranging the integers in the corresponding string. For instance, if type 1 customers have service priority over other customers within a batch J, then the integer 1 is placed ahead of others in the string J (e.g., $J = 1114423$). In fact, HE [17] has shown that the string representation provides great flexibility with respect to the service order within a batch of customers.

2.3. Service times

The service times of individual customers have discrete time PH-distributions and are independent of each other and of the arrival process. For a type k customer, its service time s_k has a PH-distribution with a matrix representation $\{m_k, \boldsymbol{\alpha}_k, T_k\}$, where m_k is the number of phases of the PH-distribution, $\boldsymbol{\alpha}_k$ is the initial probability vector, and T_k is a substochastic matrix. We assume that the service time of any customer is at least one, i.e., $\boldsymbol{\alpha}_k\mathbf{e} = 1$, $1 \le k \le K$. See Neuts [26] for more about PH-distribution. Denote by $\mathbf{T}_k^0 = (I - T_k)\mathbf{e}$, where I is the identity matrix. We assume that each matrix representation of PH-distributions is irreducible, i.e., $T_k + \mathbf{T}_k^0\boldsymbol{\alpha}_k$ is irreducible. The service time of a batch is the sum of the service time of all customers within the batch. Since the set of PH-distributions is closed under convolution, the service time s_J of a type J batch also has a discrete time PH-distribution with a matrix representation $\{m_J, \boldsymbol{\alpha}_J, T_J\}$, where, for $J = j_1 j_2 \ldots j_n$,

$$m_J = \sum_{i=1}^{n} m_{j_i}; \quad \boldsymbol{\alpha}_J = \left(\boldsymbol{\alpha}_{j_1}, 0, \ldots, 0\right);$$

$$T_J = \begin{pmatrix} T_{j_1} & \mathbf{T}_{j_1}^0\boldsymbol{\alpha}_{j_2} & & & \\ & T_{j_2} & \mathbf{T}_{j_2}^0\boldsymbol{\alpha}_{j_3} & & \\ & & \ddots & \ddots & \\ & & & T_{j_{n-1}} & \mathbf{T}_{j_{n-1}}^0\boldsymbol{\alpha}_{j_n} \\ & & & & T_{j_n} \end{pmatrix}, \qquad \mathbf{T}_J^0 = \begin{pmatrix} 0 \\ \vdots \\ \vdots \\ 0 \\ \mathbf{T}_{j_n}^0 \end{pmatrix}. \tag{2.6}$$

The mean service time of a type k customer is given by $E(s_k) = \boldsymbol{\alpha}_k(I - T_k)^{-1}\mathbf{e}$. The mean service time of a type J batch is given by $E(s_J) = \sum_{i=1}^{|J|} E(s_{j_i})$, where $|J|$ is the number of integers in the string J. The service rate of a batch J is defined as $\mu_J = (E(s_J))^{-1}$.

The traffic intensity of the queueing system can be defined in terms of batch arrival rates and batch service rates:

$$\begin{aligned} \rho &= \sum_{J \in \aleph} \lambda_J/\mu_J = \sum_{J \in \aleph} \lambda_J \sum_{i=1}^{|J|} E\left(s_{j_i}\right) = \sum_{J \in \aleph} \lambda_J \sum_{k=1}^{K} N(J,k)E(s_k) \\ &= \sum_{k=1}^{K} \sum_{J \in \aleph} \lambda_J N(J,k)E(s_k) = \sum_{k=1}^{K} \lambda_{(k)}/\mu_k. \end{aligned} \tag{2.7}$$

By Loynes [24], the queueing system is stable if and only if $\rho < 1$. Therefore, throughout this paper, we assume $\rho < 1$ to ensure system stability. We note that many classical discrete time queueing systems, such as the *GI*[*K*]/*PH*[*K*]/1 queue, the *MMAP* [*K*]/*PH*[*K*]/1 queue and the *GI*/*PH*/1 queue, are special cases of the *SM*[*K*]/*PH*[*K*]/1 queue.

3. Analysis of the generalized age process

3.1. The generalized age process

The basic idea to analyze the sojourn times is originated from the following fundamental relationship for waiting times in queues. Let w_n be the (actual) waiting time of the n-th batch. Then we have

$$w_{n+1} = \max\left\{0, w_n + s_{J_n} - \tau_{n+1}\right\}, \quad n \geq 0, \tag{3.1}$$

where τ_{n+1} is the length of the time between the n-th batch and the $(n+1)$-st batch, J_n is the type of the n-th batch, and $s_{J_n}$is the service time of the n-th batch. The process $\{(w_n, \xi_n), n \geq 0\}$ is a Markov chain. In fact, $\{(w_n, \xi_n), n \geq 0\}$ is a Markov chain with a *block repeating* structure. Unfortunately, previous study of such Markov chains (see Zhao et al. [38]) shows that it is difficult to analyze its steady state distribution or to develop efficient algorithms for computing relevant performance measures. On the other hand, with certain conditions on the arrival or the service process, some processes can be constructed, which are closely related to $\{w_n, n \geq 0\}$ and are analytically and numerically tractable. Some examples of such processes are the age process and the total workload (i.e., the virtual waiting time) process. In this paper, we use the age process and the total workload process to study sojourn times and waiting times. In HE [18], the total workload process is used in the study of waiting times.

The system status is observed at integer epochs $t = 0, 1, 2, \ldots$. We define the *age* of a batch in time period $[t, t+1)$, which will be called period t, as the total time the batch has been in the queueing system, given that the batch is in the system in period t. The *generalized age* process $\{a_g(t), t \geq 0\}$ of the batch in service or to be served next (if the system is empty) is defined as

$$a_g(t) = w_{n(t)} + s_{J_{n(t)}} - \tau_{n(t)+1} + t - \eta_{n(t)}, \tag{3.2}$$

where $n(t)$ is the ordinal number of the last batch served in or before period t and $\eta_{n(t)}$ is the departure period of the $n(t)$-th batch. The values of $n(t)$, $w_{n(t)}$, and $\eta_{n(t)}$ are updated when a departure occurs in period t, where $w_{n(t)}$ can be computed by using equation (3.1). Figure 1 shows the relationship between these variables.

The process $\{a_g(t), t \geq 0\}$ evolves as follows. During the service time of a batch, $a_g(t)$ increases its value by one per period of time. At each service completion epoch, the waiting time of the batch entering service is computed, which is $w_n + s_{J_n} - \tau_{n+1}$ after the n-th departure. If $w_n + s_{J_n} - \tau_{n+1}$ is nonnegative, it is the waiting time of

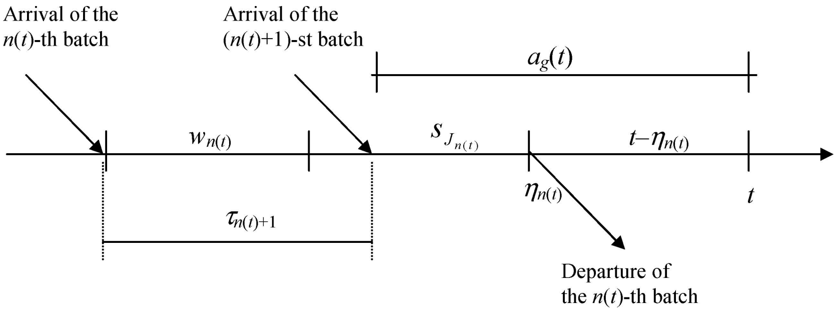
Arrival of the $n(t)$-th batch
Arrival of the $(n(t)+1)$-st batch
$a_g(t)$
$w_{n(t)}$
$s_{J_{n(t)}}$
$t-\eta_{n(t)}$
$\eta_{n(t)}$
t
$\tau_{n(t)+1}$
Departure of the $n(t)$-th batch

Figure 1. Variables $a_g(t)$, $n(t)$, $\eta_{n(t)}$, $w_{n(t)}$, $s_{J_{n(t)}}$ and $\tau_{n(t)+1}$ when $a_g(t) > 0$.

the entering batch or the age of the entering batch in that period, which is the value of $a_g(t)$ in that period. During the service of the $(n+1)$-st batch, $a_g(t)$ increases its value by one after each period of service, which will last for $s_{J_{n+1}}$ periods of time until the service is completed. For this case, $a_g(t)$ is the age of the batch in service. If $w_n + s_{J_n} - \tau_{n+1}$ is negative, then the next batch to be served has not arrived yet. It will take $-a_g(t) = -(w_n + s_{J_n} - \tau_{n+1})$ periods until the next service is initialized. During a period in which $a_g(t)$ is negative, the queueing system is empty and $-a_g(t)$ is the remaining time of the idle period. In summary, the variable $a_g(t)$ records the age of the batch currently in service if $a_g(t) \geq 0$ at time t. If $a_g(t) < 0$, $-a_g(t)$ records the remaining time of the idle period. We call $a_g(t)$ the *generalized age* of the batch in service. Equations (3.1) and (3.2) show that the process $\{w_n, n \geq 0\}$ is an embedded process of $\{a_g(t), t \geq 0\}$ at the departure epochs of batches. A typical sample path of $a_g(t)$ is shown in figure 2.

At a departure epoch, the sojourn time of the next batch to be served is computed as, by equation (3.1), $w_n + s_{J_n}$. That is: the process $\{w_n, n \geq 0\}$ jumps from one departure epoch to the next. For $\{a_g(t), t \geq 0\}$, it takes s_{J_n} periods to complete the journey, if $w_n + s_{J_n} - \tau_{n+1} \geq 0$. In each of these s_{J_n} periods, the value of $a_g(t)$ increases by one.

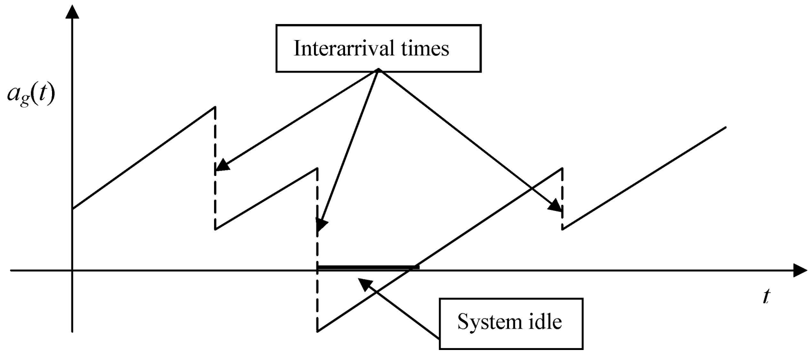

Figure 2. A sample path of $a_g(t)$.

Based on these analyses, it is easy to see that $a_g(t)$ satisfies the following equation:

$$a_g(t+1) = \begin{cases} a_g(t) + 1, & \text{if service continues at time } t+1; \\ a_g(t) + 1 - \tau_{n(t+1)+1}, & \text{if service completes at time } t+1. \end{cases} \tag{3.3}$$

In order to construct a Markov chain associated with $a_g(t)$, we introduce some auxiliary variables related to the phase of the arrival and service processes. We define a process $\{I_a(t),\ t \geq 0\}$ from the Markov chain $\{\xi_n, n \geq 0\}$ defined in Section 2 as: $I_a(t) = \xi_n$ if the n-th batch is the last batch departed in or before period t, i.e., $I_a(t)$ may change its value only at service completion epochs. Let $I_s(t)$ be the phase of the service in period t (if any) and $J(t)$ be the type of batch in service in period t (if any). If there is no service in period t, $J(t)$ is the type of the next batch to be served and $I_s(t)$ the initial service phase of the next batch to be served.

Putting these variables together, we obtain a process $\{(a_g(t), I_a(t), J(t), I_s(t)), t \geq 0\}$. This process has the so called Markovian property during service periods, since the service time of any batch is governed by an underlying Markov chain (PH-distribution), the phase of the arrival process is not changed, and $a_g(t)$ increases its value by one. In a service completion period, the value of $I_a(t)$ is updated according to a Markov chain, the interarrival time is then determined and the value of $a_g(t)$ is updated, the value of $J(t)$ is determined by a Markov chain and the value of $I_s(t)$ is determined by the initial distribution of the service time. Therefore, the process $\{(a_g(t),\ I_a(t),\ J(t),\ I_s(t)),\ t \geq 0\}$ has the Markovian property in service completion periods as well. Hence, $\{(a_g(t), I_a(t),$ $J(t),\ I_s(t)),\ t \geq 0\}$ is a Markov chain. We call $a_g(t)$ the *level variable* that takes integer values and $\{I_a(t),\ J(t),\ I_s(t)\}$ *auxiliary variables* that take a finite number of values in the set

$$\{(i,\ J,\ j)\colon 1 \leq i \leq m_a,\ J \in \aleph,\ 1 \leq j \leq m_J\}, \tag{3.4}$$

in which the states are ordered lexicographically. Equation (3.3) shows that the process $\{a_g(t),\ t \geq 0\}$ is skip-free to the right. Thus, $\{(a_g(t), I_a(t), J(t), I_s(t)),\ t \geq 0\}$ is a Markov chain of *GI*/*M*/1 type with no boundary at the level 0. Denote by

$$\begin{aligned} m_{\text{tot}} &= \sum_{n=1}^{N} m_{J_n}; \\ \boldsymbol{\alpha}(J_n) &= \big(0, \ldots, 0, \boldsymbol{\alpha}_{J_n}, 0, \ldots, 0\big), \quad 1 \leq n \leq N; \\ T_{\text{tot}} &= \begin{pmatrix} T_{J_1} & & \\ & \ddots & \\ & & T_{J_N} \end{pmatrix}, \qquad \mathbf{T}^0_{\text{tot}} = (I - T_{\text{tot}})\mathbf{e} = \begin{pmatrix} \mathbf{T}^0_{J_1} \\ \vdots \\ \mathbf{T}^0_{J_N} \end{pmatrix}, \end{aligned} \tag{3.5}$$

where $\boldsymbol{\alpha}(J_n)$ is a row vector of the size m_{tot}, $1 \leq n \leq N$, T_{tot} is an $m_{\text{tot}} \times m_{\text{tot}}$ matrix. The vector $\boldsymbol{\alpha}(J_n)$ is obtained by putting the vector $\boldsymbol{\alpha}_{J_n}$ in the positions from $\sum_{i=1}^{n-1} m_{J_i} + 1$ to $\sum_{i=1}^{n} m_{J_i}$and zero in all other positions in a vector of the size m_{tot}. (Note: when there is no confusion, we shall frequently write $\alpha(J)$ for the batch J.) The transition probability matrix of the Markov chain $\{(a_g(t), I_a(t), J(t), I_s(t)), t \geq 0\}$ can be given as follows:

$$P_g = \begin{array}{c|ccccccc|l} & \cdots & -2 & -1 & 0 & 1 & 2 & \cdots & \\ \hline \vdots & \ddots & \ddots & & & & & & \\ -2 & \cdots & 0 & I & & & & & \\ -1 & \cdots & 0 & 0 & I & & & & \\ 0 & \cdots & A_3 & A_2 & A_1 & A_0 & & & \leftarrow \text{level zero, i.e.,} \ a_g(t) = 0. \\ 1 & \cdots & A_4 & A_3 & A_2 & A_1 & A_0 & & \\ \vdots & \vdots & \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \end{array} \tag{3.6}$$

where

$$A_0 = I \otimes T_{\text{tot}}; \qquad A_n = \sum_{i=1}^{N} D_{a,J_i}(n) \otimes \left(\mathbf{T}^0_{\text{tot}} \alpha(J_i)\right), \quad n \geq 1, \tag{3.7}$$

where the notation "$\otimes$" is for Kronecker product of matrix. Note that the age of a batch is zero in the arrival period. The transition probabilities in equation (3.6) can be verified as follows. For $a_g(t) = x \geq 0$, using definitions and equation (3.3), we have

$$\begin{aligned} &P\{a_g(t+1) = x+1, I_a(t+1) = j', J(t+1) = J', I_s(t+1) = i' \\ &\qquad |a_g(t) = x, I_a(t) = j, J(t) = J, I_s(t) = i\} \\ &\qquad = \begin{cases} (T_J)_{i,i'}, & \text{if } j = j', J' = J; \\ 0, & \text{otherwise.} \end{cases} \\ &P\{a_g(t+1) = x+1-s, I_a(t+1) = j', J(t+1) = J', I_s(t+1) = i' \\ &\qquad |a_g(t) = x, I_a(t) = j, J(t) = J, I_s(t) = i\} \\ &\qquad = P\left\{\tau_{n(t+1)+1} = s, \xi_{n(t)+1} = j' | \xi_{n(t)} = j\right\} \\ &\qquad \cdot P\{\text{The service of a type } J \text{ batch is completed in phase } i\} \\ &\qquad \cdot P\{\text{Initial service phase of a type } J' \text{ batch is } i'\} \\ &\qquad = (D_{a,J'}(s))_{j,j'} \left(\mathbf{T}^0_J\right)_i (\alpha_{J'})_{i'}, \quad s \geq 1. \end{aligned} \tag{3.8}$$

Note that, for the second case, $\eta_{n(t+1)} = t+1$ since there is a service completion in period $t+1$. For $a_g(t) = x < 0$, there is no service. Then $a_g(t+1) = a_g(t) + 1$ with probability one, and all auxiliary variables remain the same. Thus, we have verified the transition probability matrix given in equation (3.6).

Apparently, the Markov chain $\{(a_g(t), I_a(t), J(t), I_s(t)), t \geq 0\}$ has rich information about the age process and idle periods. Further, it is shown in Section 4 that the Markov chain can provide information about the distributions of sojourn times and waiting times.

To find the steady state distributions of the age process, sojourn times, and waiting times, we first find the steady state distribution of the Markov chain $\{(a_g(t), I_a(t), J(t), I_s(t)), t \geq 0\}$.

Remark 3.1. The stability analysis of $\{(a_g(t), I_a(t), J(t), I_s(t)), t \geq 0\}$ depends largely on its irreducibility and the irreducibility of the matrix A (defined in equation (3.9)). Unfortunately, neither A nor $\{(a_g(t), I_a(t), J(t), I_s(t)), t \geq 0\}$ is guaranteed to be irreducible under the irreducibility assumptions on the matrices D_a and the service time distributions given in Section 2. For instance, let $n_{\max} = \max\{t\colon D_J(t) \neq 0, \text{for some } J \in \aleph\}$. If $n_{\max} < \infty$, then these states with $a_g(t) < -n_{\max} + 1$ is not reachable from states with $a_g(t) \geq -n_{\max} + 1$. Furthermore, it is possible for A and $\{(a_g(t), I_a(t), J(t), I_s(t)), t \geq 0\}$ to have several closed classes of states or transient states. (This was reported in Van Houdt and Blondia [35–37]). To analyze the Markov chain, we need to identify the closed classes of states and remove transient states. Then we concentrate on some irreducible subsets. Therefore, in the rest of the paper, we shall assume that the Markov chain $\{(a_g(t), I_a(t), J(t), I_s(t)), t \geq 0\}$ is irreducible and aperiodic.

Remark 3.2. In Van Houdt and Blondia [35], a slightly different approach was used to introduce the Markov chain $\{(a_g(t), I_a(t), J(t), I_s(t)), t \geq 0\}$. That approach may lead to a Markov chain with a fewer number of states in each level. Nonetheless, the improvement is limited.

3.2. Ergodicity of the generalized age process

As the first step to analyze the Markov chain $\{(a_g(t), I_a(t), J(t), I_s(t)), t \geq 0\}$, we show that if the queueing system is stable (i.e., $\rho < 1$), the Markov chain is ergodic. Define

$$A^*(z) = \sum_{n=0}^{\infty} z^n A_n, \quad z \geq 0; \qquad A = A^*(1) = \sum_{n=0}^{\infty} A_n, \tag{3.9}$$

if the summations are well-defined. Denote by $\chi(z)$ the Perron-Frobenius eigenvalue of the nonnegative matrix $A^*(z)$ (i.e., the eigenvalue with the largest modulus). Let $\mathbf{u}(z)$ and $\mathbf{v}(z)$ be the left and right eigenvectors corresponding to $\chi(z)$, respectively, i.e., $\mathbf{u}(z)A^*(z) = \chi(z)\mathbf{u}(z)$ and $A^*(z)\mathbf{v}(z) = \chi(z)\mathbf{v}(z)$. The two vectors $\mathbf{u}(z)$ and $\mathbf{v}(z)$ are normalized by $\mathbf{u}(z)\mathbf{v}(z) = 1$ and $\mathbf{u}(z)\mathbf{e} = 1$. If A is irreducible, it is easy to see that $A^*(z)$ is irreducible and all the elements of the vectors $\mathbf{u}(z)$ and $\mathbf{v}(z)$ are positive for $z > 0$. According to Neuts [26], the function $\chi(z)$ and the vectors $\mathbf{u}(z)$ and $\mathbf{v}(z)$ can be chosen as differentiable functions.

Denote by $\boldsymbol{\beta}_k$ the invariant probability vector of $T_k + \mathbf{T}_k^0\boldsymbol{\alpha}_k$, i.e., $\boldsymbol{\beta}_k(T_k + \mathbf{T}_k^0\boldsymbol{\alpha}_k) = \boldsymbol{\beta}_k$, $\boldsymbol{\beta}_k\mathbf{e} = 1$, $1 \leq k \leq K$. Since $T_k + \mathbf{T}_k^0\boldsymbol{\alpha}_k$ is irreducible, every element of the vector $\boldsymbol{\beta}_k$ is positive (Gantmacher [13]). By Neuts [26], $\boldsymbol{\beta}_k\mathbf{T}_k^0 = \mu_k$. Denote by $\boldsymbol{\beta}_J$ the invariant probability vector of $T_J + \mathbf{T}_J^0\boldsymbol{\alpha}_J$, for $J \in \aleph$. By $\boldsymbol{\beta}_k T_k + \mu_k\boldsymbol{\alpha}_k = \boldsymbol{\beta}_k$, it can be verified, for $J \in \aleph$,

$$\boldsymbol{\beta}_J = \mu_J \left(\frac{\boldsymbol{\beta}_{j_1}}{\mu_{j_1}}, \frac{\boldsymbol{\beta}_{j_2}}{\mu_{j_2}}, \ldots, \frac{\boldsymbol{\beta}_{j_{|J|}}}{\mu_{j_{|J|}}} \right),$$
$$\text{where } \mu_J = \left(\frac{1}{\mu_{j_1}} + \frac{1}{\mu_{j_2}} + \cdots + \frac{1}{\mu_{j_{|J|}}} \right)^{-1}, \tag{3.10}$$

and $|J|$ is the number of integers in the string J. Similar to the definition of $\boldsymbol{\alpha}(J)$ in equation (3.5), we define $\beta(J)$ as $\beta(J) = (0, \ldots, 0, \boldsymbol{\beta}_J, 0, \ldots, 0)$, which is a row vector of the size m_{tot}. Denote by

$$\boldsymbol{\theta}_{\text{tot}} = \frac{\lambda}{\rho} \sum_{i=1}^{N} \left(\boldsymbol{\theta}_a D_{a,J_i} \right) \otimes \left(\frac{\beta(J_i)}{\mu_{J_i}} \right). \tag{3.11}$$

Lemma 3.1. Assume that A is irreducible. The vector $\boldsymbol{\theta}_{\text{tot}}$ is the invariant probability vector of A.

Proof. By the definition of the traffic intensity ρ given in equation (2.7), it is easy to verify that $\boldsymbol{\theta}_{\text{tot}}\mathbf{e} = 1$. Note that

$$A = I \otimes T_{\text{tot}} + \sum_{n=1}^{\infty} \sum_{J \in \aleph} D_{a,J}(n) \otimes \left(\mathbf{T}_{\text{tot}}^0 \boldsymbol{\alpha}(J) \right) = I \otimes T_{\text{tot}} + \sum_{J \in \aleph} D_{a,J} \otimes \left(\mathbf{T}_{\text{tot}}^0 \boldsymbol{\alpha}(J) \right). \tag{3.12}$$

Then we have the following calculations:

$$\begin{aligned}
\boldsymbol{\theta}_{\text{tot}} A &= \frac{\lambda}{\rho} \left(\sum_{i=1}^{N} \left(\boldsymbol{\theta}_a D_{a,J_i} \right) \otimes \left(\frac{\beta(J_i)}{\mu_{J_i}} \right) \right) \left(I \otimes T_{\text{tot}} + \sum_{j=1}^{N} D_{a,J_j} \otimes \left(\mathbf{T}_{\text{tot}}^0 \boldsymbol{\alpha}(J_j) \right) \right) \\
&= \frac{\lambda}{\rho} \sum_{i=1}^{N} \left(\boldsymbol{\theta}_a D_{a,J_i} \right) \otimes \left(\frac{\beta(J_i) T_{\text{tot}}}{\mu_{J_i}} \right) + \frac{\lambda}{\rho} \sum_{i=1}^{N} \sum_{j=1}^{N} \left(\boldsymbol{\theta}_a D_{a,J_i} D_{a,J_j} \right) \\
&\quad \otimes \left(\frac{\beta(J_i)}{\mu_{J_i}} \mathbf{T}_{\text{tot}}^0 \boldsymbol{\alpha}(J_j) \right) \\
&= \frac{\lambda}{\rho} \sum_{i=1}^{N} \left(\boldsymbol{\theta}_a D_{a,J_i} \right) \otimes \left(0, \ldots, 0, \frac{\boldsymbol{\beta}_{J_i} T_{J_i}}{\mu_{J_i}}, 0, \ldots, 0 \right) \\
&\quad + \frac{\lambda}{\rho} \sum_{j=1}^{N} \left(\left(\sum_{i=1}^{N} \boldsymbol{\theta}_a D_{a,J_i} \right) D_{a,J_j} \right) \otimes \boldsymbol{\alpha}(J_j) \\
&= \frac{\lambda}{\rho} \sum_{i=1}^{N} \left(\boldsymbol{\theta}_a D_{a,J_i} \right) \otimes \left(0, \ldots, 0, \frac{\boldsymbol{\beta}_{J_i} T_{J_i}}{\mu_{J_i}} + \boldsymbol{\alpha}_{J_i}, 0, \ldots, 0 \right)
\end{aligned}$$

$$\begin{aligned}
&= \frac{\lambda}{\rho}\sum_{i=1}^{N}\left(\boldsymbol{\theta}_a D_{a,J_i}\right)\otimes\left(0,\ldots,0,\frac{\boldsymbol{\beta}_{J_i}T_{J_i}+\boldsymbol{\beta}_{J_i}\mathbf{T}^0_{J_i}\boldsymbol{\alpha}_{J_i}}{\mu_{J_i}},0,\ldots,0\right)\\
&= \frac{\lambda}{\rho}\sum_{i=1}^{N}\left(\boldsymbol{\theta}_a D_{a,J_i}\right)\otimes\left(0,\ldots,0,\frac{\boldsymbol{\beta}_{J_i}}{\mu_{J_i}},0,\ldots,0\right)\\
&= \frac{\lambda}{\rho}\sum_{i=1}^{N}\left(\boldsymbol{\theta}_a D_{a,J_i}\right)\otimes\left(\frac{\boldsymbol{\beta}(J_i)}{\mu_{J_i}}\right)=\boldsymbol{\theta}_{\text{tot}}. \qquad (3.13)
\end{aligned}$$

Note that $\boldsymbol{\beta}_J\mathbf{T}^0_J = 1/E(s_J) = \mu_J$ (Neuts [26]) and $\boldsymbol{\beta}_J(T_J+\mathbf{T}^0_J\boldsymbol{\alpha}_J) = \boldsymbol{\beta}_J$. Therefore, $\boldsymbol{\theta}_{\text{tot}}$ is the invariant probability vector of A. This completes the proof of Lemma 3.1. □

Lemma 3.2. Assume that A is irreducible. At $z=1$, we have $\chi(1)=1$ and $\boldsymbol{\theta}_{\text{tot}}\Sigma_n nA_n\mathbf{e} = \chi^{(1)}(1) = 1/\rho$. Consequently, $\chi^{(1)}(1) > 1$ if and only if $\rho < 1$. (Note that $\chi^{(1)}(1)$ is the first derivative of the function $\chi(z)$ at $z=1$.)

Proof. It is easy to see $\chi(1)=1$. Furthermore, we have $\mathbf{u}(1)=\boldsymbol{\theta}_{\text{tot}}$ (by definition and Lemma 3.1) and $\mathbf{v}(1)=\mathbf{e}$. By $\mathbf{u}(z)\mathbf{e}=1$, we obtain $\mathbf{u}^{(1)}(z)\mathbf{e} = 0$. By taking derivatives on both sides of $\mathbf{u}(z)A^*(z)=\chi(z)\mathbf{u}(z)$, we obtain $\mathbf{u}^{(1)}(z)A^*(z)+\mathbf{u}(z)A^{*(1)}(z)=\chi^{(1)}(z)\mathbf{u}(z)+\chi(z)\mathbf{u}^{(1)}(z)$. Letting $z=1$ and multiplying $\mathbf{e}$ on both sides of the equation, we obtain $\mathbf{u}(1)A^{*(1)}(1)\mathbf{e}=\chi^{(1)}(1)$, i.e.,

$$\boldsymbol{\theta}_{\text{tot}}\left(\sum_{n=0}^{\infty}nA_n\right)\mathbf{e} = \chi^{(1)}(1).$$

Note that $\boldsymbol{\theta}_{\text{tot}}(\sum_{n=0}^{\infty}nA_n)\mathbf{e}$ is finite since $\lambda > 0$. Using equations (3.7) and (3.11), we have

$$\begin{aligned}
\boldsymbol{\theta}_{\text{tot}}\sum_{n=0}^{\infty}nA_n\mathbf{e} &= \frac{\lambda}{\rho}\left(\sum_{i=1}^{N}\left(\boldsymbol{\theta}_a D_{a,J_i}\right)\otimes\left(\frac{\boldsymbol{\beta}(J_i)}{\mu_{J_i}}\right)\right)\left(\sum_{j=1}^{N}\sum_{n=1}^{\infty}\left(nD_{a,J_j}(n)\mathbf{e}\right)\otimes\mathbf{T}^0_{\text{tot}}\right)\\
&= \frac{\lambda}{\rho}\sum_{j=1}^{N}\left(\sum_{i=1}^{N}\left(\boldsymbol{\theta}_a D_{a,J_i}\sum_{n=1}^{\infty}\left(nD_{a,J_j}(n)\mathbf{e}\right)\right)\cdot\left(\frac{\boldsymbol{\beta}_{J_i}\mathbf{T}^0_{J_i}}{\mu_{J_i}}\right)\right)\\
&= \frac{\lambda}{\rho}\sum_{j=1}^{N}\left(\sum_{n=1}^{\infty}\boldsymbol{\theta}_a\left(nD_{a,J_j}(n)\mathbf{e}\right)\right)\\
&= \frac{\lambda}{\rho}\sum_{n=1}^{\infty}\boldsymbol{\theta}_a\left(nD_a(n)\mathbf{e}\right) = \frac{\lambda}{\rho}E_{\boldsymbol{\theta}_a}(\tau) = \frac{1}{\rho}.
\end{aligned} \qquad (3.14)$$

Note that the definition of λ (equation (2.4)) is used to obtain the last equality. Therefore, $\chi^{(1)}(1)=1/\rho$. Then $\chi^{(1)}(1) > 1$ if and only if $\rho < 1$. This completes the proof of Lemma 3.2. □

Theorem 3.3. If the Markov chain $\{(a_g(t),\, I_a(t),\, J(t),\, I_s(t)),\ t \geq 0\}$ is irreducible, it is positive recurrent if and only if $\rho < 1$.

Proof. If the Markov chain $\{(a_g(t), I_a(t), J(t), I_s(t)), t \geq 0\}$ is irreducible, by the structure of the transition probability matrix P_g, the matrix A must be irreducible. Thus, by Lemma 3.2, the vector $\boldsymbol{\theta}_{\text{tot}}$ is the invariant probability vector of A. We use the mean-drift method (Fayolle et al. [10]) to prove the ergodicity result.

First, if the Markov chain is positive recurrent, we consider a Markov chain with the following transition probability matrix:

$$P_{GI/M/1} = \begin{pmatrix} \sum_{n=1}^{\infty} A_n & A_0 & & & \\ \sum_{n=2}^{\infty} A_n & A_1 & A_0 & & \\ \sum_{n=3}^{\infty} A_n & A_2 & A_1 & A_0 & \\ \vdots & \ddots & \ddots & \ddots & \ddots \end{pmatrix}. \tag{3.15}$$

Note that $P_{GI/M/1}$ is the transition probability matrix of the age process to be defined in Section 4. The Markov chain is obtained by removing the levels $\{-1, -2, \ldots\}$ of the generalized age process. It is true that, from any level n (>0), the Markov chain can reach the level zero after, on average, a finite number of transitions. Since $P_{GI/M/1}$ is an irreducible *GI/M/*1 type Markov chain, according to Neuts [26], we must have $\boldsymbol{\theta}_{\text{tot}} \Sigma_n n A_n \mathbf{e} > 1$, which is equivalent to $\rho < 1$ by Lemma 3.2.

Next, we prove the sufficiency of $\rho < 1$. By Lemma 3.2, $\rho < 1$ implies $\chi^{(1)}(1) > 1$. If $\chi^{(1)}(1) > 1$, there exists z such that $0 < z < 1$, z is close to 1, and $\chi(z) < z$. Then $A^*(z)\mathbf{v}(z) = \chi(z)\mathbf{v}(z) < z\mathbf{v}(z)$. We define the following (vector form) Lyapunov function for all the states in the level n:

$$\mathbf{f}(n) = \begin{cases} z^{-n}\mathbf{v}(z), & \text{for } n \geq 0; \\ -n\mathbf{v}(z), & \text{for } n < 0. \end{cases} \tag{3.16}$$

(With an abuse of notation in the definition of $\mathbf{f}(n)$, we use "n" to represent all states of the level n.) It is true that $\mathbf{f}(n) \to \infty$ when $|n| \to \infty$, since $0 < z < 1$ and every element of $\mathbf{v}(z)$ is positive. Denote by $\varepsilon_1 = \min_i\{(\mathbf{v}(z))_i\}$ (>0) and $\varepsilon_2 = (z - \chi(z))/2(> 0)$. We choose n_0 such that, for any $n \geq n_0$,

$$\left(\sum_{i=1}^{\infty} i A_{n+1+i}\right) \mathbf{v}(z) \leq \varepsilon_2 \mathbf{v}(z). \tag{3.17}$$

We choose $\varepsilon_3 = \min\{1, \varepsilon_1\varepsilon_2\}/2$ (>0). Next, we calculate the mean-drift to level zero of the Markov chain with respect to the Lyapunov function $\mathbf{f}(n)$. If $n \geq n_0$,

$$\begin{aligned} &E[\mathbf{f}(a_g(t+1)) - \mathbf{f}(a_g(t)) \,|\, a_g(t) = n, I_a(t), J(t), I_s(t)] \\ &\quad = \left(\sum_{i=0}^{n+1} z^{i-n-1} A_i\right)\mathbf{v}(z) + \left(\sum_{i=n+2}^{\infty} (i-n-1)A_i\right)\mathbf{v}(z) - z^{-n}\mathbf{v}(z) \\ &\quad \leq \left(\sum_{i=0}^{\infty} z^i A_i\right) z^{-n-1}\mathbf{v}(z) + \varepsilon_2\mathbf{v}(z) - z^{-n}\mathbf{v}(z) \\ &\quad = z^{-n-1}\chi(z)\mathbf{v}(z) + \varepsilon_2\mathbf{v}(z) - z^{-n}\mathbf{v}(z) = -z^{-n-1}\left(z - \chi(z) - \varepsilon_2 z^{n+1}\right)\mathbf{v}(z) \\ &\quad \leq -z^{-n-1}\left(z - \chi(z) - \varepsilon_2\right)\mathbf{v}(z) \leq -z^{-n-1}\varepsilon_2\varepsilon_1\mathbf{e}/2 \leq -\varepsilon_2\varepsilon_1\mathbf{e}/2 \leq -\varepsilon_3\mathbf{e}. \end{aligned} \tag{3.18}$$

Note that $0 < z < 1$. If $n < 0$,

$$\begin{aligned} &E[\mathbf{f}(a_g(t+1)) - \mathbf{f}(a_g(t)) \,|\, a_g(t) = n, I_a(t), J(t), I_s(t)] \\ &\quad = (-n-1)\mathbf{v}(z) - (-n)\mathbf{v}(z) = -\mathbf{v}(z) \leq -\varepsilon_3\mathbf{e}. \end{aligned} \tag{3.19}$$

For $0 \leq n \leq n_0$, we have

$$\begin{aligned} &E[\mathbf{f}(a_g(t+1)) | a_g(t) = n, I_a(t), J(t), I_s(t)] \\ &\quad = \left(\sum_{i=0}^{n+1} z^{i-n-1} A_i\right)\mathbf{v}(z) + \left(\sum_{i=n+2}^{\infty} (i-n-1)A_i\right)\mathbf{v}(z) \\ &\quad \leq \left(\sum_{i=0}^{\infty} z^i A_i\right) z^{-n-1}\mathbf{v}(z) + \left(\sum_{i=0}^{\infty} iA_i\right)\mathbf{v}(z) \\ &\quad = z^{-n-1}\chi(z)\mathbf{v}(z) + \left(\sum_{i=0}^{\infty} iA_i\right)\mathbf{v}(z) \\ &\quad \leq z^{-n-1}\mathbf{v}(z) + \left(\sum_{i=0}^{\infty} iA_i\right)\mathbf{v}(z) \leq z^{-n_0-1}\mathbf{v}(z) + \left(\sum_{i=0}^{\infty} iA_i\right)\mathbf{v}(z) < \infty. \end{aligned} \tag{3.20}$$

Note that $\sum_{i=0}^{\infty} iA_i$ is finite since $\lambda > 0$ and all elements of $\boldsymbol{\theta}_{\text{tot}}$ and $\mathbf{e}$ are positive. Therefore, we have shown that, with respect to the Lyapunov function $\mathbf{f}(n)$, the mean-drift away from the level zero is less than $-\varepsilon_3$ for all but a finite number of states. Therefore, the Markov chain is positive recurrent by Foster's criterion (see Theorem 2.2.3 in Fayolle, et al. [10]). This completes the proof of Theorem 3.3. □

3.3. The steady state distribution of the generalized age process

We assume that the Markov chain $\{(a_g(t), I_a(t), J(t), I_s(t)), t \geq 0\}$ is irreducible and aperiodic and $\rho < 1$ so that the Markov chain is ergodic. Denote by $\boldsymbol{\pi} = (\ldots, \boldsymbol{\pi}(-1), \boldsymbol{\pi}(0), \boldsymbol{\pi}(1), \ldots)$ the steady state distribution of $\{(a_g(t), I_a(t), J(t), I_s(t)), t \geq 0\}$, where $\boldsymbol{\pi}(n) = (\ldots, \pi_{i,J,j}(n), \ldots)$ is a row vector of the size $m_a m_{\text{tot}}$ and

$$\pi_{i,J,j}(n) = \lim_{t\to\infty} P\{a_g(t) = n, I_a(t) = i, J(t) = J, I_s(t) = j \mid a_g(0), I_a(0), J(0), I_s(0)\}. \tag{3.21}$$

Then we must have $\boldsymbol{\pi} = \boldsymbol{\pi} P_g$ and $\boldsymbol{\pi}\mathbf{e} = 1$. Since $\{(a_g(t), I_a(t), J(t), I_s(t)), t \geq 0\}$ has the so called *GI/M/*1 structure for nonnegative levels (i.e., $a_g(t) \geq 0$), it is well known that its steady state distribution has a matrix geometric solution (Neuts [26]):

$$\boldsymbol{\pi}(n) = \boldsymbol{\pi}(0)R^n, \quad n \geq 0, \tag{3.22}$$

where R is an $(m_a m_{\text{tot}}) \times (m_a m_{\text{tot}})$ matrix and is the minimal nonnegative solution to equation

$$R = \sum_{n=0}^{\infty} R^n A_n. \tag{3.23}$$

We refer to Gail et al. [11,12] and Neuts [26] for more about the matrix R. Since the traffic intensity $\rho < 1$, $\boldsymbol{\theta}_{\text{tot}} \Sigma_n n A_n \mathbf{e} = 1/\rho > 1$. According to Neuts [26], the spectral radius of the matrix R is less than one, i.e., all eigenvalues of R are within the unit circle. For $\{\boldsymbol{\pi}(0), \boldsymbol{\pi}(-1), \ldots\}$, we have, for $n \geq 0$,

$$\boldsymbol{\pi}(-n) = \boldsymbol{\pi}(-n-1) + \sum_{i=0}^{\infty} \boldsymbol{\pi}(i) A_{n+1+i} = \boldsymbol{\pi}(-n-1) + \boldsymbol{\pi}(0) \sum_{i=0}^{\infty} R^i A_{n+1+i}. \tag{3.24}$$

By induction, we obtain, for $n \geq 1$,

$$\boldsymbol{\pi}(-n) = \boldsymbol{\pi}(0) - \boldsymbol{\pi}(0) \sum_{s=1}^{\infty} \left(\sum_{t=\max\{0,s-n\}}^{s-1} R^t \right) A_s. \tag{3.25}$$

Using the above equations, an explicit expression can be found for $\boldsymbol{\pi}(0)$.

Theorem 3.4. If the Markov chain $\{(a_g(t), I_a(t), J(t), I_s(t)), t \geq 0\}$ is irreducible and aperiodic and $\rho < 1$, then

$$\boldsymbol{\pi}(0) = \rho \boldsymbol{\theta}_{\text{tot}}(I - R), \quad \sum_{n=1}^{\infty} \boldsymbol{\pi}(-n) = \rho \boldsymbol{\theta}_{\text{tot}} \left(\sum_{n=1}^{\infty} n A_n - (I - R)^{-1}(A - R) \right), \quad \text{and}$$

$$\sum_{n=1}^{\infty} \boldsymbol{\pi}(-n)\mathbf{e} = 1 - \rho.$$

Proof. Since the Markov chain is ergodic, we must have $\boldsymbol{\pi}(-n) \to 0$ as $n \to 0$. Equation (3.25) leads to

$$\begin{aligned}\boldsymbol{\pi}(0) &= \boldsymbol{\pi}(0)\sum_{s=1}^{\infty}\left(\sum_{t=0}^{s-1} R^t\right)A_s = \boldsymbol{\pi}(0)(I-R)^{-1}\sum_{s=1}^{\infty}(I-R^s)A_s \\ &= \boldsymbol{\pi}(0)(I-R)^{-1}(A-R), \end{aligned} \tag{3.26}$$

which yields $\boldsymbol{\pi}(0)(I-R)^{-1} = \boldsymbol{\pi}(0)(I-R)^{-1}A$. By Lemma 3.1, $\boldsymbol{\pi}(0) = c_1\boldsymbol{\theta}_{\text{tot}}(I-R)$, where c_1 is a constant to be determined. Note that $\boldsymbol{\theta}_{\text{tot}}(I-R) \geq 0$ since it can be proved that $\boldsymbol{\theta}_{\text{tot}} \geq \boldsymbol{\theta}_{\text{tot}}R$, since $\boldsymbol{\theta}_{\text{tot}} = \boldsymbol{\theta}_{\text{tot}}A$ and $\boldsymbol{\theta}_{\text{tot}} \geq \boldsymbol{\theta}_{\text{tot}}A_0$. Define $\boldsymbol{\pi}^*_{-1}(z) = \sum_{n=1}^{\infty} z^n\boldsymbol{\pi}(-n)$. By equation (3.24) and routine calculations, we obtain

$$\boldsymbol{\pi}^*_{-1}(z) = \frac{\boldsymbol{\pi}(0)}{(1-z)}\left(I - \sum_{n=1}^{\infty}\left(\sum_{i=0}^{n-1} z^{n-1-i}R^i\right)A_n\right). \tag{3.27}$$

By l'Hôpital's Rule and Lemma 3.2, we have

$$\begin{aligned}\sum_{n=1}^{\infty}\boldsymbol{\pi}(-n) = \lim_{z\to 1}\boldsymbol{\pi}^*_{-1}(z) &= \boldsymbol{\pi}(0)\left(\sum_{n=1}^{\infty}\left(\sum_{i=0}^{n-1}(n-1-i)R^i\right)A_n\right) \\ &= c_1\boldsymbol{\theta}_{\text{tot}}\left(\sum_{n=1}^{\infty}(I-R)\left(\sum_{i=0}^{n-1}(n-1-i)R^i\right)A_n\right) \\ &= c_1\boldsymbol{\theta}_{\text{tot}}\left(\sum_{n=1}^{\infty}\left(nI - \sum_{i=0}^{n-1}R^i\right)A_n\right) \\ &= c_1\boldsymbol{\theta}_{\text{tot}}\left(\sum_{n=1}^{\infty}(nI-(I-R)^{-1}(I-R^n))A_n\right) \\ &= c_1\boldsymbol{\theta}_{\text{tot}}\left(\sum_{n=1}^{\infty}nA_n - (I-R)^{-1}(A-R)\right). \end{aligned} \tag{3.28}$$

Postmultiplying $\mathbf{e}$ on both sides of equation (3.28), yields

$$\begin{aligned}\boldsymbol{\pi}^*_{-1}(1)\mathbf{e} &= c_1\boldsymbol{\theta}_{\text{tot}}\left(\sum_{n=1}^{\infty}nA_n - (I-R)^{-1}(A-R)\right)\mathbf{e} \\ &= c_1\left(\boldsymbol{\theta}_{\text{tot}}\left(\sum_{n=1}^{\infty}nA_n\right)\mathbf{e} - 1\right) = c_1\left(\frac{1}{\rho} - 1\right). \end{aligned} \tag{3.29}$$

By equation (3.29), $1 = \boldsymbol{\pi}\mathbf{e} = \boldsymbol{\pi}^*_{-1}(1)\mathbf{e} + \boldsymbol{\pi}(0)(I-R)^{-1}\mathbf{e} = c_1(1/\rho - 1) + c_1 = c_1/\rho$. Thus, $c_1 = \rho$. All results follow directly. This completes the proof of Theorem 3.4. □

The first result in Theorem 3.4 provides explicit solution for the steady state distribution. The last result in Theorem 3.4 says that the probability that the system is empty is $1 - \rho$, which is consistent with intuition. Note that if $a_g(t) = 0$, the batch in service has just arrived so that the system is not empty.

Remark 3.3. The results obtained in this section may be valid for more general cases. Recall that $n_{\max} = \max\{t : C_J(t) \neq 0, \text{ for some } J \in \aleph\}$ introduced in Remark 3.1. For instance, if $n_{\max} < \infty$ and $\boldsymbol{\pi}(0)$ can be computed, then equations (3.22) and (3.25) can be used for computing $\boldsymbol{\pi} = (\ldots, \boldsymbol{\pi}(-1), \boldsymbol{\pi}(0), \boldsymbol{\pi}(1), \ldots)$. For this case, the states with $a_g(t) < -n_{\max} + 1$ are transient states. By equation (3.24) or equation (3.25), if $n \geq n_{\max}$, $\boldsymbol{\pi}(-n) = 0$.

4. Age, total workload, sojourn times, and waiting times

In this section, we show how to use the steady state distribution $\boldsymbol{\pi}$ of the generalized age process to derive the distributions of age, the total workload, waiting times, and sojourn times. In this section, we assume that all conditions given in Theorem 3.4 are satisfied.

4.1. Distributions of age and total workload

Let a_g be the generic random variable of the generalized age of the batch in service in an arbitrary period. By equations (3.22) and (3.25), the distribution of a_g is obtained easily as

$$P\{a_g = n\} = \begin{cases} \rho\boldsymbol{\theta}_{\text{tot}}(I - R)\left(I - \sum_{s=1}^{\infty}\left(\sum_{t=\max\{0,s-n\}}^{s-1} R^t\right)A_s\right)\mathbf{e}, & n \leq -1; \\ \rho\boldsymbol{\theta}_{\text{tot}}(I - R)R^n\mathbf{e}, & n \geq 0. \end{cases} \tag{4.1}$$

Note that the nonnegative part of the distribution of a_g is a PH-distribution whose matrix representation can be constructed in a way similar to that of the sojourn times (see Section 4.2).

The *total workload* (virtual waiting time) is defined as the total service time of all batches waiting plus the remaining service time of the batch in service (if any). Based on equation (3.1), we introduce the *generalized total workload process*:

$$v_g(t) = w_{l(t)} + s_{J_{l(t)}} - (t - \zeta_{l(t)}), \tag{4.2}$$

where $l(t)$ is the ordinal number of the last batch arrived in or before period t, and $\zeta_{l(t)}$ is the arrival time of the $l(t)$-th batch. It is readily seen that $l(t)$, $w_{l(t)}$, $J_{l(t)}$, $s_{J_{l(t)}}$, and $\zeta_{l(t)}$ update their values if a batch arrives in period t. The relationship between $l(t)$, $w_{l(t)}$, $\zeta_{l(t)}$, $s_{J_{l(t)}}$, and $v_g(t)$ is shown in figure 3. For more details about the process $v_g(t)$, see HE [18].

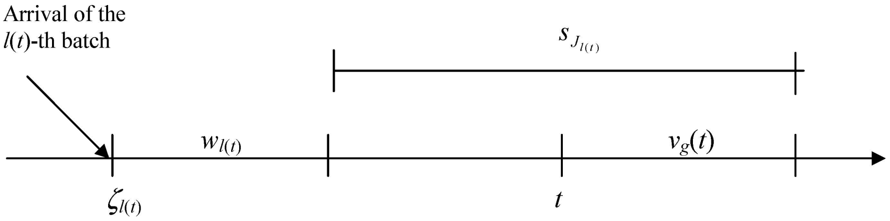

Figure 3. Variables $v_g(t)$, $l(t)$, $\zeta_{l(t)}$, $w_{l(t)}$, and $s_{J_{l(t)}}$, if $v_g(t) > 0$.

A typical sample path of $v_g(t)$ is shown in figure 4.

The relationship between the generalized age process and the generalized total workload process is shown in the following lemma. Note that the following lemma holds for continuous time queueing models as well (see HE [19]).

Lemma 4.1. In a *busy cycle* (starting from the first service in a busy period and ending right before the beginning of the next busy period), the number of times $a_g(t)$ up-crossing x equals the number of times that $v_g(t)$ down-crossing x, for any real number x.

Proof. First note that $v_g(t)$ is always non-increasing, except at batch arrival epochs; $a_g(t)$ is always non-decreasing, except at batch service completion epochs. Next, we compare the two processes. For each batch, the total workload right after its arrival epoch (a jump up epoch of $v_g(t)$) equals the sojourn time of that batch ($a_g(t)$ just before its departure). For each batch, its waiting time ($v_g(t)$ just before its arrival) equals the age just before it begins its service (a jump down epoch of $a_g(t)$). Also note (see figure 5), $v_g(t)$ at the end of a busy cycle equals $a_g(t)$ at the end of the corresponding busy period. Thus, if we draw a horizontal line at x, then this line crosses $v_g(t)$ and $a_g(t)$ for the same number of times. This completes the proof of Lemma 4.1. □

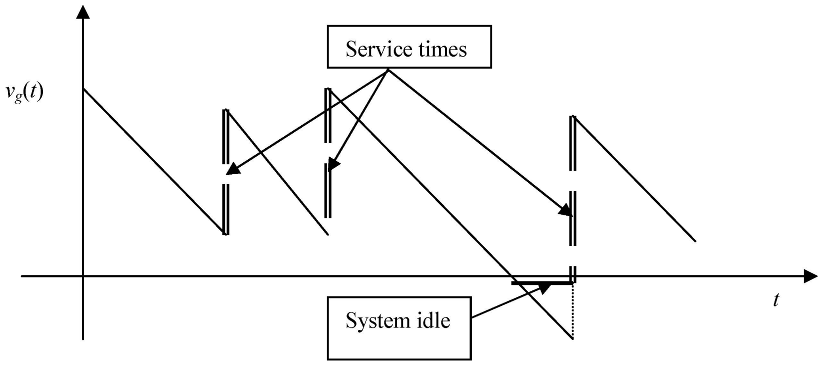

Figure 4. A sample path of $v_g(t)$.

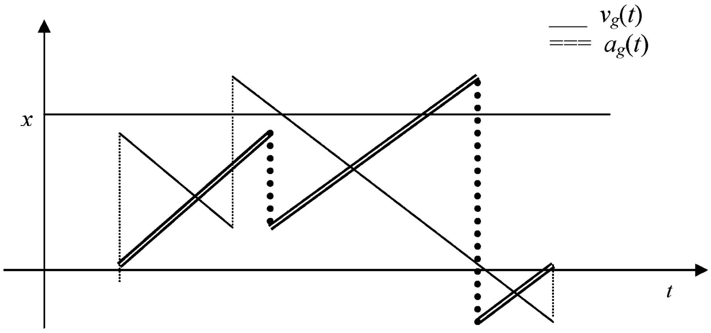

Figure 5. $v_g(t)$ and $a_g(t)$ in a busy cycle.

Let v_g be the generic random variable of the total workload in the queueing system in an arbitrary period. From Lemma 4.1, it is clear that v_g and a_g have the same distribution.

Corollary 4.2. In steady state, the generalized total workload v_g and the generalized age a_g in an arbitrary period have the same distribution, which is given in equation (4.1).

Denote by $a(t)$ the age of the batch in service at an arbitrary time. The process $a(t)$ can be obtained by only observing the generalized age process $a_g(t)$ when $a_g(t) \geq 0$. It is easy to see that $\{(a(t), I_a(t), J(t), I_s(t)), t \geq 0\}$ is a Markov chain with a transition probability matrix $P_{GI/M/1}$ given by equation (3.16), which is of the *GI/M/*1 type. Using Lemmas 3.1 and 3.2 and Neuts condition, it can be shown that the Markov chain $\{(a(t), I_a(t), J(t), I_s(t)), t \geq 0\}$ is ergodic if and only if $\rho < 1$. The steady state distribution $(\pi_a(0), \pi_a(1), \ldots, \pi_a(n), \ldots)$ of $\{(a(t), I_a(t), J(t), I_s(t)), t \geq 0\}$ can be obtained as the conditional distribution of $\{(a_g(t), I_a(t), J(t), I_s(t)), t \geq 0\}$, given that $a_g(t) \geq 0$:

$$\pi_a(n) = \left(\sum_{i=0}^{\infty} \pi(i) \right)^{-1} \pi(n) = \theta_{\text{tot}}(I - R)R^n, \quad n \geq 0. \tag{4.3}$$

To study the age process of the batch in service and the sojourn times, one can concentrate on $\{(a(t), I_a(t), J(t), I_s(t)), t \geq 0\}$. In fact, this approach was taken by Van Houdt and Blondia [35] while studying the sojourn times (i.e., delay times) in the discrete time *MMAP*[*K*]/*PH*[*K*]/1 queue.

4.2. Distributions of sojourn times

We define the *sojourn time* of an arbitrary batch (an arbitrary type J batch or an arbitrary type k customer) as the time between its arrival and its service completion (of all customers in the batch or that type k customer). Let d be the generic random variable for the sojourn time of an arbitrary batch in steady state. Let d_J be the generic random variable for the sojourn time of a type J batch in steady state. Let $d_{(k)}$ be the generic random variable for the sojourn time of a type k customer in steady state.

We introduce some extra notation first. We decompose the vector $\mathbf{T}^0_{\text{tot}}$ into vectors $\{\mathbf{T}^0_{\text{tot},J}\colon J \in \aleph\}$, where $\mathbf{T}^0_{\text{tot},J}$ is obtained by setting all elements in $\mathbf{T}^0_{\text{tot}}$ to zero, except these corresponding to the batch J (i.e., $\mathbf{T}^0_J$). Apparently, we have $\mathbf{T}^0_{\text{tot}} = \sum_{J\in\aleph} \mathbf{T}^0_{\text{tot},J}$. We construct column vectors $\{\mathbf{T}^0_{\text{tot},(k)}\colon 1 \le k \le K\}$ by putting the vector $\mathbf{T}^0_k$ into a column vector of the size m_{tot} on these places corresponding to the service of a type k customer. That is: if we divide $\mathbf{T}^0_{\text{tot},(k)}$ into (columns) vectors $\{\mathbf{T}^0_{\text{tot},(k)}(1, J_1, j_1), \ldots, \mathbf{T}^0_{\text{tot},(k)}(i, J_n, j_t), \ldots, \mathbf{T}^0_{\text{tot},(k)}(m_a, J_N, j_{|J_N|})\}$, then these vectors are zero except $\mathbf{T}^0_{\text{tot},(k)}(i, J_n, j_t) = \mathbf{T}^0_k$ if $j_t = k$.

Lemma 4.3. $\boldsymbol{\theta}_{\text{tot}}(\mathbf{e}\otimes\mathbf{T}^0_{\text{tot}}) = \lambda/\rho$, $\boldsymbol{\theta}_{\text{tot}}(\mathbf{e}\otimes\mathbf{T}^0_{\text{tot},J}) = \lambda_J/\rho$, and $\boldsymbol{\theta}_{\text{tot}}(\mathbf{e}\otimes\mathbf{T}^0_{\text{tot},(k)}) = \lambda_{(k)}/\rho$.

Proof. By definition,

$$\begin{aligned}\boldsymbol{\theta}_{\text{tot}}(\mathbf{e}\otimes\mathbf{T}^0_{\text{tot}}) &= \frac{\lambda}{\rho}\left(\sum_{i=1}^{N}(\boldsymbol{\theta}_a D_{a,J_i})\otimes\left(\frac{\boldsymbol{\beta}(J_i)}{\mu_{J_i}}\right)\right)\left(\mathbf{e}\otimes\mathbf{T}^0_{\text{tot}}\right)\\ &= \frac{\lambda}{\rho}\sum_{i=1}^{N}\frac{(\boldsymbol{\theta}_a D_{a,J_i}\mathbf{e})\big(\boldsymbol{\beta}_{J_i}\mathbf{T}^0_{J_i}\big)}{\mu_{J_i}} = \frac{\lambda}{\rho}\sum_{i=1}^{N}\boldsymbol{\theta}_a D_{a,J_i}\mathbf{e} = \frac{\lambda}{\rho}.\end{aligned} \tag{4.4}$$

Similarly, it can be shown that $\boldsymbol{\theta}_{\text{tot}}(\mathbf{e}\otimes\mathbf{T}^0_{\text{tot},J}) = \lambda_J/\rho$, and $\boldsymbol{\theta}_{\text{tot}}(\mathbf{e}\otimes\mathbf{T}^0_{\text{tot},(k)}) = \lambda_{(k)}/\rho$. □

By Lemma 4.3, λ/ρ, λ_J/ρ, and $\lambda_{(k)}/\rho$ can be interpreted as the average service completion rates of arbitrary batches, type J batches, and type k customers, respectively, given that the server is busy. For instance, the service completion rate of an arbitrary batch can be calculated as: $\sum_{n=0}^{\infty}\boldsymbol{\pi}_a(n)(\mathbf{e}\otimes\mathbf{T}^0_{\text{tot}}) = \boldsymbol{\theta}_{\text{tot}}(\mathbf{e}\otimes\mathbf{T}^0_{\text{tot}}) = \lambda/\rho$, where $\{\boldsymbol{\pi}_a(0), \boldsymbol{\pi}_a(1), \ldots\}$ is given in equation (4.3).

Consider the sojourn time of an arbitrary type J batch. Conditioning on the service completion of a batch (customer), in steady state, we have, for $n \ge 1$,

$$\begin{aligned}P\{d_J = n\} &= P\{a_g(t) = n-1|\text{Departure of type } J \text{ batch occurs next}\}\\ &= \frac{P\{a_g(t) = n-1, \text{Departure of type } J \text{ batch occurs next}\}}{P\{\text{ Departure of type } J \text{ batch occurs next}\}}\\ &= \frac{1}{\lambda_J}\sum_{I_a(t),J(t),I_s(t)} P\{a_g(t) = n-1,\, I_a(t),\, J(t),\, I_s(t),\\ &\quad \text{Departure of type J batch occurs next}\}\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{\lambda_J} \sum_{I_a(t),J(t),I_s(t)} P\{a_g(t) = n-1, I_a(t), J(t), I_s(t)\} \\
&\quad \cdot P\{\text{Departure of type } J \text{ batch occurs next } | I_a(t), J(t), I_s(t)\} \\
&= \frac{1}{\lambda_J} \boldsymbol{\pi}(n-1)\left(\mathbf{e} \otimes \mathbf{T}^0_{\text{tot},J}\right) = \frac{\rho}{\lambda_J} \boldsymbol{\theta}_{\text{tot}}(I-R)R^{n-1}\left(\mathbf{e} \otimes \mathbf{T}^0_{\text{tot},J}\right). \qquad (4.5)
\end{aligned}
$$

Note that in equation (4.5), we used the fact that the departure rate of type J batches equals the arrival rate of type J batches. Similarly, we have, for $1 \le k \le K$ and $n \ge 1$,

$$
\begin{aligned}
P\{d = n\} &= \frac{1}{\lambda} \boldsymbol{\pi}(n-1)\left(\mathbf{e} \otimes \mathbf{T}^0_{\text{tot}}\right) = \frac{\rho}{\lambda} \boldsymbol{\theta}_{\text{tot}}(I-R)R^{n-1}\left(\mathbf{e} \otimes \mathbf{T}^0_{\text{tot}}\right); \\
P\left\{d_{(k)} = n\right\} &= \frac{1}{\lambda_{(k)}} \boldsymbol{\pi}(n-1)\left(\mathbf{e} \otimes \mathbf{T}^0_{\text{tot},(k)}\right) = \frac{\rho}{\lambda_{(k)}} \boldsymbol{\theta}_{\text{tot}}(I-R)R^{n-1}\left(\mathbf{e} \otimes \mathbf{T}^0_{\text{tot},(k)}\right). \qquad (4.6)
\end{aligned}
$$

Note that in equation (4.6), we used the fact that the arrival rate of batches (type k customers) equals the service completion rate of batches (type k customers). By Lemma 4.3, it is easy to verify that the distributions given in equations (4.5) and (4.6) are proper probability distributions. Equations (4.5) and (4.6) indicate that the sojourn times have matrix geometric distributions. In Asmussen and O'Cinneide [3] and Sengupta [30], it was shown that the waiting time of an arbitrary customer and the sojourn times have continuous time PH-distributions for the continuous time *GI*/*PH*/*c* queue. These results can be extended to our queueing model.

Theorem 4.4. In steady state, for an arbitrary batch, the random variable d has a PH-distribution with a matrix representation

$$
\left\{ m_{d,\text{all}} = m_a m_{\text{tot}}, \quad \boldsymbol{\alpha}_{d,\text{all}} = \frac{\rho}{\lambda} \boldsymbol{\theta}_{\text{tot}}(I-R)\Delta_{d,\text{all}}, T_{d,\text{all}} = (\Delta_{d,\text{all}})^{-1} R \Delta_{d,\text{all}} \right\}, \qquad (4.7)
$$

where $\Delta_{d,\text{all}} = \text{diag}(\boldsymbol{\delta}_{d,\text{all}})$, $\boldsymbol{\delta}_{d,\text{all}} = (I-R)^{-1}(\mathbf{e} \otimes \mathbf{T}^0_{\text{tot}})$, and $\text{diag}(\boldsymbol{\delta}_{d,\text{all}})$ is a square matrix whose diagonal elements are from the vector $\boldsymbol{\delta}_{d,\text{all}}$ and all other elements are zero. For an arbitrary batch $J \in \aleph$, the random variable d_J has a PH-distribution with a matrix representation

$$
\left\{ m_{d,J} = m_a m_{\text{tot}}, \quad \boldsymbol{\alpha}_{d,J} = \frac{\rho}{\lambda_J} \boldsymbol{\theta}_{\text{tot}}(I-R)\Delta_{d,J}, T_{d,J} = (\Delta_{d,J})^{-1} R \Delta_{d,J} \right\}, \qquad (4.8)
$$

where $\Delta_{d,J} = \text{diag}(\boldsymbol{\delta}_{d,J})$ and $\boldsymbol{\delta}_{d,J} = (I-R)^{-1}(\mathbf{e} \otimes \mathbf{T}^0_{\text{tot},J})$. For an arbitrary type k customer, $1 \le k \le K$, the random variable $d_{(k)}$ has a PH-distribution with a matrix representation

$$
\left\{ m_{d,(k)} = m_a m_{\text{tot}}, \quad \boldsymbol{\alpha}_{d,(k)} = \frac{\rho}{\lambda_{(k)}} \boldsymbol{\theta}_{\text{tot}}(I-R)\Delta_{d,(k)}, T_{d,(k)} = (\Delta_{d,(k)})^{-1} R \Delta_{d,(k)} \right\}, \qquad (4.9)
$$

where $\Delta_{d,(k)} = \text{diag}(\boldsymbol{\delta}_{d,(k)})$ and $\boldsymbol{\delta}_{d,(k)} = (I-R)^{-1}(\mathbf{e} \otimes \mathbf{T}^0_{\text{tot},(k)})$.

Proof. We only give details for the sojourn time of an arbitrary type J batch. First, we show that the matrix representation given in equation (4.8) is for a discrete time PH-distribution. For that purpose, we only need to verify that $\Delta_{d,J}$ is invertible and the vector and the matrix in equation (4.8) are substochastic. According to Chapter 2 in Neuts [26], the elements of the matrix $I+R+R^2+\cdots=(I-R)^{-1}$ are the numbers of times that age of the batch in service is positive before it becomes zero. Thus, the vector $(I-R)^{-1}(\mathbf{e}\otimes\mathbf{T}^0_{\text{tot},J})$ is related to the probabilities that there will be a type J service completion (which is positive). Thus, every element of $(I-R)^{-1}(\mathbf{e}\otimes\mathbf{T}^0_{\text{tot},J})$ is positive. Then the matrix $\Delta_{d,J}$ is invertible. It is easy to see that the vector and the matrix in equation (4.8) are nonnegative. By Lemma 4.3 and routine calculations, we have

$$\begin{aligned}\frac{\rho}{\lambda_J}\boldsymbol{\theta}_{\text{tot}}(I-R)\Delta_{d,J}\mathbf{e}&=\frac{\rho}{\lambda_J}\boldsymbol{\theta}_{\text{tot}}\big(\mathbf{e}\otimes\mathbf{T}^0_{\text{tot},J}\big)=1;\\ (\Delta_{d,J})^{-1}R\Delta_{d,J}\mathbf{e}&=\mathbf{e}-(\Delta_{d,J})^{-1}\big(\mathbf{e}\otimes\mathbf{T}^0_{\text{tot},J}\big)\le\mathbf{e}.\end{aligned}\tag{4.10}$$

Therefore, the matrix representation given in (4.8) is for a discrete time PH-distribution. Second, we show that the PH-distribution is the same as that of the random variable $d_J$given in equation (4.5) by following calculations: for $n\ge 1$,

$$\begin{aligned}P\{d_J=n\}&=\frac{\rho}{\lambda_J}\boldsymbol{\pi}_a(n-1)\left(\mathbf{e}\otimes\mathbf{T}^0_{\text{tot},J}\right)=\frac{\rho}{\lambda}\boldsymbol{\theta}_{\text{tot}}(I-R)R^{n-1}\left(\mathbf{e}\otimes\mathbf{T}^0_{\text{tot},J}\right)\\ &=\frac{\rho}{\lambda_J}\boldsymbol{\theta}_{\text{tot}}(I-R)\Delta_{d,J}((\Delta_{d,J})^{-1}R\Delta_{d,J})^{n-1}(\Delta_{d,J})^{-1}\left(\mathbf{e}\otimes\mathbf{T}^0_{\text{tot},J}\right)\\ &=\frac{\rho}{\lambda_J}\boldsymbol{\theta}_{\text{tot}}(I-R)\Delta_{d,J}((\Delta_{d,J})^{-1}R\Delta_{d,J})^{n-1}(\mathbf{e}-(\Delta_{d,J})^{-1}R\Delta_{d,J}\mathbf{e}).\end{aligned}\tag{4.11}$$

This proves the conclusion for the sojourn time of an arbitrary type J batch. The other two cases can be shown similarly. This completes the proof of Theorem 4.4. □

4.3. *Distributions of waiting times*

In this section, we use the fact that the waiting time of a batch equals $a_g(t)$ just before that batch enters the server to find the distributions of waiting times and sojourn times of batches and individual types of customers. We focus on the waiting time w_J of an arbitrary type J batch. By definition, we have

$$\begin{aligned}P\{w_J=0\}&=P\{\nu_g\le 1|\text{An arrival of type } J \text{ batch occurs next}\}\\ &=\frac{P\{\nu_g\le 1,\text{ An arrival of type } J \text{ batch occurs next}\}}{P\{\text{An arrival of type } J \text{ batch occurs next}\}}\\ &=\frac{1}{\lambda_J}P\{a_g\le 0 \text{ after a departure } \textit{and} \text{ a type } J \text{ batch arrives}\}\end{aligned}\tag{4.12}$$

$$
\begin{aligned}
&= \frac{1}{\lambda_J} \sum_{n=0}^{\infty} \sum_{t=n+1}^{\infty} \boldsymbol{\pi}(n)\big(D_{a,J}(t)\mathbf{e} \otimes \mathbf{T}^0_{\text{tot}}\big) \\
&= \frac{1}{\lambda_J} \rho \boldsymbol{\theta}_{\text{tot}}(I - R) \sum_{t=1}^{\infty} \sum_{n=0}^{t-1} R^n \big(D_{a,J}(t)\mathbf{e} \otimes \mathbf{T}^0_{\text{tot}}\big) \\
&= \frac{\rho}{\lambda_J} \boldsymbol{\theta}_{\text{tot}} \sum_{t=1}^{\infty} (I - R^t)\big(D_{a,J}(t)\mathbf{e} \otimes \mathbf{T}^0_{\text{tot}}\big) = 1 - \frac{\rho}{\lambda_J} \boldsymbol{\theta}_{\text{tot}} R_J \mathbf{e},
\end{aligned}
$$

where $R_J = \sum_{t=1}^{\infty} R^t (D_{a,J}(t) \otimes \mathbf{T}^0_{\text{tot}} \boldsymbol{\alpha}(J))$ and we used $\boldsymbol{\theta}_{\text{tot}}(D_{a,J}\mathbf{e} \otimes \mathbf{T}^0_{\text{tot}}) = \lambda_J/\rho$ in the last equality, which can be proved as

$$
\begin{aligned}
\boldsymbol{\theta}_{\text{tot}}\left(D_{a,J}\mathbf{e} \otimes \mathbf{T}^0_{\text{tot}}\right) &= \frac{\lambda}{\rho} \left(\sum_{i=1}^{N} (\boldsymbol{\theta}_a D_{a,J_i} D_{a,J}\mathbf{e}) \left(\frac{\boldsymbol{\beta}_{J_i} \mathbf{T}^0_{J_i}}{\mu_{J_i}} \right) \right) \\
&= \frac{\lambda}{\rho} \sum_{i=1}^{N} \boldsymbol{\theta}_a D_{a,J_i} D_{a,J}\mathbf{e} = \frac{\lambda}{\rho} \boldsymbol{\theta}_a D_{a,J}\mathbf{e} = \frac{\lambda_J}{\rho}. \qquad (4.13)
\end{aligned}
$$

Similarly, we have, for $n \geq 1$,

$$
P\{w_J = n\} = \frac{1}{\lambda_J} \sum_{t=1}^{\infty} \boldsymbol{\pi}(n - 1 + t)\big(D_{a,J}(t)\mathbf{e} \otimes \mathbf{T}^0_{\text{tot}}\big) = \frac{\rho}{\lambda_J} \boldsymbol{\theta}_{\text{tot}}(I - R)R^{n-1}R_J\mathbf{e}. \qquad (4.14)
$$

It is easy to verify that the probabilities given in equations (4.12) and (4.14) constitute a proper probability distribution. From equations (4.12) and (4.14), it can be shown that w_J has a PH-distribution and its matrix representation can be constructed explicitly.

Theorem 4.5. In steady state, the random variable w_J has a PH-distribution with a matrix representation

$$
\left\{ m_{w,J} = m_a m_{\text{tot}}, \boldsymbol{\alpha}_{w,J} = \frac{\rho}{\lambda_J} \boldsymbol{\theta}_{\text{tot}}(I - R)\Delta_{w,J}, T_{w,J} = (\Delta_{w,J})^{-1} R \Delta_{w,J} \right\}, \qquad (4.15)
$$

where $\Delta_{w,J} = \text{diag}(\boldsymbol{\delta}_{w,J})$ and $\boldsymbol{\delta}_{w,J} = (I - R)^{-1} R_J \mathbf{e}$. Note that $\boldsymbol{\alpha}_{w,J}\mathbf{e} = 1 - P\{w_J = 0\} = \rho \boldsymbol{\theta}_{\text{tot}} R_J \mathbf{e}/\lambda_J$. For the waiting time w of an arbitrary batch, it has a PH-distribution with a matrix representation

$$
\left\{ m_{w,\text{all}} = m_a m_{\text{tot}}, \quad \boldsymbol{\alpha}_{w,\text{all}} = \frac{\rho}{\lambda} \boldsymbol{\theta}_{\text{tot}}(I - R)\Delta_{w,\text{all}}, T_{w,\text{all}} = \left(\Delta_{w,\text{all}}\right)^{-1} R \Delta_{w,\text{all}} \right\}, \qquad (4.16)
$$

where $\Delta_{w,\text{all}} = \text{diag}(\boldsymbol{\delta}_{w,\text{all}})$ and $\boldsymbol{\delta}_{w,\text{all}} = (I - R)^{-1}(R - A_0)\mathbf{e}$. Note that $\boldsymbol{\alpha}_{w,\text{all}}\mathbf{e} = \rho \boldsymbol{\theta}_{\text{tot}} (R - A_0)\mathbf{e}/\lambda$.

Proof. The proof is similar to that of Theorem 4.4. □

The sojourn time d_J of a type J batch is the sum of its waiting time and its service time, i.e., $d_J = w_J + s_J$. Thus, we can find the distribution of d_J by using w_J and s_J. Since both w_J and s_J have PH-distributions, their sum has a PH-distribution whose matrix representation can be constructed easily from that of w_J and s_J:

$$\begin{aligned}
m_{d(w),J} &= m_a m_{\text{tot}} + m_J;\\
\boldsymbol{\alpha}_{d(w),J} &= \left(\frac{\rho}{\lambda_J}\boldsymbol{\theta}_{\text{tot}}(I-R)\Delta_{w,J},\ \left(1-\frac{\rho}{\lambda_J}\boldsymbol{\theta}_{\text{tot}}R_J\mathbf{e}\right)\boldsymbol{\alpha}_J\right); \qquad (4.17)\\
T_{d(w),J} &= \begin{pmatrix} (\Delta_{w,J})^{-1}R\Delta_{w,J} & (I-(\Delta_{w,J})^{-1}R\Delta_{w,J})\mathbf{e}\boldsymbol{\alpha}_J \\ 0 & T_J \end{pmatrix}.
\end{aligned}$$

The construction of the above matrix representation is straightforward. Therefore, once the matrices R and R_J are obtained, the distributions of the waiting times and sojourn times can be computed. The distribution of the waiting time of an arbitrary type k customer is the sum of the waiting times of its corresponding batch and the service times of all customers in the same batch who are served first. Therefore, the waiting time of an arbitrary type k customer has a PH-distribution. By a similar argument, the sojourn time of an arbitrary type k customer has a PH-distribution. We do not intend to present all the details. Instead, we give two general formulas for computing the distributions of waiting times and sojourn times of individual types of customers. Let $w_{(k)}$ be the generic random variable for the waiting time of a type k customer in steady state. For $1 \le k \le K$, we have

$$\begin{aligned}
P\{w_{(k)} = n\} &= \sum_{J\in\aleph}\frac{\lambda_J}{\lambda_{(k)}}\sum_{t=1}^{|J|}P\{w_J + s_{j_1} + s_{j_2} + \cdots + s_{j_{t-1}} = n\}\delta_{\{j_t=k\}}, \quad n \ge 0;\\
& \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (4.18)\\
P\{d_{(k)} = n\} &= \sum_{J\in\aleph}\frac{\lambda_J}{\lambda_{(k)}}\sum_{t=1}^{|J|}P\{w_J + s_{j_1} + s_{j_2} + \cdots + s_{j_t} = n\}\delta_{\{j_t=k\}}, \quad n \ge 1.
\end{aligned}$$

where $\boldsymbol{\delta}_{\{.\}}$ is the indicator function, i.e., $\boldsymbol{\delta}_{\{j=k\}} = 1$ if $j = k$; 0, otherwise.

To end this section, we show the consistency of the results obtained in this section and that of Section 4.2. More specifically, we show that the distributions of the sojourn times obtained in Section 4.2 is consistent with that of $w_J + s_J$, where the distribution of w_J is given in this section. By equations (4.6), (4.12), and (4.14), we need to prove, for $n \ge 1$,

$$\begin{aligned}
P\{d_J = n\} = \frac{\rho}{\lambda_J}\boldsymbol{\theta}_{\text{tot}}(I-R)R^{n-1}\left(\mathbf{e}\otimes\mathbf{T}^0_{\text{tot},J}\right) &= P\{w_J + s_J = n\}\\
&= \left(1-\frac{\rho}{\lambda_J}\boldsymbol{\theta}_{\text{tot}}R_J\mathbf{e}\right)\left(\boldsymbol{\alpha}_J T_J^{n-1}\mathbf{T}^0_J\right)\\
&\quad + \sum_{t=1}^{n-1}\frac{\rho}{\lambda_J}\boldsymbol{\theta}_{\text{tot}}(I-R)R^{n-1-t}R_J\mathbf{e}\left(\boldsymbol{\alpha}_J T_J^{t-1}\mathbf{T}^0_J\right). \qquad (4.19)
\end{aligned}$$

To show equation (4.19), by definitions (3.5) and (3.7), first note

$$\begin{aligned}(D_{a,J'}(t)\otimes \mathbf{T}^0_{\text{tot}}\boldsymbol{\alpha}(J'))\big(\mathbf{e}\otimes \mathbf{T}^0_{\text{tot},J}\big) = 0, \quad \text{if} \quad J \neq J'; \\ R_J\big(\mathbf{e}\otimes \mathbf{T}^{0,t-1}_{\text{tot},J}\big) = R_J\mathbf{e}\left(\boldsymbol{\alpha}_J T_J^{t-1}\mathbf{T}^0_J\right), \quad t \geq 1. \qquad (4.20)\end{aligned}$$

where the vector $\mathbf{T}^{0,n}_{\text{tot},J}$ is obtained by replacing $\mathbf{T}^0_J$ in the vector $\mathbf{T}^0_{\text{tot},J}$ by $(T_J)^n\mathbf{T}^0_J$. By equation (4.20) and induction, it can be shown that

$$\begin{aligned}\big(D_{a,J}(t)\mathbf{e}\otimes \mathbf{T}^0_{\text{tot}}\big)\left(\boldsymbol{\alpha}_J\mathbf{T}^0_J\right) = \big(D_{a,J}(t)\otimes \mathbf{T}^0_{\text{tot}}\boldsymbol{\alpha}(J)\big)\big(\mathbf{e}\otimes \mathbf{T}^0_{\text{tot},J}\big); \\ R^n\left(\mathbf{e}\otimes \mathbf{T}^0_{\text{tot},J}\right) = \sum_{t=1}^{n} R^{n-t}R_J\big(\mathbf{e}\otimes \mathbf{T}^{0,t-1}_{\text{tot},J}\big) + \mathbf{e}\otimes \mathbf{T}^{0,n}_{\text{tot},J}, n \geq 1,\end{aligned} \qquad (4.21)$$

For $n = 1$ in equation (4.19), we have

$$\begin{aligned}P\{w_J + s_J = 1\} &= P\{w_J = 0\}P\{s_J = 1\} \\ &= \frac{\rho}{\lambda_J}\boldsymbol{\theta}_{\text{tot}}\sum_{t=1}^{\infty}(I - R^t)\big(D_{a,J}(t)\mathbf{e}\otimes \mathbf{T}^0_{\text{tot}}\big)\left(\boldsymbol{\alpha}_J\mathbf{T}^0_J\right) \\ &= \frac{\rho}{\lambda_J}\boldsymbol{\theta}_{\text{tot}}\sum_{t=1}^{\infty}(I - R^t)\big(D_{a,J}(t)\otimes \mathbf{T}^0_{\text{tot}}\boldsymbol{\alpha}(J)\big)\left(\mathbf{e}\otimes \mathbf{T}^0_{\text{tot},J}\right) \\ &= \frac{\rho}{\lambda_J}\boldsymbol{\theta}_{\text{tot}}\sum_{t=1}^{\infty}(I - R^t)\left(\sum_{J'\in\aleph} D_{a,J'}(t)\otimes \mathbf{T}^0_{\text{tot}}\boldsymbol{\alpha}(J')\right)\left(\mathbf{e}\otimes \mathbf{T}^0_{\text{tot},J}\right) \\ &= \frac{\rho}{\lambda_J}\boldsymbol{\theta}_{\text{tot}}\sum_{t=1}^{\infty}(I - R^t)A_t\left(\mathbf{e}\otimes \mathbf{T}^0_{\text{tot},J}\right) = \frac{\rho}{\lambda_J}\boldsymbol{\theta}_{\text{tot}}(A - R)\left(\mathbf{e}\otimes \mathbf{T}^0_{\text{tot},J}\right) \\ &= \frac{\rho}{\lambda_J}\boldsymbol{\theta}_{\text{tot}}(I - R)\left(\mathbf{e}\otimes \mathbf{T}^0_{\text{tot},J}\right) = P\{d_J = 1\}. \qquad (4.22)\end{aligned}$$

For $n \geq 2$, we have

$$\begin{aligned}P\{w_J + s_J = n\} &= \frac{\rho}{\lambda_J}\boldsymbol{\theta}_{\text{tot}}(I - R)\sum_{t=1}^{n-1} R^{n-1-t}R_J\mathbf{e}\left(\boldsymbol{\alpha}_J T_J^{t-1}\mathbf{T}^0_J\right) \\ &\quad + \frac{\rho}{\lambda_J}\boldsymbol{\theta}_{\text{tot}}\sum_{t=1}^{\infty}(I - R^t)(D_{a,J}(t)\mathbf{e}\otimes \mathbf{T}^0_{\text{tot}})\left(\boldsymbol{\alpha}_J T_J^{n-1}\mathbf{T}^0_J\right) \\ &= \frac{\rho}{\lambda_J}\boldsymbol{\theta}_{\text{tot}}(I - R)\sum_{t=1}^{n-1} R^{n-1-t}R_J\mathbf{e}\left(\boldsymbol{\alpha}_J T_J^{t-1}\mathbf{T}^0_J\right) \\ &\quad + \frac{\rho}{\lambda_J}\boldsymbol{\theta}_{\text{tot}}(A - R)\big(\mathbf{e}\otimes \mathbf{T}^{0,n-1}_{\text{tot},J}\big)\end{aligned}$$

$$
\begin{aligned}
&= \frac{\rho}{\lambda_J}\boldsymbol{\theta}_{\mathrm{tot}}(I - R)\sum_{t=1}^{n-1} R^{n-1-t} R_J\big(\mathbf{e}\otimes \mathbf{T}^{0,t-1}_{\mathrm{tot},J}\big)\\
&\quad + \frac{\rho}{\lambda_J}\boldsymbol{\theta}_{\mathrm{tot}}(I - R)\big(\mathbf{e}\otimes \mathbf{T}^{0,n-1}_{\mathrm{tot},J}\big)\\
&= \frac{\rho}{\lambda_J}\boldsymbol{\theta}_{\mathrm{tot}}(I - R)R^{n-1}\left(\mathbf{e}\otimes \mathbf{T}^{0}_{\mathrm{tot},J}\right) = P\{d_J = n\}. \qquad (4.23)
\end{aligned}
$$

Therefore, the results obtained in Section 4.2 and this section are consistent. Note that in Sengupta [28], the consistency between d and $w + s$ was not shown (see Remark 7 in [28]). That issue (for the continuous case) is resolved in HE [19].

5. The discrete time *MMAP*[*K*]/*PH*[*K*]/1 Queue

In this section, we consider the discrete time *MMAP*[*K*]/*PH*[*K*]/1 queue, a special case of the model introduced in Section 2. This queueing system has a batch Markov arrival process with matrix representation $\{D_0, D_J, J \in \aleph\}$, where D_0 and D_J are $m_a \times m_a$ substochastic matrices. The matrix D_J is the (matrix) arrival rate of type J batches. For more about *MMAP*[*K*], see Asmussen and Koole [4], HE [15,17], and HE and Neuts [21]. The relationship between the two sets of parameters of the arrival process is: $D_{a,J}(t) = D_0^{t-1} D_J$, for $J \in \aleph, t \geq 1$. Let $D = D_0 + \Sigma_J D_J$, i.e., the transition probability matrix of the underlying Markov chain of the arrival process. We assume that D is irreducible and $D \neq D_0$. Denote by $\boldsymbol{\theta}$ the invariant probability vector of the stochastic matrix D. It is easy to see $\boldsymbol{\theta}_a = \boldsymbol{\theta}(I - D_0)/\lambda$, where $\lambda = \boldsymbol{\theta}(I - D_0)\mathbf{e} = \boldsymbol{\theta}\Sigma_J D_J \mathbf{e}$. In addition, we have $\lambda_J = \boldsymbol{\theta} D_J \mathbf{e}$. We assume that λ and $\{\lambda_J, J \in \aleph\}$ are all positive.

For this queueing system, both the arrival and service processes are governed by Markov chains. By utilizing detailed information about these Markov chains, a QBD process can be introduced to describe the age process of the batch in service and the total workload process. This method was used in Van Houdt and Blondia [37]. In this section, we link the QBD process to the total workload process and obtain more detailed results on the steady state distributions.

5.1. *The Markov chain* $\{(r(t), I_a(t), J(t), I_s(t)), t \geq 0\}$

We construct a *fictitious* process $r(t)$ that increases with service (like the age process) and decreases with arrival (like the total workload process) alternatively. If the process is in a service leg in period t and the service continues into period $t + 1$, then $r(t + 1) = r(t) + 1$. If the process is in a service leg and a service is completed in period t, then the process switches to an arrival leg in period $t + 1$ and $r(t + 1) = r(t)$. If the process is in an arrival leg in period t and there is no arrival in period t, then $r(t + 1) = \max\{0, r(t) - 1\}$. If the process is in an arrival leg and there is a batch arrival in period t, then the process switches to a service leg in period $t + 1$ and $r(t + 1) = r(t)$.

We compare the processes $\{r(t), t \geq 0\}$ and $\{w_n, n \geq 0\}$. For $\{w_n, n \geq 0\}$, the process jumps from an arrival epoch to the next: $\max\{0, w_n + s_{J_n} - \tau_{n+1}\}$. For $\{r(t), t \geq 0\}$, it takes two legs to go from an arrival epoch to the next. The first leg is the *service leg* of the length s_{J_n} and the second leg is the *arrival leg* of the length τ_{n+1}. During the service leg,$r(t)$ increases its value by one in each period. At the end of this leg, the value of $r(t)$ becomes $w_n + s_{J_n}$. During the arrival leg, $r(t)$ decreases its value by one in each period (if nonnegative). At the end of this leg, the value of $r(t)$ becomes $w_{n+1} = \max\{0, w_n + s_{J_n} - \tau_{n+1}\}$ and the next cycle begins. Since the arrival and service legs are considered separately, unlike the generalized age process introduced in Section 3, the time t for $r(t)$ is not for the real operating time of the queueing system.

Auxiliary variables $\{I_a(t), J(t),I_s(t)\}$ are defined as follows. During a service leg, the phase of the arrival process $I_a(t)$ remains constant, which is the initial phase of the next arrival leg. During an arrival leg, the phase of the service process $(J(t), I_s(t))$ takes value zero. The initial phase of the next service leg is determined by the initial phase distribution of the next PH-service time. Therefore, the set of values of $\{I_a(t), J(t), I_s(t)\}$ is extended from equation (3.4) to

$$\{(i, 0) : \quad 1 \leq i \leq m_a\} \cup \{(i, J_n, j) : \quad 1 \leq i \leq m_a, 1 \leq j \leq m_{j_n}, 1 \leq n \leq N\}. \tag{5.1}$$

Note that if $r(t) = 0$, there is no service in the system and $(J(t), I_s(t))$ takes value zero.

Since both the arrival process and the service times are governed by Markov chains, it is easy to see that $\{(r(t), I_a(t), J(t), I_s(t)), t \geq 0\}$ is a Markov chain. It is readily seen that $\{r(t), t \geq 0\}$ is skip-free to both left and right. Therefore, $\{(r(t), I_a(t), J(t), I_s(t)), t \geq 0\}$ is a quasi-birth-and-death (QBD) process. A typical sample path of $r(t)$ is shown in figure 6. When $r(t) = 0$, there is no customer in the system. Thus, if $r(t) = 0$, there is no service. The transition probability matrix of the Markov chain $\{(r(t), I_a(t), J(t), I_s(t)),$

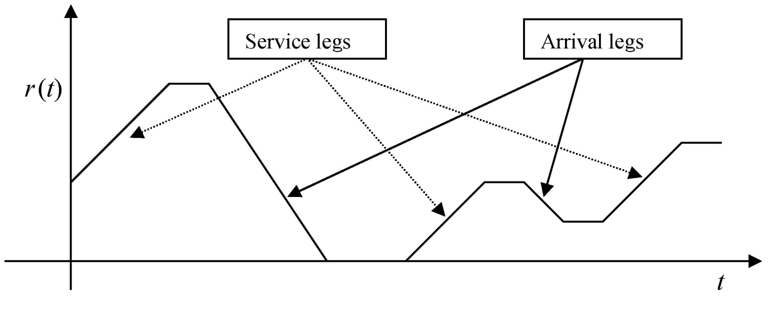

Figure 6. A sample path of $r(t)$.

$t \geq 0\}$ is given by

$$P_{\text{QBD}} = \begin{pmatrix} D_0 & \tilde{A}_0 & & & \\ \tilde{A}_2 & A_1 & A_0 & & \\ & A_2 & A_1 & A_0 & \\ & & \ddots & \ddots & \ddots \end{pmatrix}, \tag{5.2}$$

where

$$A_0 = \begin{pmatrix} 0 & 0 \\ 0 & I \otimes T_{\text{tot}} \end{pmatrix}, \quad A_1 = \begin{pmatrix} 0 & \sum_{i=1}^{N} D_{J_i} \otimes \boldsymbol{\alpha}(J_i) \\ I \otimes \mathbf{T}_{\text{tot}}^0 & 0 \end{pmatrix}, \quad A_2 = \begin{pmatrix} D_0 & 0 \\ 0 & 0 \end{pmatrix},$$
$$\tilde{A}_2 = \begin{pmatrix} D_0 \\ 0 \end{pmatrix}, \quad \tilde{A}_0 = \begin{pmatrix} 0 & \sum_{i=1}^{N} D_{J_i} \otimes \boldsymbol{\alpha}(J_i) \end{pmatrix}, \tag{5.3}$$

and A_0, A_1, and A_2 are $(m_a(1+m_{\text{tot}})) \times (m_a(1+m_{\text{tot}}))$ matrices. The vectors $\{\boldsymbol{\alpha}(J_i), 1 \leq i \leq N\}$ were defined in Section 3. For the Markov chain $\{(r(t), I_a(t), J(t), I_s(t)), t \geq 0\}$ with transition probability matrix P_{QBD}, $r(t)$ represents the total workload if $(J(t), I_s(t)) = 0$ and $r(t) - 1$ represents the age of the batch in service if $(J(t), I_s(t)) \neq 0$. First, we show that the Markov chain $\{(r(t), I_a(t), J(t), I_s(t)), t \geq 0\}$ is ergodic if the queueing system is stable. Denote by

$$\begin{aligned} \boldsymbol{\theta}_{\text{tot},a} &= \frac{1}{\rho} \sum_{J \in \aleph} (\boldsymbol{\theta} D_J) \otimes \left(\frac{\boldsymbol{\beta}(J)}{\mu_J} \right); \\ \boldsymbol{\theta}_{\text{QBD}} &= \frac{\left(\boldsymbol{\theta}, \sum_{J \in \aleph} (\boldsymbol{\theta} D_J) \otimes (\frac{\boldsymbol{\beta}(J)}{\mu_J})\right)}{(1+\rho)} = \frac{(\boldsymbol{\theta}, \rho \boldsymbol{\theta}_{\text{tot},a})}{(1+\rho)}. \end{aligned} \tag{5.4}$$

It is easy to verify that both $\boldsymbol{\theta}_{\text{tot},a}$ and $\boldsymbol{\theta}_{\text{QBD}}$ are probability vectors. Let $A = A_0 + A_1 + A_2$. Then

$$A = \begin{pmatrix} D_0 & \sum_{i=1}^{N} D_{J_i} \otimes \boldsymbol{\alpha}(J_i) \\ I \otimes \mathbf{T}_{\text{tot}}^0 & I \otimes T_{\text{tot}} \end{pmatrix}. \tag{5.5}$$

Theorem 5.1. Assume that the Markov chain $\{(r(t), I_a(t), J(t), I_s(t)), t \geq 0\}$ is irreducible and aperiodic. The vector $\boldsymbol{\theta}_{\text{QBD}}$ is the invariant probability vector of A and $\boldsymbol{\theta}_{\text{QBD}} A_2 \mathbf{e} - \boldsymbol{\theta}_{\text{QBD}} A_0 \mathbf{e} = (1-\rho)/(1+\rho)$. The Markov chain is positive recurrent if and only if $\rho < 1$.

Proof. By routine calculations, we have, for $\boldsymbol{\theta}_{\text{QBD}}A$,

$$\begin{aligned}(1+\rho)^{-1}\left(\boldsymbol{\theta}D_0+\sum_{i=1}^{N}\left(\boldsymbol{\theta}D_{J_i}\right)\left(\frac{\boldsymbol{\beta}_{J_i}\mathbf{T}^0_{J_i}}{\mu_{J_i}}\right)\right)&=(1+\rho)^{-1}\left(\boldsymbol{\theta}D_0+\sum_{i=1}^{N}\boldsymbol{\theta}D_{J_i}\right)\\&=(1+\rho)^{-1}\,\boldsymbol{\theta},\end{aligned}\tag{5.6}$$

and

$$\begin{aligned}&(1+\rho)^{-1}\left(\boldsymbol{\theta}\sum_{i=1}^{N}D_{J_i}\otimes\boldsymbol{\alpha}(J_i)+\sum_{i=1}^{N}(\boldsymbol{\theta}D_{J_i})\otimes\left(\frac{\boldsymbol{\beta}(J_i)T_{\text{tot}}}{\mu_{J_i}}\right)\right)\\&=(1+\rho)^{-1}\left(\sum_{i=1}^{N}\left(\boldsymbol{\theta}D_{J_i}\right)\otimes\left(0,\ldots,0,\boldsymbol{\alpha}_{J_i}+\frac{\boldsymbol{\beta}_{J_i}T_{J_i}}{\mu_{J_i}},0,\ldots,0\right)\right)\\&=(1+\rho)^{-1}\left(\sum_{i=1}^{N}\left(\boldsymbol{\theta}D_{J_i}\right)\otimes\left(0,\ldots,0,\frac{\boldsymbol{\beta}_{J_i}}{\mu_{J_i}},0,\ldots,0\right)\right)\\&=(1+\rho)^{-1}\left(\sum_{i=1}^{N}\left(\boldsymbol{\theta}D_{J_i}\right)\otimes\left(\frac{\boldsymbol{\beta}(J_i)}{\mu_{J_i}}\right)\right)=\frac{\rho}{1+\rho}\boldsymbol{\theta}_{\text{tot}}.\end{aligned}\tag{5.7}$$

Equations (5.6) and (5.7) imply $\boldsymbol{\theta}_{\text{QBD}}A=\boldsymbol{\theta}_{\text{QBD}}$, i.e., $\boldsymbol{\theta}_{\text{QBD}}$ is the invariant probability vector of A. By routine calculations,

$$\begin{aligned}\boldsymbol{\theta}_{\text{QBD}}A_2\mathbf{e}&=(1+\rho)^{-1}\boldsymbol{\theta}D_0\mathbf{e}=(1+\rho)^{-1}(1-\lambda);\\\boldsymbol{\theta}_{\text{QBD}}A_0\mathbf{e}&=(1+\rho)^{-1}\left(\sum_{i=1}^{N}(\boldsymbol{\theta}D_{J_i})\otimes\left(\frac{\boldsymbol{\beta}(J_i)}{\mu_{J_i}}\right)\right)\left(\mathbf{e}\otimes\left(\mathbf{e}-\mathbf{T}^0_{\text{tot}}\right)\right)\\&=(1+\rho)^{-1}\sum_{i=1}^{N}\frac{(\boldsymbol{\theta}D_{J_i}\mathbf{e})\left(\boldsymbol{\beta}_{J_i}\left(\mathbf{e}-\mathbf{T}^0_{J_i}\right)\right)}{\mu_{J_i}}=(1+\rho)^{-1}(\rho-\lambda).\end{aligned}\tag{5.8}$$

Therefore, $\boldsymbol{\theta}_{\text{QBD}}A_2\mathbf{e}-\boldsymbol{\theta}_{\text{QBD}}A_0\mathbf{e}=(1-\rho)/(1+\rho)$ and $\boldsymbol{\theta}_{\text{QBD}}A_2\mathbf{e}>\boldsymbol{\theta}_{\text{QBD}}A_0\mathbf{e}$ if and only if $\rho<1$. According to Neuts condition (Neuts [26]), the QBD process $\{(r(t), I_a(t), J(t), I_s(t)), t\geq 0\}$ is positive recurrent if and only if $\boldsymbol{\theta}_{\text{QBD}}A_0\mathbf{e} < \boldsymbol{\theta}_{\text{QBD}}A_2\mathbf{e}$. Therefore, the QBD process is positive recurrent if and only if $\rho < 1$. This completes the proof of Theorem 5.1. □

5.2. *Steady state distributions*

If the Markov chain $\{(r(t), I_a(t), J(t), I_s(t)), t\geq 0\}$ is ergodic, its steady state distribution has a matrix geometric form and can be found by routine methods. Denote by $\boldsymbol{\pi}=(\boldsymbol{\pi}(0), \boldsymbol{\pi}(1), \boldsymbol{\pi}(2), \ldots)$ the stationary distribution of $\{(r(t), I_a(t), J(t), I_s(t)), t\geq 0\}$, i.e., $\boldsymbol{\pi} = \boldsymbol{\pi}P_{\text{QBD}}$ and $\boldsymbol{\pi}\mathbf{e}=1$, where $\boldsymbol{\pi}(n)=(\pi_{i,J,j}(n))$ is a row vector of the size

$m_a(m_{\text{tot}}+1)$, $n \geq 0$. Let R be an $(m_a(m_{\text{tot}}+1)) \times (m_a(m_{\text{tot}}+1))$ matrix that is the minimal nonnegative solution to

$$R = A_0 + RA_1 + R^2 A_2. \tag{5.9}$$

We refer to Bini and Meini [5], Latouche and Ramaswami [22,23] and Neuts [26] for more about the matrix R. Since the Markov chain is positive recurrent, the spectral radius of R is less than one, which implies that the matrix $I - R$ is invertible. By the structure of the matrix A_0 (equation (5.3)), the matrix R also has a special structure

$$R = \begin{pmatrix} 0 & 0 \\ R_\nu & R_a \end{pmatrix}, \tag{5.10}$$

where R_ν and R_a are the minimal nonnegative solutions to the equations:

$$\begin{aligned} R_\nu &= R_a\big(I \otimes \mathbf{T}^0_{\text{tot}}\big) + R_a R_\nu D_0; \\ R_a &= I \otimes T_{\text{tot}} + R_\nu \sum_{J \in \aleph} D_J \otimes \boldsymbol{\alpha}(J). \end{aligned} \tag{5.11}$$

Apparently, the spectral radius of the matrix $R_a$equals the spectral radius of R and is less than one. According to Neuts [26], $\{\boldsymbol{\pi}(0), \boldsymbol{\pi}(1), \ldots\}$ has a matrix geometric distribution

$$\boldsymbol{\pi}(n) = \boldsymbol{\pi}(1) R^{n-1}, \quad n \geq 1, \tag{5.12}$$

where $\boldsymbol{\pi}(0)$ and $\boldsymbol{\pi}(1)$ satisfy equations:

$$\begin{aligned} (\boldsymbol{\pi}(0), \boldsymbol{\pi}(1)) &= (\boldsymbol{\pi}(0), \boldsymbol{\pi}(1)) \begin{pmatrix} D_0 & \tilde{A}_0 \\ \tilde{A}_2 & A_1 + RA_2 \end{pmatrix}; \\ \boldsymbol{\pi}(0)\mathbf{e} + \boldsymbol{\pi}(1)(I - R)^{-1}\mathbf{e} &= 1. \end{aligned} \tag{5.13}$$

Theorem 5.2. If the queueing system is stable and the Markov chain $\{(r(t), I_a(t), J(t), I_s(t)), t \geq 0\}$ is irreducible and aperiodic, the vectors $\boldsymbol{\pi}(0)$ and $\boldsymbol{\pi}(1)$ are given explicitly as

$$\begin{aligned} \boldsymbol{\pi}(0) &= \boldsymbol{\theta}_{\text{QBD}}(I - R)\tilde{A}_2 = \frac{1}{1+\rho}(\boldsymbol{\theta} - \rho\boldsymbol{\theta}_{\text{tot},a} R_\upsilon) D_0; \\ \boldsymbol{\pi}(1) &= \boldsymbol{\theta}_{\text{QBD}}(I - R)\begin{pmatrix} I - D_0 & 0 \\ 0 & I \end{pmatrix} \\ &= \left(\frac{1}{1+\rho}(\boldsymbol{\theta} - \rho\boldsymbol{\theta}_{\text{tot},a} R_\upsilon)(I - D_0), \frac{\rho}{1+\rho}\boldsymbol{\theta}_{\text{tot},a}(I - R_a)\right). \end{aligned} \tag{5.14}$$

Proof. By rewriting equalities in equation (5.13), we obtain

$$\begin{aligned}(\boldsymbol{\pi}(0), 0) &= (\boldsymbol{\pi}(0), 0)A_2 + \boldsymbol{\pi}(1)A_2;\\ \boldsymbol{\pi}(1) &= (\boldsymbol{\pi}(0), 0)A_1 + \boldsymbol{\pi}(1)(A_1 + RA_2).\end{aligned} \tag{5.15}$$

Note $\boldsymbol{\pi}(0)$ is extended to $(\boldsymbol{\pi}(0), 0)$ by adding $m_a m_{\text{tot}}$ zeros. Let $\mathbf{u} = (\boldsymbol{\pi}(0), 0) + \boldsymbol{\pi}(1)$. We first show $\mathbf{u} = \boldsymbol{\theta}_{\text{QBD}}(I - R)$. By equation (5.15) and $(\boldsymbol{\pi}(0), 0)R = 0$, we find that $\mathbf{u}$ satisfies equation $\mathbf{u} = \mathbf{u}(A_1 + (I + R)A_2)$, which leads to $\mathbf{u} = \mathbf{u}(I - R)^{-1}[A_1 - RA_1 + A_2 - R^2A_2] = \mathbf{u}(I - R)^{-1}(A - R)$. Therefore, the vector $\mathbf{u}(I - R)^{-1}$ satisfies equation: $\mathbf{u}(I - R)^{-1} = \mathbf{u}(I - R)^{-1}A$, which implies $\mathbf{u}(I - R)^{-1} = c\boldsymbol{\theta}_{\text{QBD}}$, where c is a constant. Thus, $\mathbf{u} = c\boldsymbol{\theta}_{\text{QBD}}(I - R)$. By the second equality in equation (5.13), we must have $\mathbf{u}(I - R)^{-1}\mathbf{e} = 1$. Thus, we must have $c = 1$. Therefore, $\mathbf{u} = \boldsymbol{\theta}_{\text{QBD}}(I - R)$. By equation (5.13), we have

$$\boldsymbol{\pi}(0) = \boldsymbol{\pi}(1)\tilde{A}_2(I - D_0)^{-1} \quad \textit{or} \quad (\boldsymbol{\pi}(0), 0) = \boldsymbol{\pi}(1)\begin{pmatrix} D_0(I - D_0)^{-1} & 0 \\ 0 & 0 \end{pmatrix}. \tag{5.16}$$

Combining $\mathbf{u} = \boldsymbol{\theta}_{\text{QBD}}(I - R)$ and equation (5.16), yields,

$$\boldsymbol{\theta}_{\text{QBD}}(I - R) = \boldsymbol{\pi}(1)\begin{pmatrix} D_0(I - D_0)^{-1} & 0 \\ 0 & 0 \end{pmatrix} + \boldsymbol{\pi}(1) = \boldsymbol{\pi}(1)\begin{pmatrix} (I - D_0)^{-1} & 0 \\ 0 & I \end{pmatrix}, \tag{5.17}$$

which leads to the expressions for $\boldsymbol{\pi}(1)$ and $\boldsymbol{\pi}(0)$ in equation (5.14). This completes the proof of Theorem 5.2. □

For later use, we divide vectors $\{\boldsymbol{\pi}(0), \boldsymbol{\pi}(1), \ldots\}$ in the following way: $\boldsymbol{\pi}(n) = (\boldsymbol{\pi}_v(n), \boldsymbol{\pi}_a(n))$, $n \geq 1$, where $\boldsymbol{\pi}_v(n)$ is a vector of dimension m_a and $\boldsymbol{\pi}_a(n)$ is a vector of dimension $m_a m_{\text{tot}}$. By equation (5.12) and Theorem 5.2, we obtain, for $n \geq 2$,

$$(\boldsymbol{\pi}_v(n), \boldsymbol{\pi}_a(n)) = \boldsymbol{\pi}(1)R^{n-1} = \frac{\rho}{1+\rho}\left(\boldsymbol{\theta}_{\text{tot},a}(I - R_a)R_a^{n-2}R_v, \quad \boldsymbol{\theta}_{\text{tot},a}(I - R_a)R_a^{n-1}\right). \tag{5.18}$$

5.3. *Age process, distributions of waiting times and sojourn times*

If we observe the process $r(t)$ only at its increasing legs, the excised process is related to the age process of an arbitrary batch in service. Recall that $r(t) - 1$ represents the age of the batch in service if $(J(t), I_s(t)) \neq 0$. The steady state distribution of the age process is given in the following proposition.

Proposition 5.3. Denote by $\bar{\pi}_a = (\bar{\pi}_a(0), \bar{\pi}_a(1), \ldots)$ the steady state distribution of the age process. Then

$$\bar{\pi}_a(n) = \boldsymbol{\theta}_{\text{tot},a}(I - R_a)R_a^n, \quad n \geq 0. \tag{5.19}$$

The distributions of the age of the batch in service and the total workload are the same, i.e., $P\{a = n\} = P\{v = n\} = \bar{\pi}_a(n)\mathbf{e}$, for $n \geq 0$.

Proof. By definition, we must have

$$\bar{\pi}_a = \left(\sum_{n=1}^{\infty} \pi_a(n)\mathbf{e}\right)^{-1} (\pi_a(1), \pi_a(2), \ldots). \tag{5.20}$$

Equation (5.19) is obtained immediately from equations (5.14) and (5.18). The distributions of the age of the batch in service and the total workload are the same. This completes the proof of Proposition 5.3. □

By using equations (5.9) and (5.14) and the relationship between the two sets of parameters of the batch arrival process $D_{a,J}(t) = D_0^{t-1}D_J, for J \in \aleph, t \geq 1$, it can be shown that the matrix R_a equals R defined in equation (3.23) and $\boldsymbol{\theta}_{\text{tot},a}$ equals $\boldsymbol{\theta}_{\text{tot}}$ defined in equation (3.11). Similar to Sections 4.2 and 4.3, the distributions of sojourn times and waiting times of batches and individual customers can be obtained from $\bar{\pi}_a = (\bar{\pi}_a(0), \bar{\pi}_a(1), \ldots)$. The matrix representations of these PH-distributions can be constructed by replacing R by R_a and $\boldsymbol{\theta}_{\text{tot}}$ by $\boldsymbol{\theta}_{\text{tot},a}$ in Theorems 4.2 and 4.3. Details are omitted.

5.4. Workload process, distributions of waiting times and sojourn times

In this section, we use the total workload process to find the distributions of waiting times and sojourn times. If we observe the process *r*(*t*) only in decreasing legs, the excised process represents the total workload process. The steady state distribution of this excised process is given by

$$\bar{\pi}_v = \left(\sum_{n=0}^{\infty} \pi_v(n)\mathbf{e}\right)^{-1} (\pi_v(0), \pi_v(1), \ldots). \tag{5.21}$$

By equations (5.14) and (5.15) and routine calculations, we obtain the following results.

Proposition 5.4. Denote by $\bar{\pi}_v = (\bar{\pi}_v(0), \bar{\pi}_v(1), \ldots)$ the steady state distribution of the total workload process. Then

$$\bar{\pi}_v(n) = \begin{cases} (\boldsymbol{\theta} - \rho\boldsymbol{\theta}_{\text{tot},a}R_v)D_0, & n = 0; \\ (\boldsymbol{\theta} - \rho\boldsymbol{\theta}_{\text{tot},a}R_v)(I - D_0), & n = 1; \\ \rho\boldsymbol{\theta}_{\text{tot},a}(I - R_a)R_a^{n-2}R_v, & n \geq 2. \end{cases} \tag{5.22}$$

Based on the total workload process, the distributions of the waiting times of an arbitrary batch and an arbitrary customer are given, respectively, as follows, for $J \in \aleph$, $1 \leq k \leq K$,

$$P\{v = n\} = \bar{\pi}_v(n)\mathbf{e}, \quad n \geq 0. \tag{5.23}$$

$$P\{w = n\} = \begin{cases} \frac{1}{\lambda}\left(\bar{\pi}_v(0) + \bar{\pi}_v(1)\right)(I - D_0)\mathbf{e} = 1 - \frac{\rho}{\lambda}\boldsymbol{\theta}_{\text{tot}}R_v(I - D_0)\mathbf{e}, & n = 0; \\ \frac{1}{\lambda}\bar{\pi}_v(n+1)(I - D_0)\mathbf{e} = \frac{\rho}{\lambda}\boldsymbol{\theta}_{\text{tot}}(I - R_a)R_a^{n-1}R_v(I - D_0)\mathbf{e}, & n \geq 1. \end{cases} \tag{5.24}$$

$$P\{w_J = n\} = \begin{cases} \frac{1}{\lambda_J}\left(\bar{\pi}_v(0) + \bar{\pi}_v(1)\right)D_J\mathbf{e} = 1 - \frac{\rho}{\lambda_J}\boldsymbol{\theta}_{\text{tot}}R_vD_J\mathbf{e}, & n = 0; \\ \frac{1}{\lambda_J}\bar{\pi}_v(n+1)D_J\mathbf{e} = \frac{\rho}{\lambda_J}\boldsymbol{\theta}_{\text{tot}}(I - R_a)R_a^{n-1}R_vD_J\mathbf{e}, & n \geq 1. \end{cases} \tag{5.25}$$

Note that we used $\lambda = \boldsymbol{\theta}(I - D_0)\mathbf{e}$ and $\lambda_J = \boldsymbol{\theta}D_J\mathbf{e}$ in equations (5.24) and (5.25), respectively. It is easy to verify that the distributions given in equations (5.23), (5.24), and (5.25) are proper probability distributions. Based on Proposition 5.4 and equations (5.23), (5.24), and (5.25), the matrix representations of these PH-distributions can be constructed.

Theorem 5.5. The total workload at an arbitrary time (or the age of the batch in service) has a PH-distribution with matrix representation

$$\begin{aligned} m_v &= m_a m_{\text{tot}} + 2; \\ \boldsymbol{\alpha}_v &= ((\boldsymbol{\theta} - \rho\boldsymbol{\theta}_{\text{tot},a}R_v)(I - D_0)\mathbf{e}, \quad \rho\boldsymbol{\theta}_{\text{tot},a}R_v\mathbf{e}, \quad 0); \\ T_v &= \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & (\boldsymbol{\theta}_{\text{tot},a}R_v\mathbf{e})^{-1}\boldsymbol{\theta}_{\text{tot},a}(I - R_a)\Delta_v \\ 0 & 0 & (\Delta_v)^{-1}R_a\Delta_v \end{pmatrix}, \end{aligned} \tag{5.26}$$

where $\Delta_v = \text{diag}(\boldsymbol{\delta}_v)$ and $\boldsymbol{\delta}_v = (I - R_a)^{-1} R_v \mathbf{e}$. The waiting time of an arbitrary batch has a PH-distribution with matrix representation

$$\left\{ m_{w,\text{all}} = m_a m_{\text{tot}}; \quad \boldsymbol{\alpha}_{w,\text{all}} = \frac{\rho}{\lambda} \boldsymbol{\theta}_{\text{tot},a}(I - R_a)\Delta_{w,\text{all}}; \quad T_{w,\text{all}} = (\Delta_{w,\text{all}})^{-1} R_a \Delta_{w,\text{all}} \right\}, \tag{5.27}$$

where $\Delta_{w,\text{all}} = \text{diag}(\boldsymbol{\delta}_{w,\text{all}})$ and $\boldsymbol{\delta}_{w,\text{all}} = (I - R_a)^{-1} R_v (I - D_0)\mathbf{e}$. The waiting time of an arbitrary type J batch has PH-distribution with matrix representation

$$\left\{ m_{w,J} = m_a m_{\text{tot}}; \quad \boldsymbol{\alpha}_{w,J} = \frac{\rho}{\lambda_J} \boldsymbol{\theta}_{\text{tot},a}(I - R_a)\Delta_{w,J}; \quad T_{w,J} = (\Delta_{w,J})^{-1} R_a \Delta_{w,J} \right\}, \tag{5.28}$$

where $\Delta_{w,J} = \text{diag}(\boldsymbol{\delta}_{w,J})$ and $\boldsymbol{\delta}_{w,J} = (I - R_a)^{-1} R_v D_J \mathbf{e}$.

Proof. The proofs of (5.27) and (5.28) are similar to that of Theorem 4.4. The distribution of the total workload is given explicitly as

$$P\{v = n\} = \begin{cases} (\boldsymbol{\theta} - \rho\boldsymbol{\theta}_{\text{tot},a} R_v) D_0 \mathbf{e}, & n = 0; \\ (\boldsymbol{\theta} - \rho\boldsymbol{\theta}_{\text{tot},a} R_v)(I - D_0)\mathbf{e}, & n = 1; \\ (\rho\boldsymbol{\theta}_{\text{tot},a} R_v \mathbf{e}) \dfrac{\boldsymbol{\theta}_{\text{tot},a}(I - R_a) R_a^{n-2} R_v \mathbf{e}}{\boldsymbol{\theta}_{\text{tot},a} R_v \mathbf{e}}, & n \geq 2. \end{cases} \tag{5.29}$$

Apparently, the above distribution is the mixture of two discrete time PH-distributions, which is a discrete time PH-distribution as well. The matrix representation of the mixture is obtained easily from the matrix representations of the two PH-distributions. Details are omitted. This completes the proof of Theorem 5.5. □

Once the distributions of waiting times are obtained, the distributions of sojourn times can be found. For example, the sojourn time of an arbitrary type J batch has a PH-distribution with matrix representation

$$\begin{aligned} m_{d(w),J} &= m_a m_{\text{tot}} + m_J; \\ \boldsymbol{\alpha}_{d(w),J} &= \frac{1}{\lambda_J} (\rho\boldsymbol{\theta}_{\text{tot},a}(I - R_a)\Delta_{w,J}, \quad (\boldsymbol{\theta} - \rho\boldsymbol{\theta}_{\text{tot},a} R_v) D_J \mathbf{e}\boldsymbol{\alpha}_J); \\ T_{d(w),J} &= \begin{pmatrix} (\Delta_{w,J})^{-1} R_a \Delta_{w,J} & (I - (\Delta_{w,J})^{-1} R_a \Delta_{w,J})\mathbf{e}\boldsymbol{\alpha}_J \\ 0 & T_J \end{pmatrix}. \end{aligned} \tag{5.30}$$

Furthermore, the waiting times and the sojourn times of individual types of customers have PH-distributions. Details are omitted.

In HE [18], an *M/G/*1 type Markov chain was introduced for studying the total workload process and the waiting times. Compared to that approach, the QBD process approach has some advantages in computation. For instance, the steady state distribution has a matrix geometric form and the computation of the matrix R can be done efficiently. On the other hand, the dimension of the matrices involved in the QBD approach can be as large as $m_a m_{\text{tot}}$, compared to m_a for the *M*/*G*/1 approach. Therefore, if there are many different types of customers or batches involved, the *M*/*G*/1 approach may be more efficient than the QBD approach.

6. Numerical analysis

In this section, we present a few numerical examples to demonstrate the implementability of the methods developed in this paper and to demonstrate the performances of different types of customers in a single server queueing system. First, we summarize the computational steps for the *GI*/*M*/1 case:

1. Compute transition blocks $\{A_0, A_1, A_2, \ldots\}$ by definition (3.7).
2. Compute $\boldsymbol{\theta}_{a,}$ arrival rates $\{\lambda_J, J \in \aleph\}$, service rates $\{\mu_J, J \in \aleph\}$, and ρ.
3. Compute $\boldsymbol{\beta}_k$, $\boldsymbol{\beta}_J$, and $\boldsymbol{\theta}_{\text{tot}}$ by (3.11).
4. Compute the matrix R by equation (3.23).
5. Compute the vector $\boldsymbol{\pi}(0)$ by Theorem 3.4.
6. Construct PH-distributions by Theorems 4.4 and 4.5.

Example 6.1. Consider the simplest model: the *Geo*/*Geo*/1 queue. The interarrival times have a common geometric distribution with $m_a = 1, D_0 = (0.8)$, and $D_1 = (0.2)$. The service times have a common geometric distribution with $m_s = 1$, $\boldsymbol{\alpha}_1 = (1)$, and $T_1 = (0.2)$. Then the waiting time (total workload) and the sojourn time have PH-distributions with matrix representations:

$$\begin{aligned}
&m_{w,1} = 1, \quad \boldsymbol{\alpha}_{w,1} = (1/16), \quad T_{w,1} = (1/4);\\
&m_{d(w),1} = 2, \quad \boldsymbol{\alpha}_{d(w),1} = (1/16, 15/16), \quad T_{d(w),1} = \begin{pmatrix} 1/4 & 3/4 \\ 0 & 1/5 \end{pmatrix}; \\
&m_{d,1} = 1, \quad \boldsymbol{\alpha}_{d,1} = (1), \quad T_{d,1} = (1/4).
\end{aligned} \tag{6.1}$$

respectively. Note that the matrix representation $\{m_{d(w),1}, \boldsymbol{\alpha}_{d(w),1}, T_{d(w),1}\}$ of the sojourn time obtained from $w_J + s_J$ can be reduced to $\{m_{d,1}, \boldsymbol{\alpha}_{d,1}, T_{d,1}\}$ since $\boldsymbol{\alpha}_{d(w),1} T_{d(w),1} = 0.25\boldsymbol{\alpha}_{d(w),1}$.

Example 6.2. Consider an *MMAP*[4]/*PH*[4]/1 queue with following system parameters: $K = 4$, $m_a = 3$, $m_1 = 1$, $\boldsymbol{\alpha}_1 = (1)$, $T_1 = (0.1)$;

$$m_2 = 4, \quad \boldsymbol{\alpha}_2 = (0.1, 0.1, 0.7, 0.1), \quad T_2 = \begin{pmatrix} 0 & 0.2 & 0 & 0 \\ 0.1 & 0.1 & 0 & 0 \\ 0 & 0.1 & 0.5 & 0.1 \\ 0.4 & 0 & 0 & 0 \end{pmatrix};$$

$$m_3 = 2, \quad \boldsymbol{\alpha}_3 = (0.1, 0.9), \quad T_3 = \begin{pmatrix} 0.5 & 0.1 \\ 0.6 & 0 \end{pmatrix};$$

$$m_4 = 2, \quad \boldsymbol{\alpha}_4 = (1, 0), \quad T_4 = \begin{pmatrix} 0 & 0.8 \\ 0.5 & 0.1 \end{pmatrix};$$

$$D_0 = \begin{pmatrix} 0.1 & 0.3 & 0 \\ 0.2 & 0.4 & 0.1 \\ 0.1 & 0 & 0.6 \end{pmatrix}, \quad D_1 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0.3 \end{pmatrix}, \quad D_2 = \begin{pmatrix} 0.1 & 0.1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix},$$

$$D_3 = \begin{pmatrix} 0.05 & 0.35 & 0 \\ 0.1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad D_4 = \begin{pmatrix} 0 & 0 & 0 \\ 0.1 & 0.1 & 0 \\ 0 & 0 & 0 \end{pmatrix}. \tag{6.2}$$

The traffic intensity of this queueing system is $\rho = 0.833$. The distributions of waiting times are given in Table 1.

It is interesting to see that type 1 batches have a waiting time that is significantly shorter than other types of batches. In fact, the sojourn times of type 1 batches are also significantly shorter than other types of batches. From the construction of the arrival process, it is clear that once the arrival process is in phase three, it stays there for a (relatively) long time. During that time, only type 1 batches arrive and their service times are short. On the other hand, once the arrival process is in phase 1 or 2, it stays there for a while with types 2, 3, and 4 batches. The service times of these types of batches are longer than that of type 1. Therefore, type 1 batches have a shorter waiting and sojourn time.

The matrix representations of the waiting times and sojourn times can be found for this example. But the dimension of the matrices is 27. Therefore, those matrix representations are not presented.

Table 1
Distributions of waiting times.

n	0	1	2	3	4	5	6	7	...
$P\{v=n\}$	0.1670	0.0912	0.0543	0.0449	0.0383	0.0359	0.0334	0.0313	...
$P\{w=n\}$	0.2432	0.0544	0.0456	0.0391	0.0367	0.0342	0.0320	0.0299	...
$P\{w_1=n\}$	0.4067	0.0521	0.0366	0.0324	0.0301	0.0281	0.0262	0.0246	...
$P\{w_2=n\}$	0.1833	0.0546	0.0484	0.0462	0.0423	0.0399	0.0371	0.0348	...
$P\{w_3=n\}$	0.1768	0.0551	0.0491	0.0466	0.0428	0.0402	0.0374	0.0351	...
$P\{w_4=n\}$	0.1595	0.0564	0.0508	0.0477	0.0439	0.0411	0.0382	0.0358	...

Example 6.3. Consider an *MMAP*[2]/*PH*[2]/1 queue with following system parameters: $K=2$, $m_a=2$, $N=3$,

$$D_0=\begin{pmatrix}0.7 & 0\\ 0.2 & 0.5\end{pmatrix},\quad D_1=\begin{pmatrix}0 & 0.2\\ 0 & 0\end{pmatrix},\quad D_2=\begin{pmatrix}0 & 0.1\\ 0 & 0.1\end{pmatrix},\quad D_{112}=\begin{pmatrix}0 & 0\\ 0 & 0.2\end{pmatrix};$$

$$m_1=2,\quad \boldsymbol{\alpha}_1=(0.9,0.1),\quad T_1=\begin{pmatrix}0.2 & 0.4\\ 0 & 0.1\end{pmatrix};$$

$$m_2=2,\quad \boldsymbol{\alpha}_2=(0.9,0.1),\quad T_2=\begin{pmatrix}0.1 & 0\\ 0.1 & 0.4\end{pmatrix}. \tag{6.3}$$

In this example, there are three types of batches: $J_1=1, J_2=2$, and $J_3=112$. The traffic intensity of this queueing system is $\rho=0.8162$. The distributions of waiting times $w_{(1)}$ and $w_{(2)}$ of type 1 and type 2 customers can be computed by using equation (4.17) and results are given in Table 2.

Table 2 shows that the waiting times of type 2 customers are significantly longer than that of type 1 customers. Next, we change the service order of in the batch $J_3=112$ to $J_3=121$. The distributions of waiting times of type 1 and type 2 customers are given in Table 3.

It is interesting to see that the waiting times of both types of customers are probabilistically shorter. Table 3 shows that the waiting times of type 2 customers are comparable to that of type 1 customers. Thus, a (single) change in the service order can change the queueing processes of different types of customers significantly. This demonstrates the necessity to conduct queueing analysis at the level of individual types of customers.

Table 2
Distributions of waiting times $w_{(1)}$ and $w_{(2)}$ when $J_3=112$.

n	0	1	2	3	4	5	6	7	...
$P\{w_{(1)}=n\}$	0.1640	0.6950	0.0739	0.0620	0.0559	0.0510	0.0467	0.0424	...
$P\{w_{(2)}=n\}$	0.1192	0.0276	0.0457	0.0693	0.0698	0.0603	0.0537	0.0491	...

Table 3
Distributions of waiting times $w_{(1)}$ and $w_{(2)}$ when $J_3=121$.

n	0	1	2	3	4	5	6	7	...
$P\{w_{(1)}=n\}$	0.2727	0.0828	0.1059	0.0919	0.0763	0.0627	0.0515	0.0425	...
$P\{w_{(2)}=n\}$	0.2070	0.1415	0.1168	0.0955	0.0773	0.0626	0.0510	0.0417	...

7. Discussion on the *SM*[*K*]/*PH*[*K*]/*c* case

The method used in this paper can be used to analyze the discrete time *SM*[*K*]/*PH*[*K*]/*c* queue, where customers are served by c servers, as was demonstrated in Van Houdt and Blondia [36,37]. Let $w_{n,k}$ be the workload of the k-th server after the arrival of the n-th batch, $1 \leq k \leq c$. Denote by $\mathbf{w}_n = (w_{n,1}, \ldots, w_{n,c})$. The construction of the corresponding Markov chain is based on the following basic relationship for the batch waiting times:

$$\mathrm{w}_{n+1} = O\big((w_{n,1} + s_{J_n} - \tau_{n+1})^+, (w_{n,2} - \tau_{n+1})^+, \ldots, (w_{n,c} - \tau_{n+1})^+\big), \quad n \geq 0, \tag{7.1}$$

where $x^+ = \max\{0, x\}$ and $O(\mathbf{y})$ is an operation to reorder the elements of the vector $\mathbf{y}$ in nondecreasing order. Then $w_{n,1}$ is the waiting time of the nth batch. Note that, in equation (7.1), we assumed that all customers in a batch are served by one server. If the service discipline is first-come-first-served (FCFS), we consider the age of the "youngest" batch in service in any period. If the service discipline is first-come-first-out (FCFO), we consider the age of the "youngest" batch or the "oldest" batch in service in any period. It is easy to see that the age process $a_g(t)$ constructed in such a way has the skip free to the right property. We choose variables for the arrival process and services in a way similar to that of Section 3 and put them into a vector $\mathbf{x}(t)$. Then the process $\{(a_g(t), \mathbf{x}(t)), t \geq 0\}$ is a *GI*/*M*/1 type Markov chain. This construction is similar to the construction in Asmussen and O'Cinneide [3] for the continuous time *GI*/*PH*/*c* queue.

The above discussion demonstrates that it is possible to analyze the waiting times and sojourn times of batches and customers for the discrete time *SM*[*K*]/*PH*[*K*]/*c* queue. However, the state space becomes extremely complicated, especially the organization of the level zero. Thus, it is not straightforward to obtain explicit results. We leave this problem to future research.

Acknowledgment

The research was supported by a NSERC research grant and a research grant from K.C. Wang Education Foundation and Chinese Academy of Sciences.

References

[1] S.A. Alfa, Discrete time queues and matrix-analytic methods, TOP 10(2) (2002) 147–210.

[2] S. Asmussen, Phase-type representation in random walk and queueing problems, Annals of Probability 20(2) (1992) 772–789.

[3] S. Asmussen and C. O'Cinneide, Representation for matrix-geometric and matrix-exponential steady-state distributions with applications to many-server queues, Stochastic Models 14 (1998) 369–387.

[4] S. Asmussen and G. Koole, Marked point processes as limits of Markovian arrival streams, J. Appl. Probab. 30 (1993) 365–372.

[5] D. Bini and B. Meini, On cyclic reduction applied to a class of Toeplitz-like matrices arising in queueing problems, in: *Computations with Markov Chains*, eds. W.J. Stewart (Kluwer Academic Publisher, 1996) pp. 21–38.

[6] E. Cinlar, Queues with semi-Markov arrivals, J. Appl. Prob. 4 (1967) 365–379.

[7] J.W. Cohen, *The Single Server Queue* (North-Holland Amsterdam, 1982).

[8] D.V. Cortizo, J. Garcia, C. Blondia and B. Van Houdt, FIFO by sets ALOHA (FS-ALOHA): A collision resolution algorithm for the contention channel in wireless ATM systems, Performance Evaluation 36–37 (1999) 401–427.

[9] J.H.A. De Smit, The single server semi-Markov queue, Stoch. Proc. And Appl. 22 (1986) 37–50.

[10] G. Fayolle, V.A. Malyshev and M.V. Menshikov, *Topics in the Constructive Theory of Countable Markov Chains* (Cambridge University Press, 1995).

[11] H.R. Gail, S.L. Hantler and B.A. Taylor, Solutions of the basic matrix equation for $M/G/1$ and $G/M/1$ type Markov chains, Stochastic Models 10 (1994) 1–43.

[12] H.R. Gail, S.L. Hantler and B.A. Taylor, Non-skip-free $M/G/1$ and $G/M/1$ type Markov chains, Adv. Appl. Probab. 29 (1997) 733–758.

[13] F.R. Gantmacher, *The Theory of Matrices* (Chelsea, New York, 1959).

[14] W.K. Grassmann and J.L. Jain, Numerical solutions of waiting time distribution and idle time distribution of the arithmetic *GI*/*G*/1 queue, Operations Research 37 (1989) 141–150.

[15] Qi-Ming HE, Queues with marked customers, Adv. Appl. Prob. 28 (1996) 567–587.

[16] Qi-Ming HE, Quasi-birth-and-death Markov processes with a tree structure and the *MMAP*[*K*]/ *PH*[*K*]/*N*/*LCFS* non-preemptive queue, European Journal of Operational Research 120(3) (2000) 641–656.

[17] Qi-Ming HE, The versatility of $MMAP$[*K*] and the *MMAP*[*K*]/ G[*K*]/1 queue, Queueing Systems 38(4) (2001) 397–418.

[18] Qi-Ming HE, Workload process, waiting times, and sojourn times in a discrete time *MMAP*[*K*]/ *SM*[*K*]/1/*FCFS* queue, Stochastic Models 20(4) (2004) 415–437.

[19] Qi-Ming HE, Age process, sojourn times, waiting times, and queue lengths in a continuous time *SM*[*K*]/*PH*[*K*]/1/FCFS queue (submitted for publication) (2003).

[20] Qi-Ming HE and A.S. Alfa, The *MMAP*[*K*]/*PH*[*K*]/1 queue with a last-come-first-served preemptive service discipline, Queueing Systems 28 (1998) 269–291.

[21] Qi-Ming HE and M.F. Neuts, Markov chains with marked transitions, Stochastic Processes and their Applications 74(1) (1998) 37–52.

[22] G. Latouche and V. Ramaswami, A logarithmic reduction algorithm for quasi-birth-and-death process, Journal of Applied Probability 30 (1993) 650–674.

[23] G. Latouche and V. Ramaswami, *Introduction to Matrix Analytic Methods in Stochastic Modelling* (ASA & SIAM, Philadelphia, USA, 1999).

[24] R.M. Loynes, The stability of a queue with non-independent interarrival and service times, Proc. Cambridge Philos. Soc. 58 (1962) 497–520.

[25] M.F. Neuts, Generalizations of the Pollaczek-Khinchin integral method in the theory of queues, Adv. Appl. Prob. 18 (1986) 952–990.

[26] M.F. Neuts, *Matrix-Geometric Solutions in Stochastic Models: An algorithmic Approach* (The Johns Hopkins University Press, Baltimore, 1981).

[27] M.F. Neuts, *Structured Stochastic Matrices of M/G/*1*type and Their Applications* (Marcel Dekker, New York, 1989).

[28] B. Sengupta, Markov processes whose steady state distribution is matrix-exponential with an application to the *GI*/*PH*/1 queue, Adv. Appl. Prob. 21 (1989) 159–180.

[29] B. Sengupta, Phase-type representations for matrix-geometric solutions, Stochastic Models 6 (1990) 163–167.

[30] B. Sengupta, The semi-Markovian queue: Theory and applications, Stochastic Models 6 (1990) 383–413.

[31] T. Takine, Queue length distribution in a FIFO single-server queue with multiple arrival streams having different service time distributions, Queueing System 39 (2001) 349–375.

[32] T. Takine, A recent progress in algorithmic analysis of FIFO queues with Markovian arrival streams, J. Korean Math. Soc. 38(4) (2001) 807–842.

[33] T. Takine and T. Hasegawa, The workload in a *MAP*/*G*/1 queue with state-dependent services: Its applications to a queue with preemptive resume priority, Stochastic Models 10(1) (1994) 183–204.

[34] B. Van Houdt and C. Blondia, Stability and performance of stack algorithms for random access communication modeled as a tree structured QBD Markov chain, Stochastic Models 17 (2001) 247–270.

[35] B. Van Houdt and C. Blondia, The delay distribution of a type k customer in a FCFS $MMAP[K]/PH[K]/1$ queue, Journal of Applied Probability 39(1) (2002) 213–222.

[36] B. Van Houdt and C. Blondia, The waiting time distribution of a type k customer in a FCFS $MMAP[K]/PH[K]/2$ queue (manuscript), (2002).

[37] B. Van Houdt and C. Blondia, The waiting time distribution of a type k customer in a discrete time $MMAP[K]/PH[K]/c$ ($c = 1, 2$) queue using QBDs, Stochastic models 20 (2004) 55–69.

[38] Y.Q. Zhao, W. Li and W.J. Braun, Censoring, factorization, and spectral analysis for transition matrices with block-repeating entries, Technical report (No. 355), Laboratory for Research in Statistics and Probability, Carleton University and University of Ottawa, (2001).

[39] T. Yang and M. Chaudhry, On the steady-state queue size distributions of discrete-time *GI*/*G*/1 queue, Adv. Appl. Probab. 28 (1996) 1177–1200.