J Syst Sci Complex (2012) 25: 133–155

ANALYSIS OF A CONTINUOUS TIME $SM[K]/PH[K]/1$/FCFS QUEUE: AGE PROCESS, SOJOURN TIMES, AND QUEUE LENGTHS*

Qiming HE

DOI: 10.1007/s11424-012-9138-0
Received: 18 May 2009 / Revised: 28 October 2010

Abstract This paper studies a continuous time queueing system with multiple types of customers and a first-come-first-served service discipline. Customers arrive according to a semi-Markov arrival process and the service times of individual types of customers have PH-distributions. A $GI/M/1$ type Markov process for a generalized age process of batches of customers is constructed. The stationary distribution of the $GI/M/1$ type Markov process is found explicitly and, consequently, the 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. The paper gives the matrix representations of the PH-distributions of waiting times and sojourn times. Some results are obtained for the distributions of queue lengths at departure epochs and at an arbitrary time. These results can be used to analyze not only the queue length, but also the composition of the queue. Computational methods are developed for calculating steady state distributions related to the queue lengths, sojourn times, and waiting times.

Key words $GI/M/1$ type Markov process, matrix analytic methods, queueing systems, queue length, semi-Markov chain, waiting times.

1 Introduction

Waiting times and sojourn times are important performance measurements for queueing systems. The study of waiting times and sojourn times was extensive and, in some cases, thorough for many queueing models. For instance, for the classical queueing models $M/M/1$, $M/G/1$, $MAP/G/1$, and $MMAP[K]/G[K]/1$, the distributions of waiting times were found in the form of (generalized) Pollaczek-Khintchine formula[1−13]. For the classical $GI/M/c$, $GI/PH/1$, $GI/PH/c$, and $SM/PH/1$ queue, it has shown that the waiting times and sojourn times have matrix exponential distributions[2,14−18]. The objective of this paper is to extend some of these results in [16–18] to a queueing model with multiple types of customers.

For the classical $GI/M/1$ queue, it is well known that the waiting time has a distribution similar to the exponential distribution[2]. That result was extended to the $GI/PH/1$ queue by Sengupta[16], who showed that the waiting times and sojourn times have matrix exponential distributions. The queue length distributions at departure epochs and at an arbitrary time

Qiming HE
Department of Management Sciences, University of Waterloo, Waterloo, Ontario, N2L 3G1, Canada.
Email: q7he@uwaterloo.ca.
*This research is supported by NSERC.
◇ *This paper was recommended for publication by Editor Shouyang WANG.*

were obtained as by-products as well. In [18], the above results were further generalized to a queueing model with a semi-Markov arrival process. In [15], Sengupta's results were extended to queues with a semi-Markov arrival process and multiple servers ($SM/PH/c$). In this paper, we consider a queueing system with multiple types of customers, which is an extension of the model defined in Examples 4 and 6 in [18]. The results obtained in [16, 18] are extended to the queueing model of interest. In addition to the extensions, some new issues arise and are addressed. For instance, the waiting time distributions are found not only for an arbitrary customer, but also for individual types of customers. A generalized age process that is more general than the Markov processes used in previous papers is constructed and analyzed.

The study of queueing systems with multiple types of customers and an FCFS service discipline can be summarized as follows. For the continuous time $MMAP[K]/G[K]/1$ queue for which the arrival process is Markovian and the service times have general distribution, Pollaczek-Khintchine type results were found for the waiting time distributions and algorithms were developed for computing these distributions[3−4,11−12]. The queue length distributions are extremely difficult to obtain because there are different types of customers with different service time distributions in the system. By utilizing the distributions of waiting times, Takine[11] found the queue length distributions for such queueing models. For the discrete time $MMAP[K]/G[K]/1$ queue, He[19] used the generalized total workload process to obtain distributions of the waiting times. For the discrete time $SM[K]/PH[K]/1$ queue, He[20] and Van Houdt and Blondia[21] utilized the (generalized) age process to obtain the distributions of waiting times and sojourn times. This paper can be considered as an extension of [20–22] from the discrete time case to the continuous time case. As mentioned before, this paper can also be considered as an extension of [16, 18] from the one type customer case to the multiple type customer case. In this paper, a number of results are obtained for the distributions of waiting times, sojourn times, and queue lengths for individual types of customers. Algorithms are developed for computing these distributions and related performance measures. For some special cases, the algorithms are involved with only matrix operations and can be implemented in a straightforward manner. The results about the waiting time, sojourn times, and queue strings can be used to analyze not only the queue length, but also the relationship between different types of customers in the queue as well as the composition of the queue (see examples in Section 6). Applications of some of the algorithms developed in the paper can be found in [23].

In summary, on one hand, this paper extends the main results in [20–22] from the discrete time case to the continuous time case. This paper also studies the queue length processes that were not investigated in [20–22]. On the other hand, this paper extends some of the results in [16, 18] from queueing systems with a single type of customers to queueing systems with multiple types of customers. Compared to [16, 18], this paper studies a queueing model with a fairly general formalism for the arrival process with multiple types of customers. This paper demonstrates that a queueing system with that arrival process is still tractable. In addition, this paper introduces a generalized age process that plays a central role in the study of queueing performance measures. An open problem in [16] is resolved in this paper as well.

The rest of the paper is organized as follows. In Section 2, the queueing model of interest is introduced. In Section 3, a generalized age process is introduced and its stationary distribution is obtained. Section 4 utilizes the results obtained in Section 3 to find distributions of waiting times and sojourn times. In Section 5, some results are obtained for the queue length distributions. In Sections 2 to 5, we use the $MMAP[K]/PH[K]/\ 1$ queue as an example to show that more detailed results can be obtained. Finally, in Section 6, three numerical examples are analyzed to show the usefulness of the theory developed in this paper and to gain insight into the queueing model of interest.

2 The Continuous Time $SM[K]/PH[K]/1$/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 give a detailed description of the customer arrival process and service process. A discrete time version of this model was introduced in [20].

The customer arrival process The customer arrival process is a continuous time semi-Markov process with marked transitions[1,4,24–26]. Customers are distinguished into K types and arrive in batches. To characterize the batches of customers, we define a set of strings of integers:

$$\aleph = \{J_n : J_n = j_1 j_2 \cdots j_{|J_n|},\ 1 \le j_i \le K,\ 1 \le i \le |J_n|,\ 1 \le n \le N\}, \tag{1}$$

where N is the total number of different strings in set $\aleph$ and $|J|$ is the number of integers in the string J, which is called the length of J. For the queueing system, a string $J = j_1 j_2 \cdots j_n \in \aleph$ represents a batch that has n customers. These n customers are of types j_1, j_2, $\cdots$, and j_n, respectively. For instance, if $K = 6$ and $J = 35525$, i.e., j_1=3, j_2=5, j_3=5, j_4=2, and j_5=5, then there are five customers in the batch J and the types of these customers are 3, 5, 5, 2, and 5, respectively. We call J a string representation of that batch. Thus, there are in total N different types of batches.

Consider a continuous time semi-Markov chain $\{(\xi_n,\ \tau_n),\ n \ge 0\}$ with m_a phases. The variable ξ_n is the phase of the semi-Markov chain right after the n-th transition. The variable τ_n is the time between the $(n–1)$-th 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. Define, for $t \ge 0$,

$$P\{\xi_n = j,\ \tau_n \le t,\ J_n = J | \xi_{n-1} = i\} = d_{J,i,j}(t),\quad 1 \le i,\quad j \le m_a,\quad n \ge 1,\quad J \in \aleph. \tag{2}$$

The variable $d_{J,i,j}(t)$ is the probability that a batch J arrives before time t after the arrival of the last batch at time zero and the phase of the underlying semi-Markov process becomes j right after the transition, given that the phase was i at time 0. We assume that $d_{J,i,j}(0)$=0. Let $D_{a,J}(t)$ be an $m_a \times m_a$ matrix with $(i,\ j)$-th element $d_{J,i,j}(t)$. Matrices $\{D_{a,J}(t),\ t \ge 0,\ J \in \aleph\}$ provide all information about the semi-Markov arrival process with marked transitions. Define

$$D_a(t) = \sum_{J \in \aleph} D_{a,J}(t);\quad D_{a,J} = \lim_{t \to \infty} D_{a,J}(t),\quad J \in \aleph;\quad D_a = \sum_{J \in \aleph} D_{a,J} = \lim_{t \to \infty} D_a(t). \tag{3}$$

The matrix D_a is the probability transition matrix of the embedded Markov chain at transition epochs of the semi-Markov process $\{(\xi_n,\ \tau_n),\ n \ge 0\}$. We assume that D_a is irreducible. Let θ_a be the invariant probability vector of the stochastic matrix D_a, i.e., $\theta_a D_a = \theta_a$ and $\theta_a \boldsymbol{e} = 1$, where $\boldsymbol{e}$ 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:

$$E_{\theta_a}[\tau] = \theta_a \int_0^\infty t D_a(dt) \boldsymbol{e}. \tag{4}$$

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

$$\lambda_{(k)} = \sum_{J\in\aleph} N(J,k)\lambda_J, \quad 1 \le k \le K, \tag{5}$$

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.

The Service Process All batches are served by a single server on an FCFS basis. After a batch of customers arrived, all customers in that batch join the queue according to the order in the batch. In most part of this paper, the services of the customers in a batch are considered together. The service times of individual customers have continuous time PH-distributions and are independent of each other and the arrival process. For a type k customer, its service time s_k has a PH-distribution with a matrix representation $\{m_k,\ \alpha_k,\ T_k\}$, where m_k is the number of phases of the PH-distribution, α_k is a probability vector, and T_k is a subgenerator (i.e., the diagonal elements of T_k are negative, all other elements are nonnegative, and $T_k\boldsymbol{e} \le 0$). We assume that the service time of any customer is positive, i.e., $\alpha_k\boldsymbol{e} = 1$, $1\le k\le K$. See [7] for more about PH-distribution. Denote by $\boldsymbol{T}_k{}^0 = -T_k\ \boldsymbol{e}$. We assume that each PH-distribution is irreducible, i.e., $T_k + \boldsymbol{T}_k{}^0\alpha_k$ is irreducible. The service time of a batch is the sum of the service times 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 continuous time PH-distribution with a matrix representation $\{m_J,\ \alpha_J,\ T_J\}$, where, for $J = j_1 j_2 \cdots j_n \in \aleph$,

$$m_J = \sum_{i=1}^{n} m_{j_i}; \qquad \alpha_J = (\alpha_{j_1},\ 0,\ \cdots,\ 0);$$

$$T_J = \begin{pmatrix} T_{j_1} & \boldsymbol{T}^0_{j_1}\alpha_{j_2} & & & \\ & T_{j_2} & \boldsymbol{T}^0_{j_2}\alpha_{j_3} & & \\ & & \ddots & \ddots & \\ & & & T_{j_{n-1}} & \boldsymbol{T}^0_{j_{n-1}}\alpha_{j_n} \\ & & & & T_{j_n} \end{pmatrix}, \qquad \boldsymbol{T}^0_J = \begin{pmatrix} 0 \\ \vdots \\ \vdots \\ 0 \\ \boldsymbol{T}^0_{j_n} \end{pmatrix}. \tag{6}$$

The mean service time of a type k customer is given by $E[s_k] = -\alpha_k T_k{}^{-1}\boldsymbol{e}$. The mean service time of a type J batch is given by $E[s_J] = \sum_{i=1}^{|J|} E[s_{j_i}]$. The service rate of a batch J is defined as $\mu_J = (E[s_J])^{-1}$.

The service rate of type k customers is $\mu_k = (E[s_k])^{-1}$. The traffic intensity of the queueing system can be defined in terms of batch arrival rates and batch service rates (or customer arrival rates and service rates):

$$\rho = \sum_{J\in\aleph} \frac{\lambda_J}{\mu_J} = \sum_{k=1}^{K} \frac{\lambda_{(k)}}{\mu_k}. \tag{7}$$

By Loynes[27], 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 queueing

systems, such as the $MMAP[K]/PH[K]/1$ queue, the $GI/PH/1$ queue, and the $GI/M/1$ queue, are special cases of the $SM[K]/PH[K]/1$ queue.

Example 2.1 The Continuous Time $MMAP[K]/PH[K]/1$ Queue This queueing system has a batch Markovian arrival process with matrix representation $\{D_0,\ D_J,\ J\in\aleph\}$, where D_0 is an $m_a\times\ m_a$ subgenerator, $\{D_J,\ J\in\aleph\}$ are $m_a\ \times m_a$ nonnegative matrices. The matrix D_J is the (matrix) arrival rate of type J batches. For more about $MMAP[K]$, see [3–4, 24, 26]. The relationship between the two sets of parameters of the arrival process is: $D_{a,J}(dt) = \exp\{D_0t\}dtD_J,\ J \in \aleph,\ t \geq 0$. Let $D = D_0 + \Sigma_J D_J$ be the infinitestimal generator of the underlying Markov process of the arrival process. We assume that D is irreducible and $D\neq\ D_0$. Denote by θ the invariant probability vector of the stochastic matrix D. It is easy to see $\theta_a = -\theta D_0/\ \lambda$, where $\lambda = -\theta D_0\boldsymbol{e} = \theta\Sigma_J D_J\boldsymbol{e}$. In addition, we have $\lambda_J = \theta D_J\boldsymbol{e}$. As shall be shown, more detailed results can be obtained for this special case.

3 Analysis of the Generalized Age Process

The following analysis of the generalized age process is parallel to that of the discrete time case studied in [20]. Thus, some details will be omitted and some proofs are given in the appendix.

3.1 The Generalized Age Process

The basic idea to analyze the sojourn times 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

$$w_{n+1} = \max\{0,\ w_n + s_{J_n} - \tau_{n+1}\},\quad n \geq 0, \tag{8}$$

where τ_{n+1} is the length of the time between the n-th batch and the $(n+1)$-th batch, J_n is the type of the n-th batch, and $s_{J_n}$is the service time of the n-th batch. We define the age of a batch at time t as the total time the batch has been in the queueing system, given that the batch is in the system at time t. The generalized age process $\{a_g(t),\ t\geq0\}$ of the batch in service or to be served next (if the system is empty) is defined as

$$a_g(t) = w_{n(t)} + s_{J_{n(t)}} - \tau_{n(t)+1} + t - \eta_{n(t)}, \tag{9}$$

where $n(t)$ is the ordinal number of the last batch served before time t and $\eta_{n(t)}$ is the departure time of the $n(t)$-th batch. Detailed discussion on the process $\{a_g(t),\ t \geq0\}$ can be found in [20] for the discrete time case. In general, 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 current idle period. It is easy to see that $a_g(t)$ satisfies the following equation, for small δt,

$$a_g(t+\delta t) = \begin{cases} a_g(t) + \delta t, & \text{if no service is completed in } (t,\ t+\delta t), \\ a_g(t) - \tau_{n(t+\delta t)+1} + \delta t, & \text{if a service is completed in } (t,\ t+\delta t). \end{cases} \tag{10}$$

In order to construct a Markov chain, we introduce some auxiliary variables related to the phase of the arrival and service processes. We define a process $\{I_a(t),\ t\geq0\}$ from the Markov chain $\{\xi_n,\ n\geq0\}$ defined in Section 2 as: $I_a(t) = \xi_n$ if the n-th batch is the last batch departed at or before 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 at time t (if any) and $J(t)$ be the type of the batch in service at time t (if any). If there is no service at t, $J(t)$ is the type of the next batch to be served and $I_s(t)$

the initial service phase of the batch to be served. Putting these variables together, we obtain a Markov process $\{(a_g(t),\ I_a(t),\ J(t),\ I_s(t)),\ t\geq 0\}$. We call $a_g(t)$ the level variable and $\{I_a(t),\ J(t),\ I_s(t)\}$ auxiliary variables that take a finite number of values in the set:

$$\{(i, J_n, j):\quad 1\leq i\leq m_a,\ \ 1\leq j\leq m_{J_n},\ 1\leq n\leq N\}, \tag{11}$$

in which the states are ordered lexicographically. Equation (10) 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 process of $GI/M/1$ type with no boundary at the level 0. Denote by

$$\begin{aligned}
&m_{tot}=\sum_{n=1}^{N} m_{J_n},\\
&\alpha(J_n)=(0,\cdots,0,\ \alpha_{J_n},0,\cdots,0),\quad 1\leq n\leq N,\\
&T_{tot}=\begin{pmatrix} T_{J_1} & & & \\ & T_{J_2} & & \\ & & \ddots & \\ & & & T_{J_N}\end{pmatrix},\qquad \boldsymbol{T}^0_{tot}=-T_{tot}\boldsymbol{e}=\begin{pmatrix}\boldsymbol{T}^0_{J_1}\\ \boldsymbol{T}^0_{J_2}\\ \vdots\\ \boldsymbol{T}^0_{J_N}\end{pmatrix},
\end{aligned} \tag{12}$$

where $\alpha(J_n)$ is a row vector of the size m_{tot}, $1\leq n\leq N$, and T_{tot} is an $m_{tot}\times m_{tot}$ matrix. The vector $\alpha(J_n)$ is obtained by putting the vector α_{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}.

Similar to [16], the transition of the Markov chain $\{(a_g(t),\ I_a(t),\ J(t),\ I_s(t)),\ t\geq 0\}$ is specified as follows:

1) The process $a_g(t)$ increases linearly at rate one, except when downward jumps occur. So $a_g(t)$ is skip free to the right.

2) The downward jumps in $a_g(t)$ correspond to service completions. If a batch completes its service at time t, the magnitude of the downward jump of $a_g(t)$ is $\tau_{n(t-)+1}$.

The transition probabilities are given as follows: If $a_g(t)=x\ \geq 0$, for $u>0$,

$$\begin{aligned}
&P\{a_g(t+\delta t)=x+\delta t,\ I_a(t+\delta t)=i',\ J(t+\delta t)=J',\ I_s(t+\delta t)=j'\\
&\qquad\qquad |a_g(t)=x,\ I_a(t)=i,\ J(t)=J,\ I_s(t)=j\}\\
&=\begin{cases} 1-(\mathrm{diag}(T_{tot}))_{(i,J,j),(i,J,j)}\,\delta t+o(\delta t), & \text{if}\quad (i,J,j)=(i',J',j');\\ (I\otimes(T_{tot}-\mathrm{diag}(T_{tot})))_{(i,J,j),(i',J',j')}\,\delta t+o(\delta t), & \text{otherwise;}\end{cases}
\end{aligned} \tag{13}$$

$$\begin{aligned}
&P\{x-u\leq a_g(t+\delta t)<x,\ I_a(t+\delta t)=i',\ J(t+\delta t)=J',\ I_s(t+\delta t)=j'\\
&\qquad\qquad |a_g(t)=x,\ I_a(t)=i,\ J(t)=J,\ I_s(t)=j\}\\
&=\left(\sum_{n=1}^{N} D_{a,J_n}(u)\otimes(\boldsymbol{T}^0_{tot}\alpha(J_n))\right)_{(i,J,j),(i',J',j')}\delta t+o(\delta t),
\end{aligned} \tag{14}$$

where diag (T_{tot}) is a matrix whose diagonal elements are the diagonal elements of the matrix T_{tot} and all other elements are zero, the notation "$\otimes$" is for Kronecker product of matrix (see [28–29]). If $a_g(t)=x<0$, for $x+\delta t<0$,

$$\begin{aligned}
&P\{a_g(t+\delta t)=x+\delta t,\ I_a(t+\delta t)=i',\ J(t+\delta t)=J',\ I_s(t+\delta t)=j'\\
&\qquad |a_g(t)=x,\ I_a(t)=i,\ J(t)=J,\ I_s(t)=j\}\\
&=\begin{cases} 1, & \text{if}\ \ i=i',\ J=J',\ j=j';\\ 0, & \text{otherwise.}\end{cases}
\end{aligned} \tag{15}$$

3.2 Ergodicity of the Generalized Age Process

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

$$\begin{aligned} A(t) &= I \otimes T_{tot} + \sum_{i=1}^{N} D_{a,J_i}(t) \otimes (\boldsymbol{T}_{tot}^0 \alpha(J_i)), \quad t \geq 0; \\ A^*(s) &= \int_0^\infty \exp\{-st\} A(dt); \\ A &= \lim_{t\to\infty} A(t) = \lim_{s\to 0} A^*(s) = I \otimes T_{tot} + \sum_{n=1}^{N} D_{a,J_n} \otimes (\boldsymbol{T}_{tot}^0 \alpha(J_n)). \end{aligned} \tag{16}$$

The off-diagonal elements of $A(t)$ can be interpreted as the rates of transitions. It is easy to see that A is a generator, i.e., $A\boldsymbol{e} = 0$. We assume that A is irreducible. Denote by $\chi(s)$ the Perron-Frobenius eigenvalue of the matrix $A^*(s)$ (i.e., the eigenvalue with the largest modulus).

Denote by β_k the invariant probability vector of $T_k + \boldsymbol{T}_k{}^0\alpha_k$, i.e., $\beta_k(T_k + \boldsymbol{T}_k{}^0\alpha_k) = 0$, $1{\leq}k{\leq}K$. Since $T_k + \boldsymbol{T}_k{}^0\alpha_k$ is irreducible, every element of the vector β_k is positive[28–29]. By Neuts[7], $\beta_k \boldsymbol{T}_k{}^0 = \mu_k$. Denote by β_J the invariant probability vector of $T_J + \boldsymbol{T}_J{}^0\alpha_J$, for $J{\in}\aleph$. By $\beta_k T_k + \mu_k\alpha_k = 0$, it can be verified, for $J{\in}\aleph$,

$$\beta_J = \mu_J \left(\frac{\beta_{j_1}}{\mu_{j_1}},\, \frac{\beta_{j_2}}{\mu_{j_2}},\, \cdots,\, \frac{\beta_{j_{|J|}}}{\mu_{j_{|J|}}} \right), \quad \text{where } \mu_J = \left(\frac{1}{\mu_{j_1}} + \frac{1}{\mu_{j_2}} + \cdots + \frac{1}{\mu_{j_{|J|}}} \right)^{-1}. \tag{17}$$

Recall that $|J|$ is the number of integers in the string J. Similar to the definition of $\alpha(J)$ in Equation (12), we define $\beta(J)$ as $\beta(J) = (0, \cdots, 0, \beta_J, 0, \cdots, 0)$, which is a row vector of the size m_{tot}. Denote by

$$\theta_{tot} = \frac{\lambda}{\rho} \sum_{n=1}^{N} (\theta_a D_{a,J_n}) \otimes \left(\frac{\beta(J_n)}{\mu_{J_n}} \right). \tag{18}$$

The following lemmas are useful in the proof of Theorem 3.3 and in the next few sections. Their proofs are given in the appendix.

Lemma 3.1 *Assume that A is irreducible. The vector θ_{tot} is the unique invariant probability vector of A.*

Lemma 3.2 *Assume that the matrix A is irreducible. At $s = 0$, we have $\chi(0) = 0$ and $\theta_{tot}A^{*(1)}(0)\boldsymbol{e} = \chi^{(1)}(0) = 1/\rho$. Consequently, $\chi^{(1)}(0) > 1$ if and only if $\rho < 1$ (Note that $\chi^{(1)}(0)$ is the first derivative of the function $\chi(s)$ at $s{=}0$).*

Theorem 3.3 *Assume that the Markov process $\{(a_g(t),\, I_a(t),\, J(t),\, I_s(t)),\, t \geq 0\}$ and the matrix A are irreducible. The Markov chain is positive recurrent if and only if $\rho < 1$.*

Proof By Theorem 2.3 in [16], the Markov process is positive recurrent if and only if $\theta_{tot}A^{*(1)}(0)\boldsymbol{e} > 1$. In [16], the level variable of the age process takes nonnegative values. In our case, $a_g(t)$ can be negative. Since there is no service if $a_g(t) < 0$, the process will definitely reach the level zero from any negative level in finite time. Therefore, the result in [16] can be applied to the Markov process of interest. By Lemma 3.2, $\theta_{tot}A^{*(1)}(0)\boldsymbol{e} > 1$ is equivalent to $\rho < 1$, which leads to the desired result. This completes the proof of Theorem 3.3. ∎

3.3 The Steady State Distribution of the Generalized Age Process

Now, we assume $\rho<1$ so that the Markov process $\{(a_g(t),\, I_a(t),\, J(t),\, I_s(t)),\, t\geq 0\}$ is positive recurrent. Denote by $\pi(x)$ the density function of the steady state distribution of $\{(a_g(t),\, I_a(t),\, J(t),\, I_s(t)),\, t\geq 0\}$, where $\pi(x) = (\cdots,\, \pi_{i,J,j}(x),\, \cdots)$ is a row vector of the size $m_a m_{tot}$ and

$$\pi_{i,J,j}(x)dx = \lim_{t\to\infty} \frac{1}{t} \int_0^t P\{x < a_g(u) < x + dx,\ I_a(u) = i,\ J(u) = J,\ I_s(u) = j\ \}du. \tag{19}$$

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 $(a_g(t)\geq 0)$, by Lemma 2.4 in [16], its steady state distribution has a matrix exponential solution:

$$\pi(x) = \pi(0)\exp\{Tx\}, \quad x \geq 0, \tag{20}$$

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

$$T = I \otimes T_{tot} + \sum_{J\in\aleph} \int_0^\infty \exp\{Tx\} \left(D_{a,J}(dx) \otimes \boldsymbol{T}_{tot}^0 \alpha(J)\right). \tag{21}$$

We refer to [16] for more about Equation (21) and the matrix T. A stable iterative algorithm was developed in [16] for computing the matrix T as well. Using the above equations, explicit expressions can be found for $\pi(0)$ and $\pi(x)$ for $x < 0$.

Theorem 3.4 *Assume $\rho< 1$. We have*

$$\pi(x) = \rho\theta_{tot} T \int_0^\infty \exp\{Tt\} A(t-x)dt, \quad x < 0; \qquad \pi(0) = -\rho\theta_{tot} T; \tag{22}$$

$$\int_{-\infty}^0 \pi(x)dx = \rho\theta_{tot} \left(\int_0^\infty xA(dx) + T^{-1}A - I\right); \qquad \int_{-\infty}^0 \pi(x)dx\boldsymbol{e} = 1 - \rho.$$

Proof When $a_g(t) < 0$, $a_g(t)$ is increasing linearly in t at the rate one. Thus, $\pi(x)$ is increasing in x if $x < 0$. The difference $\pi(x)$– $\pi(x–\delta)$ comes from the downward jumps of $a_g(t)$. Therefore, we have, for $x<0$,

$$\pi^{(1)}(x) = \int_0^\infty \pi(t)A(dt - x) = \pi(0)\int_0^\infty \exp\{Tt\}A(dt - x), \tag{23}$$

where $\pi^{(1)}(x)$ is for the derivative of $\pi(x)$ at x. Equation (23) leads to

$$\begin{aligned}
\pi(0) - \pi(x) &= \int_x^0 \pi^{(1)}(u)du \\
&= \pi(0)\int_x^0 \int_0^\infty \exp\{Tt\}A(dt - u)du \\
&= \pi(0)\int_0^\infty \exp\{Tt\}\int_x^0 (-A(t - du))dt \\
&= \pi(0)\int_0^\infty \exp\{Tt\}\left(A(t-x) - A(t)\right)dt \\
&= \pi(0)\int_0^\infty \exp\{Tt\}A(t-x)dt - \pi(0)\int_0^\infty \exp\{Tt\}A(t)dt.
\end{aligned} \tag{24}$$

Since

$$\begin{aligned}\int_0^\infty \exp\{Tt\}A(t)dt &= -T^{-1}\int_0^\infty (d\exp\{Tt\})\,A(t)\\ &= T^{-1}\left(-A(0)-\int_0^\infty \exp\{Tt\}A(dt)\right)\\ &= -T^{-1}\left(I\otimes T_{tot}+\int_0^\infty \exp\{Tt\}A(dt)\right)\\ &= -T^{-1}T = -I. \end{aligned} \tag{25}$$

Equation (24) leads to $\pi(x) = -\pi(0)\int_0^\infty \exp\{Tt\}A(t-x)dt$. Letting $x \to -\infty$ in that expression, we obtain $\pi(0)T^{-1}A = 0$, since $\pi(x) \to 0$ as $x \to -\infty$. Therefore, $\pi(0) = c\theta_{tot}T$. To determine the constant c, we evaluate the following integration of $\pi(x)$:

$$\begin{aligned}\int_{-\infty}^0 \pi(x)dx &= -\int_{-\infty}^0 \pi(0)\int_0^\infty \exp\{Tt\}A(t-x)dtdx\\ &= -\pi(0)T^{-1}\int_{-\infty}^0\int_0^\infty (d\exp\{Tt\})\,A(t-x)dx\\ &= -\pi(0)T^{-1}\int_{-\infty}^0\left(-A(-x)-\int_0^\infty \exp\{Tt\}A(dt-x)\right)dx\\ &= \pi(0)T^{-1}\left(\int_0^\infty A(x)dx+\int_0^\infty \exp\{Tt\}(A-A(t))dt\right)\\ &= \pi(0)T^{-1}\left(\int_0^\infty A(x)dx - T^{-1}A + I\right), \end{aligned} \tag{26}$$

where Equation (25) is used in the last equality. Since $\int_{-\infty}^\infty \pi(x)dx = 1$, Equation (25) leads to

$$\begin{aligned}1 &= \int_{-\infty}^\infty \pi(x)dx\boldsymbol{e}\\ &= \left(\pi(0)T^{-1}\left(\int_0^\infty A(x)dx - T^{-1}A + I\right)+\int_0^\infty \pi(x)dx\right)\boldsymbol{e}\\ &= \left(\pi(0)T^{-1}\left(\int_0^\infty A(x)dx - T^{-1}A + I\right)-\pi(0)T^{-1}\right)\boldsymbol{e}\\ &= c\theta_{tot}\left(\int_0^\infty A(x)dx\boldsymbol{e} - T^{-1}A\boldsymbol{e}\right)\\ &= -c\theta_{tot}\int_0^\infty xA(dx)\boldsymbol{e} = -c/\rho, \end{aligned} \tag{27}$$

where we used Lemma 3.2 and $A\boldsymbol{e}= 0$ for the last equality. By Equation (27), $c = -\rho$. The rest of the results follow easily. This completes the proof of Theorem 3.4. ∎

Theorem 3.4 shows that the system is busy with probability ρ, which is consistent with intuition. Summarizing the results in Equation (20) and Theorem 3.4, we obtain the following expressions for the density function of the stationary distribution of the Markov process $\{(a_g(t), I_a(t), J(t), I_s(t)), t\geq 0\}$:

$$\pi(x) = \begin{cases} \rho\theta_{tot}T\displaystyle\int_0^\infty \exp\{Tt\}A(t-x)dt, & x < 0,\\ -\rho\theta_{tot}T\exp\{Tx\}, & x \geq 0. \end{cases} \tag{28}$$

Example 3.1 The Continuous Time *MMAP*[*K*]/*PH*[*K*]/1 Queue (Example 2.1 continued) For this case, Equation (21) for the matrix T can be simplified to

$$\begin{aligned} T &= I \otimes T_{tot} + \int_0^\infty \exp\{Tt\} \left(\exp\{D_0 t\} \otimes I\right) dt \left(\sum_{J \in \aleph} D_J \otimes \boldsymbol{T}_{tot}^0 \alpha(J)\right) \\ &= I \otimes T_{tot} + L \left(\sum_{J \in \aleph} D_J \otimes \boldsymbol{T}_{tot}^0 \alpha(J)\right), \end{aligned} \tag{29}$$

where

$$\begin{aligned} L \;=\; & -T^{-1} - T^{-1} \int_0^\infty \exp\{Tt\} \left(\exp\{D_0 t\} \otimes I\right) dt \left(D_0 \otimes I\right) \\ \Rightarrow\; & TL + L\left(D_0 \otimes I\right) = -I. \end{aligned} \tag{30}$$

Denote by $\phi(L)$ the direct-sum of the matrix L, i.e., putting the rows of the matrix L (from top to bottom) into a single row vector (see page 19 in [7]). Then the last equality in Equation (30) is equivalent to

$$\phi(L) = -\phi(I) \left(T' \otimes I + I \otimes D_0 \otimes I\right)^{-1}, \tag{31}$$

where T' is the transpose of the matrix T and $\phi\left(I\right)$ is the direct-sum of the matrix I. Therefore, for computing the matrix T by the iteration method given in [16], each iteration is reduced to solve a linear equation. This property makes the computation of T easier, since integration is avoided.

We would like to point out that Equation (30) is a (generalized) Sylvester matrix equation, it can be solved more efficiently by using a Hessenberg/Schur algorithm (see [30]). In addition, the vector θ_{tot} can be simplified to

$$\theta_{tot} = \frac{\theta}{\rho} \sum_{J \in \aleph} \left(D_J\right) \otimes \left(\frac{\beta(J)}{\mu_J}\right). \tag{32}$$

4 Age, Total Workload, Sojourn Times, and Waiting Times

4.1 Distributions of Age and Total Workload

Let a_g be the generic random variable of the generalized age at an arbitrary epoch. By Equation (28), the distribution of a_g is obtained easily as

$$P\{a_g < x\} = \begin{cases} \rho\theta_{tot} \left(\displaystyle\int_{-x}^{\infty} A(u)du - \int_0^\infty \exp\{Tt\} A(t-x)dt\right) \boldsymbol{e}, & x < 0, \\ 1 - \rho\theta_{tot} \exp\{Tx\}\boldsymbol{e}, & x \geq 0. \end{cases} \tag{33}$$

The total workload (virtual waiting time) is defined as the total service time of all waiting batches plus the remaining service time of the batch in service (if any). Based on Equation (8), we introduce the generalized total workload process:

$$v_g(t) = w_{n(t)} + s_{J_{n(t)}} - (t - \xi_{n(t)}), \tag{34}$$

where $n(t)$ be the ordinal number of the last batch arrived at or before time t, and $\xi_{n(t)}$ is the arrival time of the $n(t)$-th batch. Then the total workload is given by $\max\{0,\, v_g(t)\}$. It is

well-known in the literature that $\max\{0, v_g(t)\}$ and $\max\{0, a_g(t)\}$ have the same distribution in steady state. For more details about the process $v_g(t)$, see [19].

The relationship between the generalized age process and the generalized total workload process is shown in the following lemma. The following lemma was proved for discrete time queueing models in [20] (Lemma 4.1 in [20]), which is also valid for continuous time cases.

Lemma 4.1 *In a busy cycle (starting with the first service in a busy period and ending at the beginning of the next busy period), the number of times that $a_g(t)$ up-crosses x equals the number of times that $v_g(t)$ down-crosses x, for any real number x. Consequently, in steady state, $v_g(t)$ and $a_g(t)$ have the same distribution.*

4.2 Distributions of Sojourn Times

We define the sojourn time of an arbitrary batch (an arbitrary type J batch) as the time between its arrival and its service completion (of all customers in the batch). 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 an arbitrary type J batch in steady state. Let $d_{(k)}$ be the generic random variable for the sojourn time of an arbitrary type k customer in steady state, which is the time between its arrival (i.e., the arrival of its btach) and its service completion (not the service completion of its batch).

We decompose $\boldsymbol{T}^0{}_{tot}$ into $\{\boldsymbol{T}^0{}_{tot,J}\colon\ J\in\aleph\}$, where $\boldsymbol{T}^0{}_{tot,J}$ is obtained by setting all $\boldsymbol{T}^0{}_H$ in the vector $\boldsymbol{T}^0{}_{tot}$ to a zero vector if $H \neq J$. Apparently, we have $\boldsymbol{T}^0_{tot} = \sum\limits_{J\in\aleph} \boldsymbol{T}^0_{tot,J}$. For $1\leq k\leq K$, we construct column vector $\boldsymbol{T}^0{}_{tot,(k)}$ from $\boldsymbol{T}^0{}_{tot}$ by setting all $\boldsymbol{T}^0{}_j$ in $\boldsymbol{T}^0{}_{tot}$ to a zero vector if $j \neq k$. We also have $\boldsymbol{T}^0_{tot} = \sum_{k=1}^K \boldsymbol{T}^0_{tot,(k)}$.

Lemma 4.2 *$\theta_{tot}(\boldsymbol{e}\otimes\boldsymbol{T}^0{}_{tot}) = \lambda/\rho$, $\theta_{tot}(\boldsymbol{e}\otimes\boldsymbol{T}^0{}_{tot,J}) = \lambda_J/\rho$, $J\in\aleph$, and $\theta_{tot}(\boldsymbol{e}\otimes\boldsymbol{T}^0{}_{tot,(k)}) = \lambda_{(k)}/\rho$, $1\leq k\leq K$.*

Consider the sojourn time of an arbitrary type J batch. Note that the sojourn time of a batch equals the age of that batch at its service completion epoch. Conditioning on the service completion of a type J batch, we obtain, for $x \geq 0$,

$$\begin{aligned} P\{d_J \leq x\} &= \frac{1}{\lambda_J}\int_0^x \pi(t)\left(\boldsymbol{e}\otimes\boldsymbol{T}^0_{tot,J}\right)dt \\ &= -\frac{\rho}{\lambda_J}\theta_{tot}T\int_0^x \exp\{Tt\}dt\left(\boldsymbol{e}\otimes\boldsymbol{T}^0_{tot,J}\right) \\ &= 1-\frac{\rho}{\lambda_J}\theta_{tot}\exp\{Tx\}\left(\boldsymbol{e}\otimes\boldsymbol{T}^0_{tot,J}\right). \end{aligned} \tag{35}$$

Note that in Equation (35), we used the fact that the departure rate of type J batches equals the arrival rate of type J batches. Similarly, we obtain, for $1\leq k\leq K$ and $x\geq 0$,

$$\begin{aligned} P\{d \leq x\} &= 1-\frac{\rho}{\lambda}\theta_{tot}\exp\{Tx\}\left(\boldsymbol{e}\otimes\boldsymbol{T}^0_{tot}\right), \\ P\{d_{(k)} \leq x\} &= 1-\frac{\rho}{\lambda_{(k)}}\theta_{tot}\exp\{Tx\}\left(\boldsymbol{e}\otimes\boldsymbol{T}^0_{tot,(k)}\right). \end{aligned} \tag{36}$$

By Lemma 4.2, it is easy to verify that the distributions given in Equations (35) and (36) are probability distributions with no mass at zero. Equations (35) and (36) indicate that the sojourn times have matrix exponential distributions[15]. In [18], it has shown that the waiting time of an arbitrary customer and the sojourn times have continuous time PH-distributions for the continuous time $SM/PH/1$ queue (also see [15–17]). These results can be extended to our queueing model.

Theorem 4.3 *Assume that the queueing system is stable. Let* $\Delta = \text{diag}\,(\theta_{tot})$ *and* $Q = \Delta^{-1}T'\Delta$. *For an arbitrary batch, the random variable* d *has a PH-distribution with a matrix representation* $\left\{m_a m_{tot},\ \rho\left(\Delta(\boldsymbol{e}\otimes\boldsymbol{T}^0_{tot})\right)'/\lambda,\ Q\right\}$. *For an arbitrary type* J *batch, the random variable* d_J *has a PH-distribution with a matrix representation* $\left\{m_a m_{tot},\ \rho\left(\Delta(\boldsymbol{e}\otimes\boldsymbol{T}^0_{tot,J})\right)'/\lambda_J,\ Q\right\}$. *For an arbitrary type* k *customer, the random variable* $d_{(k)}$ *has a PH-distribution with a matrix representation* $\left\{m_a m_{tot},\ \rho(\Delta(\boldsymbol{e}\otimes\boldsymbol{T}^0_{tot,(k)}))'/\lambda_{(k)}, Q\right\}$.

Proof The proof is based on Equation (36) and is similar to the proof of Theorem 5 in [18]. For completeness, a proof is given the appendix. ∎

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 it enters the server to find the waiting times of batches. Equivalently, the waiting time of a batch is $a_g(t)$ right after the departure of a (arbitrary) batch. We focus on the waiting time w of an arbitrary batch and the waiting time w_J of an arbitrary type J batch. By the definition of $a_g(t)$, we have, for $x \geq 0$,

$$\begin{aligned}
P\{w_J \leq x\} = 1 - P\{w_J > x\} &= 1 - \frac{1}{\lambda_J}\int_x^\infty \pi(t)\,(D_{a,J}(t-x)\otimes I)\,dt\,\left(\boldsymbol{e}\otimes\boldsymbol{T}^0_{tot}\right)\\
&= 1 + \frac{\rho}{\lambda_J}\theta_{tot}T\int_x^\infty \exp\{Tt\}\,(D_{a,J}(t-x)\otimes I)\,dt\,\left(\boldsymbol{e}\otimes\boldsymbol{T}^0_{tot}\right)\\
&= 1 - \frac{\rho}{\lambda_J}\theta_{tot}\int_x^\infty \exp\{Tt\}\,(D_{a,J}(dt-x)\otimes I)\,\left(\boldsymbol{e}\otimes\boldsymbol{T}^0_{tot}\right)\\
&= 1 - \frac{\rho}{\lambda_J}\theta_{tot}\int_0^\infty \exp\{T(t+x)\}\,(D_{a,J}(dt)\otimes I)\,\left(\boldsymbol{e}\otimes\boldsymbol{T}^0_{tot}\right)\\
&= 1 - \frac{\rho}{\lambda_J}\theta_{tot}\exp\{Tx\}R_J\left(\boldsymbol{e}\otimes\boldsymbol{T}^0_{tot}\right), \qquad (37)
\end{aligned}$$

where $R_J = \int_0^\infty \exp\{Tt\}\,(D_{a,J}(dt)\otimes I)$. Similarly, the waiting time distribution of an arbitrary batch can be obtained as:

$$P\{w \leq x\} = 1 - \frac{\rho}{\lambda}\theta_{tot}\exp\{Tx\}R\left(\boldsymbol{e}\otimes\boldsymbol{T}^0_{tot}\right), \qquad (38)$$

where $R = \int_0^\infty \exp\{Tt\}\Big(\sum_{J\in\aleph} D_{a,J}(dt)\otimes I\Big) = \sum_{J\in\aleph} R_J$.

It is easy to verify that the distribution functions given in Equations (37) and (38) are proper probability distributions with a mass at zero. From Equations (37) and (38), it can be shown that w_J and w have *PH*-distributions and their matrix representations can be constructed explicitly.

Theorem 4.4 *Assume that the queueing system is stable, i.e.,* $\rho < 1$. *Let* $\Delta = \text{diag}\,(\theta_{tot})$ *and* $Q = \Delta^{-1}T'\Delta$. *The waiting time* w *has a PH-distribution with a matrix representation* $\{m_a m_{tot},\ \rho(\Delta R(\boldsymbol{e}\otimes\boldsymbol{T}^0{}_{tot}))'/\lambda,\ Q\}$. *For the waiting time* w_J *of an arbitrary type* J *batch, it has a PH-distribution with a matrix representation* $\{m_a m_{tot},\ \rho(\Delta R_J(\boldsymbol{e}\otimes\boldsymbol{T}^0{}_{tot}))'/\lambda_J,\ Q\}$.

Proof The proof is based on Equation (38) and is similar to that of Theorem 5 in [18]. ∎

The above results can be used to find the distributions of the waiting 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 $1 \leq k \leq K$, we have

$$P\{w_{(k)} \le t\} = \sum_{J\in\aleph} \frac{\lambda_J}{\lambda_{(k)}} \sum_{t=1}^{|J|} P\{w_J + s_{j_1} + s_{j_2} + ... + s_{j_{t-1}} \le t\}\delta_{\{j_t=k\}}, \quad t \ge 0, \tag{39}$$

where $\delta_{\{.\}}$ is the indicator function, i.e., $\delta_{\{j=k\}} = 1$ if j=k; 0, otherwise. Since w_J has a PH-distribution and service times are PH-distributed, the random variable $w_J+s_{j_1}+s_{j_2}+\cdots+s_{j_{t-1}}$ has a PH-distribution. Since the total number N of different batches is finite, Equation (39) indicates that $w_{(k)}$ has a PH-distribution. The construction of the PH-representation of $w_{(k)}$ is straightforward but tedious. Therefore, details are omitted.

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 also find the distribution of d_J by w_J and s_J. That poses a question of consistence between the results obtained in this section and that of Section 4.2, i.e., the consistence of distributions of d_J obtained in Section 4.2 and w_J+s_J. This was an unresolved problem in [16] (see Remark 7 in [16]).

Corollary 4.5 *If the queueing system is stable, then* $E[\exp\{-sd_J\}] = E[\exp\{-sw_J\}]E[\exp\{-ss_J\}]$.

The proof of Corollary 4.5 is in the appendix. Corollary 4.5 shows that the results obtained in Section 4.2 and this section are consistent. The relationship given in Corollary 4.5 is useful for checking the accuracy of computational results.

Example 4.1 The Continuous Time $MMAP[K]/PH[K]/1$ Queue (Example 2.1 continued) For this special case, $R_J = \int_0^\infty \exp\{Tt\}\,(D_{a,J}(dt) \otimes I) = L(D_J \otimes I)$, where the matrix L is given in Example 3.1. We also have $R = \int_0^\infty \exp\{Tt\}\Big(\sum\limits_{J\in\aleph} D_{a,J}(dt)\otimes I\Big) = L\Big(\sum\limits_{J\in\aleph} D_J\otimes I\Big)$.

5 Distributions of Queue Lengths

In this section, some results on the queue length distributions are obtained by utilizing the results on the waiting times and sojourn times. The basic idea is that the queue at any time epoch consists of the batches arrived during the waiting time and service time of the batch currently in service. The queue can be observed at the batch level or at the customer level. Since information about the queue at the customer level can be obtained from the queue at the batch level, our focus is on the queue of batches of customers. We start with a special case where detailed results on the queue can be obtained.

Example 5.1 The $SM[K]/PH[K]/1$ Queue with $m_a = 1$ This queueing model is a direct extension of the model studied in [16]. Although this is a special case, it has not been considered before. Thus, the following results are new. Denote by

$$D_{a,H_1\cdots H_n}(t) = D_{a,H_1} * \cdots * D_{a,H_n}(t), \quad \text{for} \quad H_1, \cdots, H_n \in \aleph, \tag{40}$$

where "*" is for the convolution of functions. The probability that n batches $\{H_1, H_2, \cdots, H_n\}$ arrived in the order H_1, H_2, $\cdots$, and H_n in $[0, t]$ is given by $D_{a,H_1\cdots H_n}(t) - D_{a,H_1\cdots H_n} * D_a(t)$.

Let $x(H_1H_2\cdots H_n)$ be the probability that the queue consists of a type H_1 batch, a type H_2 batch, $\cdots$, and a type H_n batch (arrived in that order) right after the departure of an arbitrary batch; $x_J(H_1H_2\cdots H_n)$ be the probability that the queue is $H_1H_2\cdots H_n$ right after the departure of an arbitrary type J batch; $y_q(H_1H_2\cdots H_n)$ be the probability that the waiting queue is $H_1H_2\cdots H_n$ at an arbitrary time (excluding all customers in the batch in service (if any)), $y(H_1H_2\cdots H_n)$ be the probability that the queue is $H_1H_2\cdots H_n$ at an arbitrary time (including the batch (if any) in service). For all these cases, the n batches $\{H_1, H_2, \cdots, H_n\}$ arrived in the order H_1, H_2, $\cdots$, and H_n. Note that the customers in queue right after the

departure of a batch are these customers who arrived during the sojourn time of that batch. By conditioning on the age of the batch just departed, we obtain, for $J \in \aleph$,

$$\begin{aligned} & x_J(H_1H_2\cdots H_n) \\ &= -\frac{\rho}{\lambda_J}\theta_{tot}T\int_0^\infty \exp\{Tt\}\left((D_{a,H_1\cdots H_n}(t) - D_{a,H_1\cdots H_n} * D_a(t))\,\boldsymbol{e}\otimes \boldsymbol{T}^0_{tot,J}\right)dt \\ &= \frac{\rho}{\lambda_J}\theta_{tot}\int_0^\infty \exp\{Tt\}\left((D_{a,H_1\cdots H_n}(dt) - D_{a,H_1\cdots H_n} * D_a(dt))\,\boldsymbol{e}\otimes \boldsymbol{T}^0_{tot,J}\right) \\ &= \frac{\rho}{\lambda_J}\theta_{tot}(I-R)R_{H_1}\cdots R_{H_n}\left(\boldsymbol{e}\otimes \boldsymbol{T}^0_{tot,J}\right), \end{aligned} \tag{41}$$

where $R_J = \int_0^\infty \exp\{Tt\}D_{a,J}(dt)$ and $R = \sum_{J\in\aleph} R_J$. Note that the definitions of the matrices R_J and R are consistent with the definitions given right after Equations (37) and (38), respectively. Also note that the last equality in Equation (41) holds because $m_a = 1$. Similar to Equation (41), we have

$$\begin{aligned} x(H_1H_2\cdots H_n) &= \frac{\rho}{\lambda}\theta_{tot}(I-R)R_{H_1}\cdots R_{H_n}\left(\boldsymbol{e}\otimes \boldsymbol{T}^0_{tot}\right), \\ y_q(H_1H_2\cdots H_n) &= \begin{cases} 1-\rho\theta_{tot}R\boldsymbol{e}, & \text{if } n=0, \\ \rho\theta_{tot}(I-R)R_{H_1}\cdots R_{H_n}\boldsymbol{e}, & \text{if } n>0; \end{cases} \\ y(H_1H_2\cdots H_n) &= \begin{cases} 1-\rho, & \text{if } n=0, \\ \rho\theta_{tot}(I-R)R_{H_2}\cdots R_{H_n}\left(\boldsymbol{e}\otimes \boldsymbol{e}(H_1)\right), & \text{if } n>0, \end{cases} \end{aligned} \tag{42}$$

where $\boldsymbol{e}(J)$ is a column vector of the size m_{tot} such that all elements are zero except the elements corresponding to the string J, which are set to be one. With the distributions given in Equations (41) and (42), various performance measures of queue at a departure epoch and at an arbitrary time can be found at both the batch level and the customer level. The above results can be used to analyze relationship between waiting customers and the composition of the queue (see Example 6.3). Details are omitted.

Note 5.1 In Example 5.1, we found the distributions of queue string at departure epochs and at an arbitrary time if $m_a = 1$. For $m_a \geq 2$, define

$$\begin{aligned} R_{H_1H_2\cdots H_n} &= \int_0^\infty \exp\{Tt\}\left(D_{a,H_1H_2\cdots H_n}(dt)\otimes I\right); \\ R_{H_1H_2\cdots H_n,all} &= \int_0^\infty \exp\{Tt\}\left(D_{a,H_1H_2\cdots H_n} * D_a(dt)\otimes I\right). \end{aligned} \tag{43}$$

Similar to Equation (41), we have

$$x_J(H_1H_2\cdots H_n) = \frac{\rho}{\lambda_J}\theta_{tot}\left(R_{H_1\cdots H_n} - R_{H_1\cdots H_n,all}\right)\left(\boldsymbol{e}\otimes \boldsymbol{T}^0_{tot,J}\right). \tag{44}$$

According to [16, 18], elements in the matrix R_H can be interpreted as the expected number of times an arriving batch sees a type H batch in the system in a busy period. Unfortunately, unlike the case with $m_a = 1$, the computation of $R_{H_1\cdots H_n}$ is difficult since $R_{H_1\cdots H_n}$ may not be the product of the corresponding matrices $\{R_{H_1}, R_{H_2}, \cdots, R_{H_n}\}$. Thus, the extension to $m_a \geq 2$ is not straightforward. The computation of Equation (44) is an interesting problem for future research.

Since it is difficult to find product form solution if $m_a \geq 2$ for the distribution of the queue string, we consider the numbers of different types of batches in the system, for which the order

of the batches is not important. Denote by $\boldsymbol{q} = (q_1,\ q_2,\ \cdots,\ q_N)$ and $\boldsymbol{n} = (n_1,\ n_2,\ \cdots,\ n_N)$, where $\{q_1,\ q_2,\ \cdots,\ q_N\}$ and $\{n_1,\ n_2,\ \cdots,\ n_N\}$ are nonnegative integers. Let $x(\boldsymbol{q})$ be the probability that the queue consists of q_1 type J_1 batches, q_2 type J_2 batches, $\cdots$, and q_N type J_N batches right after the departure of an arbitrary batch; $x_J(\boldsymbol{q})$ be the probability that the queue is $\boldsymbol{q}$ right after the departure of an arbitrary type J batch; $y_q(\boldsymbol{q})$ be the probability that the waiting queue is $\boldsymbol{q}$ at an arbitrary time (excluding the customers in service(if any)). Let $P(\boldsymbol{n},\ t)$ be the (matrix) probability that there are n_1 type J_1 batches, n_2 type J_2 batches, $\cdots$, and n_N type J_N batches arrived in $[0,\ t]$. Note that the customers in queue right after the departure of a batch are these customers who arrived during the sojourn time of that batch. By conditioning on the age of the batch just departed, we obtain, for $\boldsymbol{q} \geq 0$,

$$\begin{aligned} x(\boldsymbol{q}) &= -\frac{\rho}{\lambda}\theta_{tot}T\int_0^\infty \exp\{Tt\}\left(P(\boldsymbol{q},t)\boldsymbol{e}\otimes \boldsymbol{T}_{tot}^0\right)dt, \\ x_J(\boldsymbol{q}) &= -\frac{\rho}{\lambda_J}\theta_{tot}T\int_0^\infty \exp\{Tt\}\left(P(\boldsymbol{q},t)\boldsymbol{e}\otimes \boldsymbol{T}_{tot,J}^0\right)dt, \\ y_q(\boldsymbol{q}) &= \begin{cases} 1-\rho-\rho\theta_{tot}T\displaystyle\int_0^\infty \exp\{Tt\}P(0,t)dt\boldsymbol{e}, & q=0, \\ -\rho\theta_{tot}T\displaystyle\int_0^\infty \exp\{Tt\}P(\boldsymbol{q},t)dt\boldsymbol{e}, & q\neq 0. \end{cases} \end{aligned} \tag{45}$$

In general, if an explicit formula of the distribution function $P(\boldsymbol{n},\ t)$ can be obtained, explicit formulas for the distributions given in Equation (45) can be obtained. Such an example is given below. The distribution of the total number of batches, the marginal distributions of the numbers of individual types of batches, and the marginal distributions of the numbers of individual types of customers and mean queue lengths can be obtained from the distributions given in Equation (45). In addition, we point out that these distributions can also be obtained directly by applying the above method (see Example 6.1).

Example 5.2 The Continuous Time $MMAP[K]/PH[K]/1$ Queue (Example 2.1 continued) For this example, we consider the computation of Equation (45). The probability matrix $P(\boldsymbol{n},\ t)$ satisfies

$$P^{(1)}(\boldsymbol{n},t) = P(\boldsymbol{n},t)D_0 + \sum_{i=1:\ n_i\geq 1}^{N} P(\boldsymbol{n}-\boldsymbol{e}(i),t)D_{J_i}, \tag{46}$$

where $\boldsymbol{e}(i)$ is a vector with all elements zero except the ith element which is one. By using Equation (46), the joint distributions of the queue lengths can be obtained. To evaluate the joint distributions of queue lengths, the key is to evaluate the following matrix, for $\boldsymbol{n} \neq 0$,

$$\begin{aligned} L(\boldsymbol{n}) &\equiv \int_0^\infty d\exp(Tt)\left(P(\boldsymbol{n},t)\otimes I\right) = -\int_0^\infty \exp(Tt)\left(P(\boldsymbol{n},dt)\otimes I\right) \\ &= -\int_0^\infty \exp(Tt)\left(\left(P(\boldsymbol{n},t)D_0 + \sum_{i=1:\ n_i\geq 1}^{N} P(\boldsymbol{n}-\boldsymbol{e}(i),t)D_{J_i}\right)\otimes I\right)dt \\ &= -T^{-1}L(\boldsymbol{n})\left(D_0\otimes I\right) - T^{-1}\sum_{i=1:\ n_i\geq 1}^{N} L(\boldsymbol{n}-\boldsymbol{e}(i))\left(D_{J_i}\otimes I\right). \end{aligned} \tag{47}$$

Thus, the matrices $\{L(\boldsymbol{n}),\, \boldsymbol{n}\geq 0\}$ can be obtained by solving the following linear equations:

$$\begin{aligned} &TL(0) + L(0)\,(D_0 \otimes I) = -T, \\ &TL(\boldsymbol{n}) + L(\boldsymbol{n})\,(D_0 \otimes I) = -\sum_{i=1:\ n_i \geq 1}^{N} L(\boldsymbol{n} - \boldsymbol{e}(i))\,(D_{J_i} \otimes I)\,, \quad \boldsymbol{n} \neq 0. \end{aligned} \tag{48}$$

Apparently, the direct-sum approach used in Example 3.1 can be used to solve Equation (48) for $L(\boldsymbol{n})$. The joint distributions of queue lengths at various time epochs can be obtained as:

$$\begin{aligned} x(\boldsymbol{q}) &= \frac{\rho}{\lambda}\theta_{tot} L(\boldsymbol{q})\left(\boldsymbol{e} \otimes \boldsymbol{T}_{tot}^{0}\right), && q \geq 0, \\ x_J(\boldsymbol{q}) &= \frac{\rho}{\lambda_J}\theta_{tot} L(\boldsymbol{q})\left(\boldsymbol{e} \otimes \boldsymbol{T}_{tot,J}^{0}\right), && q \geq 0, \\ y_q(\boldsymbol{q}) &= \begin{cases} 1 - \rho - \rho\theta_{tot} L(0)\boldsymbol{e}, & q = 0, \\ -\rho\theta_{tot} L(\boldsymbol{q})\boldsymbol{e}, & q \neq 0. \end{cases} \end{aligned} \tag{49}$$

Based on Equation (49), combining with Example 3.1, a simple algorithm can be developed for computing the queue length distributions of the number of different types of batches in the queue for the $MMAP[K]/PH[K]/1$ queue. Simple formulas can be found for other queue length related distributions and their means (see Example 6.1).

6 Numerical Examples

In this section, we report and analyze some numerical results for models introduced in Example 2.1 and Example 5.1.

Example 6.1 (Example 2.1 continued) For Example 2.1, we further assume that there are two types of customers and two types of batches $\{J_1 = 1,\, J_2 = 2\}$. System parameters are given as: $K = 2$, $\aleph = \{1,\, 2\}$, $m_a = 2$,

$$\begin{aligned} &D_0 = \begin{pmatrix} -2 & 1 \\ 0 & -5 \end{pmatrix}, \quad D_1 = \begin{pmatrix} 0 & 1 \\ 0.1 & 0 \end{pmatrix}, \quad D_2 = \begin{pmatrix} 0 & 0 \\ 1.9 & 3 \end{pmatrix}; \\ &m_1 = 2, \quad \alpha_1 = (0.8,\ 0.2),\ T_1 = \begin{pmatrix} -2 & 1.5 \\ 0 & -1 \end{pmatrix}; \\ &m_2 = 2, \quad \alpha_2 = (1,\ 0),\ T_2 = \begin{pmatrix} -25 & 5 \\ 0 & -25 \end{pmatrix}. \end{aligned} \tag{50}$$

By routine calculations, we have $\lambda_1 = 0.55$, $\lambda_2 = 2.45$, $\mu_1 = 0.833$, $\mu_2 = 20.833$, and $\rho = 0.777$. The mean sojourn time of type 1 customers is 4.079. The mean sojourn time of type 2 customers is 3.168. The mean waiting time of type 1 customers is 2.879. The mean waiting time of type 2 customers is 3.120. Since most of type 1 customers are followed by at least one type 2 customer, the waiting time of a type 2 customer is longer than that of a type 1 customer, even though the service of type 2 customers is quick. That shows the impact of type 1 customer's service times on the waiting times of type 2 customers.

For queue length distributions, equations in Example 5.2 can be used. We can also use the following formulas to compute the mean queue lengths directly. For instance, the mean total number of waiting customers at an arbitrary time is given by $E[y_{q,all}] = -\rho\theta_{tot} L_{all}\boldsymbol{e}$, where

$$\begin{cases} TP_{all} + P_{all}\,(D \otimes I) = -I, \\ TL_{all} + L_{all}\,(D \otimes I) = -P_{all} \sum_{i=1}^{N} D_{J_i} \otimes I. \end{cases} \tag{51}$$

The mean number of type 1 batches in the queue is 1.583 and the mean number of type 2 batches in the queue is 7.640. The mean total number of customers in queue (regardless of their types) is 9.223. The number of waiting type 1 customers is small, but their impact on the queueing process of type 2 customers is significant. Part of the joint distribution $y_q(q_1, q_2)$ (defined in Equation (49)) of the queue lengths of types 1 and 2 customers are given in Table 6.1.

Table 6.1 The joint queue length distribution $y_q(q_1, q_2)$

$q_1 \backslash q_2$	0	1	2	3	4
0	0.2834	0.0545	0.0390	0.0286	0.0210
1	0.0040	0.0103	0.0165	0.0187	0.0185
2	0.0004	0.0014	0.0032	0.0054	0.0073
3	0.0000	0.0002	0.0005	0.0011	0.0019
4	0.0000	0.0000	0.0001	0.0002	0.0004

Table 6.1 shows that the distributions of the queue lengths of the two types of customers are dependent. For type 1 customers, their queue length is short. But the presence of type 1 customers increases the queue length of type 2 customers significantly. This can be seen from the fact that the probability of the queue length of type 2 customers is increasing when q_2 increases from 0 to 3 if q_1 is positive.

Example 6.2 (Example 2.1 continued) In this example, we consider a queue with 5 types of customers that may arrive in batches. There are 6 different types of batches: $J_1 = 1$, $J_2 = 2$, $J_3 = 3$, $J_4 = 4$, $J_5 = 14$, and $J_6 = 12345$. The Markov arrival process is defined as: $m_a = 2$,

$$D_0 = \begin{pmatrix} -5 & 1 \\ 0 & -3 \end{pmatrix}, \quad D_1 = \begin{pmatrix} 0.05 & 0.05 \\ 0.2 & 0.1 \end{pmatrix}, \quad D_2 = \begin{pmatrix} 0.1 & 0.2 \\ 0.1 & 0.6 \end{pmatrix},$$
$$D_3 = \begin{pmatrix} 0 & 0.2 \\ 0.2 & 0.3 \end{pmatrix}, \quad D_4 = \begin{pmatrix} 2 & 1 \\ 0.5 & 0.4 \end{pmatrix}, \quad D_{14} = \begin{pmatrix} 0.05 & 0.05 \\ 0.05 & 0.05 \end{pmatrix}, \quad D_{12345} = \begin{pmatrix} 0.3 & 0 \\ 0 & 0.5 \end{pmatrix}. \tag{52}$$

The service times of the 5 types of customers are exponentially distributed with parameters $\{m_1{=}1, \alpha_1 = 1, T_1 = -4\}$, $\{m_2{=}1, \alpha_2 = 1, T_2 = -5\}$, $\{m_3{=}1, \alpha_3 = 1, T_3 = -6\}$, $\{m_4{=}1, \alpha_4 = 1, T_4 = -7\}$, and $\{m_5{=}1, \alpha_5 = 1, T_5 = -8\}$, respectively. The service time of the batch $J_5 = 14$ is the sum of the (independent) service times of the 2 customers in the batch. The service time of the batch $J_6 = 12345$ is the sum of the (independent) service times of all 5 customers in the batch. The PH-representations of the batch service times are

$$m_{14} = 2, \quad \alpha_{14} = (1,\ 0), \quad T_{14} = \begin{pmatrix} -4 & 4 \\ 0 & -7 \end{pmatrix},$$
$$m_{12345} = 5, \quad \alpha_{12345} = (1,\ 0,\ 0,\ 0,\ 0), \quad T_{12345} = \begin{pmatrix} -4 & 4 & 0 & 0 & 0 \\ 0 & -5 & 5 & 0 & 0 \\ 0 & 0 & -6 & 6 & 0 \\ 0 & 0 & 0 & -7 & 7 \\ 0 & 0 & 0 & 0 & -8 \end{pmatrix}. \tag{53}$$

For this example, the mean batch waiting times are $E[w_1] = 2.7675$, $E[w_2] = 2.7692$, $E[w_3] = 2.7687$, $E[w_4] = 2.7949$, $E[w_{14}] = 2.7778$, and $E[w_{12345}] = 2.7721$. We can further find the waiting times of individual types of customers. The mean numbers of type 1, type 2, type 3, type 4, type 14, type 12345 batches in queue are 0.6665, 1.6108, 1.1387, 4.2515, 0.2777, and 1.2221, respectively. It is easy to see that the mean waiting times are almost the same, but the numbers of different batches in queue can be quite different.

What is more interesting about this example is that, even though the service times of individual types of customers are exponential, the number of phases needed to formulate the Markov process $\{(a_g(t),\ I_a(t),\ J(t),\ I_s(t)),\ t{\geq}0\}$ is $m_{tot} = 22$. Thus, most of the matrices involved in computation are 22 by 22 matrices. The computation time of this example is significantly longer than that of Example 6.1. In fact, our numerical experimentations indicate that the computation time increases significantly if some of the service times have more phases or if there are more types of customers or types of batches. How to reduce the dimension of the matrices involved is an interesting topic for future research.

Example 6.3 (Example 5.1 continued) For Example 5.1, we further assume that the arrival process is deterministic defined by $m_a = 1$, for $1{\leq}n{\leq}N$,

$$D_{J_n}(t) = \begin{cases} 0, & t < t_n, \\ p_n, & t \geq t_n, \end{cases} \tag{54}$$

where $\{t_n,\ 1{\leq}n{\leq}N\}$ are positive constants and $\{p_n,\ 1{\leq}n{\leq}N\}$ are probabilities summing to one. It is easy to see that the total batch arrival rate $\lambda = (\Sigma_n p_n t_n)^{-1}$. The arrival rates of individual batch types are $\{\lambda_n = p_n\lambda,\ 1{\leq}n{\leq}N\}$. By routine calculations, we have $\theta_{tot} = \frac{\lambda}{\rho}\left[\sum_{n=1}^{N} \frac{p_n\beta(J_n)}{\mu_n}\right]$. Equation (21) for computing the matrix T can be simplified to $T = T_{tot} + \sum_{n=1}^{N} p_n \exp\{Tt_n\}\boldsymbol{T}_{tot}^0\alpha(J_n)$. After computing T, the distributions of sojourn times and waiting times can be found. For queue string distributions, the matrices $\{R_{J_n},\ 1{\leq}n{\leq}N\}$ can be obtained by $R_{J_n} = p_n \exp\{Tt_n\}$. By Equations (41) and (42), the probabilities for queue strings can be computed.

For instance, we assume that $K{=}2$, $N{=}3$, $\aleph{=}\{1,\ 2,\ 12\}$, $\{d_1{=}0.9,\ d_2{=}1.2,\ d_{12}{=}1.5\}$, $\{p_1{=}0.2,\ p_2{=}0.5,\ p_{12}{=}0.3\}$,

$$\begin{aligned}
&m_1 = 2, \quad \alpha_1 = (0.4,\ 0.6), \quad T_1 = \begin{pmatrix} -2 & 1 \\ 1 & -5 \end{pmatrix}; \\
&m_2 = 3, \quad \alpha_2 = (1,\ 0,\ 0), \quad T_2 = \begin{pmatrix} -1.5 & 1 & 0 \\ 1 & -3 & 0.1 \\ 0.1 & 0.2 & -5 \end{pmatrix}; \\
&m_{12} = 5, \quad \alpha_{12} = (0.4,\ 0.6,\ 0,\ 0,\ 0), \quad T_{12} = \begin{pmatrix} -2 & 1 & 1 & 0 & 0 \\ 1 & -5 & 4 & 0 & 0 \\ 0 & 0 & -1.5 & 1 & 0 \\ 0 & 0 & 1 & -3 & 0.1 \\ 0 & 0 & 0.1 & 0.2 & -5 \end{pmatrix}.
\end{aligned} \tag{55}$$

Note that the service time of a batch $J{=}12$ is the sum of the (independent) service times of a type 1 customer and a type two customer. For this system, $\rho = 0.9377$. The mean waiting times of the three types of batches are given by $E[w_1] = 6.7989$, $E[w_2] = 6.5310$, and $E[w_{12}] = 6.2730$, respectively. The mean sojourn times of the three types of batches are given by $E[d_1] = 7.2656$, $E[d_2] = 7.6810$, and $E[d_{12}] = 7.8897$, respectively. By using Equation (42), the distributions of the queue strings can be calculated. The probability $y_q(J)$ of the waiting queue string at an arbitrary time is given in the following table.

Table 6.2 The probability $y_q(H)$ of waiting queue string

H	0	1	2	12	-	-	-	-	-
$y_q(H)$	0.1949	0.0252	0.0609	0.0353	-	-	-	-	-
H_1, H_2	1,1	1,2	1,12	2,1	2,2	2,12	12,1	12,2	12,12
$y_q(H_1H_2)$	0.0045	0.0109	0.0062	0.0109	0.0261	0.0151	0.0062	0.0151	0.0087
H_1, H_2, H_3	1,1,1	1,1,2	1,2,1	1,2,2	2,1,1	2,1,2	2,2,1	2,2,12	1,1,12
$y_q(H_1H_2H_3)$	0.0008	0.0019	$\approx$0	0.0002	0.0019	0.0046	0.0002	0.0002	0.0011

Note that in Table 6.2, H, H_1, H_2, and H_3 are strings in $\aleph$. Table 6.2 provides information about the size of the queue as well as the relationship between individual customers in the queue. For instance, the queue string 112 has two constructions: {1, 1, 2} with three batches and {1, 12} with two batches. Table 6.2 shows that the probability to have a queue 112 at an arbitrary time is 0.0081. If the queue is seen as112, it is more likely to have two batches {1, 12} (with probability 0.0062) than three batches {1, 1, 2} (with probability 0.0019). Thus, if the waiting queue string is 112, it is more likely that the second type 1 customer arrived with the only type 2 customer in queue.

References

[1] E. Cinlar, Queues with semi-Markov arrivals, *J. Appl. Prob.*, 1967, **4**: 365–379.

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

[3] Q. M. He, Queues with marked customers, *Adv. Appl. Prob.*, 1996, **28**: 567–587.

[4] Q. M. He, The versatility of $MMAP[K]$ and the $MMAP[K]/G[K]/1$ queue, *Queueing Systems*, 2001, **38**(4): 397–418.

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

[6] D. M. Lucantoni, K. S. Meier-Hellstern, and M. F. Neuts, A single server queue with server vacations and a class of non-renewal arrival processes, *Adv. Appl. Prob.*, 1990, **22**: 676–705.

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

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

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

[10] V. Ramaswami, The $N/G/1$ queue and its detailed analysis, *Adv. Appl. Prob.*, 1980, **12**: 222–261.

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

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

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

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

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

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

[17] B. Sengupta, Phase-type representations for matrix-geometric solutions, *Stochastic Models*, 1990a, **6**: 163–167.

[18] B. Sengupta, The semi-Markovian queue: Theory and applications, *Stochastic Models*, 1990b, **6**: 383–413.
[19] Q. M. He, Workload process, waiting times, and sojourn times in a discrete time $MMAP[K]/SM[K]/1$/FCFS queue, *Stochastic Models*, 2004, **20**(4): 415–437.
[20] Q. M. He, Age process, total workload, sojourn times, and waiting times in a discrete time $SM[K]/PH[K]/1$/FCFS queue, *Queueing Systems*, 2005, **49**: 363–403.
[21] B. Van Houdt and C. Blondia, The delay distribution of a type k customer in an FCFS $MMAP[K]/PH[K]/1$ queue, *Journal of Applied Probability*, 2002a, **39**(1): 213–222.
[22] B. Van Houdt and C. Blondia, The waiting time distribution of a type k customer in a discrete-time FCFS $MMAP[K]/PH[K]/c$ (c=1,2) queue using QBDs, *Stochastic Models*, 2002b, **20**(1): 55–69.
[23] J. Lambert, B. Van Houdt, and C. Blondia, Queues in DOCSIS cable modem networks, *Computer and Operations Research*, 2008, **35**(8): 2482–2496.
[24] S. Asmussen and G. Koole, Marked point processes as limits of Markovian arrival streams, *J. Appl. Probab.*, 1993, **30**: 365–372.
[25] E. Cinlar, Markov renewal theory, *Adv. Appl. Prob.*, 1969, **1**: 123–187.
[26] Q. M. He and M. F. Neuts, Markov chains with marked transitions, *Stochastic Processes and Their Applications*, 1998, **74**(1): 37–52.
[27] R. M. Loynes, The stability of a queue with non-independent interarrival and service times, *Proc. Cambridge Philos. Soc.*, 1962, **58**: 497–520.
[28] F. R. Gantmacher, *The Theory of Matrices*, Chelsea, New York, 1959.
[29] H. Marcus and H. Minc, *A Survey of Matrix Theory and Matrix Inequalities*, Allyn and Bacon, Inc., Boston, 1964.
[30] J. F. Perez, J. Van Velthoven, and B. Van Houdt, Q-MAM: A tool for solving infinite queues using matrix-analytic methods, Proceedings of SMCtools'08, ACM Press, Athens (Greece), 2008.

Appendix The Proofs of Some Lemmas, Corollaries, and Theorems

The Proof of Lemma 3.1 By the definition of the traffic intensity ρ given in Equation (7), it is easy to verify that $\theta_{tot}\boldsymbol{e}= 1$. Then we have the following calculations:

$$\begin{aligned}
&\theta_{tot}A\\
&= \frac{\lambda}{\rho}\left(\sum_{i=1}^{N}(\theta_a D_{a,J_i})\otimes\left(\frac{\beta(J_i)}{\mu_{J_i}}\right)\right)\left(I\otimes T_{tot}+\sum_{j=1}^{N}D_{a,J_j}\otimes(\boldsymbol{T}^0_{tot}\alpha(J_j))\right)\\
&= \frac{\lambda}{\rho}\sum_{i=1}^{N}(\theta_a D_{a,J_i})\otimes\left(\frac{\beta(J_i)T_{tot}}{\mu_{J_i}}\right)+\frac{\lambda}{\rho}\sum_{i=1}^{N}\sum_{j=1}^{N}\left(\theta_a D_{a,J_i}D_{a,J_j}\right)\otimes\left(\frac{\beta(J_i)}{\mu_{J_i}}\boldsymbol{T}^0_{tot}\alpha(J_j)\right)\\
&= \frac{\lambda}{\rho}\sum_{i=1}^{N}(\theta_a D_{a,J_i})\otimes\left(0,\cdots,0,\frac{\beta_{J_i}T_{J_i}}{\mu_{J_i}},0,\cdots,0\right)+\frac{\lambda}{\rho}\sum_{j=1}^{N}\left(\left(\sum_{i=1}^{N}\theta_a D_{a,J_i}\right)D_{a,J_j}\right)\otimes\alpha(J_j)\\
&= \frac{\lambda}{\rho}\sum_{i=1}^{N}(\theta_a D_{a,J_i})\otimes\left(0,\cdots,0,\frac{\beta_{J_i}T_{J_i}}{\mu_{J_i}}+\alpha_{J_i},0,\cdots,0\right)\\
&= \frac{\lambda}{\rho}\sum_{i=1}^{N}(\theta_a D_{a,J_i})\otimes\left(0,\cdots,0,\frac{\beta_{J_i}T_{J_i}+\beta_{J_i}\boldsymbol{T}^0_{J_i}\alpha_{J_i}}{\mu_{J_i}},0,\cdots,0\right)\\
&= \frac{\lambda}{\rho}\sum_{i=1}^{N}(\theta_a D_{a,J_i})\otimes(0,\cdots,0,\ 0\ ,0,\cdots,0)=0. \qquad (56)
\end{aligned}$$

Note that $\beta_J\boldsymbol{T}_J{}^0 = 1/E[s_J] = \mu_J$[7] and $\beta_J(T_J + \boldsymbol{T}_J{}^0\alpha_J) = 0$. Therefore, θ_{tot} is an

invariant probability vector of A. Since A is irreducible, θ_{tot} is the unique invariant probability vector of A. This completes the proof of Lemma 3.1. ∎

Proof of Lemma 3.2 Let $\boldsymbol{u}(s)$ and $\boldsymbol{v}(s)$ be the left and right eigenvectors corresponding to $\chi(s)$, respectively, i.e., $\boldsymbol{u}(s)A^*(s) = \chi(s)\boldsymbol{u}(s)$ and $A^*(s)\boldsymbol{v}(s) = \chi(s)\boldsymbol{v}(s)$. The two vectors $\boldsymbol{u}(s)$ and $\boldsymbol{v}(s)$ are normalized by $\boldsymbol{u}(s)\boldsymbol{v}(s) = 1$ and $\boldsymbol{u}(s)\boldsymbol{e}= 1$. It is easy to see that $A^*(s)$ is irreducible and all the elements of the vectors $\boldsymbol{u}(s)$ and $\boldsymbol{v}(s)$ are positive for $s>0$. According to [7], the vectors $\boldsymbol{u}(s)$ and $\boldsymbol{v}(s)$ can be chosen as differentiable functions.

It is easy to see $\chi(0) = 0$. Furthermore, we have $\boldsymbol{u}(0) = \theta_{tot}$ (by the definition of $\boldsymbol{u}(s)$ and Lemma 3.1) and $\boldsymbol{v}(0) = \boldsymbol{e}$. By $\boldsymbol{u}(s)\boldsymbol{e}= 1$, we obtain $\boldsymbol{u}^{(1)}(s)\boldsymbol{e}= 0$. By taking derivatives on both sides of $\boldsymbol{u}(s)A^*(s) = \chi(s)\boldsymbol{u}(s)$, we obtain $\boldsymbol{u}^{(1)}(s)A^*(s) + \boldsymbol{u}(z)A^{*(1)}(s) = \chi^{(-1)}(s)\boldsymbol{u}(s) + \chi(s)\boldsymbol{u}^{(1)}(s)$. Letting $s=0$ and multiplying $\boldsymbol{e}$ on both sides of the equation, we obtain $\boldsymbol{u}(0)A^{*(1)}(0)\boldsymbol{e} = \chi^{(1)}(0)$, i.e., $\theta_{tot}\int_{\mathbf{0}}^{\infty} tdA(t)\boldsymbol{e} = \chi^{(1)}(0)$. Using Equations (17) and (18), we have

$$\begin{aligned}
\theta_{tot}\int_0^\infty tA(dt)\boldsymbol{e} &= \frac{\lambda}{\rho}\Big(\sum_{i=1}^{N}(\theta_a D_{a,J_i})\otimes\Big(\frac{\beta(J_i)}{\mu_{J_i}}\Big)\Big)\Big(\sum_{j=1}^{N}\int_0^\infty tD_{a,J_i}(dt)\boldsymbol{e}\otimes\boldsymbol{T}^0_{tot}\Big)\\
&= \frac{\lambda}{\rho}\sum_{j=1}^{N}\Big(\sum_{i=1}^{N}\Big(\theta_a D_{a,J_i}\int_0^\infty tD_{a,J_i}(dt)\boldsymbol{e}\Big)\cdot\Big(\frac{\beta_{J_i}\boldsymbol{T}^0_{J_i}}{\mu_{J_i}}\Big)\Big)\\
&= \frac{\lambda}{\rho}\sum_{j=1}^{N}\Big(\theta_a\int_0^\infty tD_{a,J_i}(dt)\boldsymbol{e}\Big)\\
&= \frac{\lambda}{\rho}\theta_a\int_0^\infty tD(dt)\boldsymbol{e} = \frac{\lambda}{\rho}E_{\theta_a}[\tau] = \frac{1}{\rho}. \qquad (57)
\end{aligned}$$

Note that the definition of λ (see Equation (4)) is used to obtain the last equality. Therefore, $\chi^{(1)}(0) = 1/\rho$. Then $\chi^{(1)}(0) > 1$ if and only if $\rho < 1$. ∎

Proof of Lemma 4.2 By definition,

$$\begin{aligned}
\theta_{tot}\left(\boldsymbol{e}\otimes\boldsymbol{T}^0_{tot}\right) &= \frac{\lambda}{\rho}\left(\sum_{i=1}^{N}(\theta_a D_{a,J_i})\otimes\left(\frac{\beta(J_i)}{\mu_{J_i}}\right)\right)\left(\boldsymbol{e}\otimes\boldsymbol{T}^0_{tot}\right)\\
&= \frac{\lambda}{\rho}\sum_{i=1}^{N}\frac{(\theta_a D_{a,J_i}\boldsymbol{e})(\beta_{J_i}\boldsymbol{T}^0_{J_i})}{\mu_{J_i}} = \frac{\lambda}{\rho}\sum_{i=1}^{N}\theta_a D_{a,J_i}\boldsymbol{e} = \frac{\lambda}{\rho}. \qquad (58)
\end{aligned}$$

Similarly, it can be shown that $\theta_{tot}(\boldsymbol{e}\otimes\boldsymbol{T}^0{}_{tot,J}) = \lambda_J/\rho$, and $\theta_{tot}(\boldsymbol{e}\otimes\boldsymbol{T}^0{}_{tot,(k)}) = \lambda_{(k)}/\rho$. This completes the proof of Lemma 4.2. ∎

Proof Theorem 4.3 First, by definition, we have

$$\begin{aligned}
P\{d < x\} &= 1 - \theta_{tot}\Delta^{-1}\exp\{\Delta T\Delta^{-1}x\}\left(\frac{\rho}{\lambda}\Delta\left(\boldsymbol{e}\otimes\boldsymbol{T}^0_{tot}\right)\right)\\
&= 1 - \boldsymbol{e}'\exp\{\Delta T\Delta^{-1}x\}\left(\frac{\rho}{\lambda}\Delta\left(\boldsymbol{e}\otimes\boldsymbol{T}^0_{tot}\right)\right)\\
&= 1 - \left(\frac{\rho}{\lambda}\Delta\left(\boldsymbol{e}\otimes\boldsymbol{T}^0_{tot}\right)\right)'\exp\{\Delta^{-1}T'\Delta x\}\boldsymbol{e}. \qquad (59)
\end{aligned}$$

Thus, we only need to verify the followings:

$$Q\boldsymbol{e} = \Delta^{-1}T'\Delta\boldsymbol{e} = \left(\theta_{tot}T\Delta^{-1}\right)' = \left(\theta_{tot}T\Delta^{-1}\right)' = -\left(\pi(0)\Delta^{-1}\right)'/\rho \leq 0;$$

$$0 \leq \left(\frac{\rho}{\lambda}\Delta\left(\boldsymbol{e}\otimes\boldsymbol{T}^0_{tot}\right)\right)';$$

$$\left(\frac{\rho}{\lambda}\Delta\left(\boldsymbol{e}\otimes\boldsymbol{T}^0_{tot}\right)\right)'\boldsymbol{e} = \frac{\rho}{\lambda}\theta_{tot}\left(\boldsymbol{e}\otimes\boldsymbol{T}^0_{tot}\right) = 1. \tag{60}$$

That proves the results for the sojourn time of an arbitrary batch. Similarly, we can prove the results for the other two cases. This completes the proof of Theorem 4.3. ∎

Proof of Corollary 4.5 The LSTs of the d_J, w_J, and s_J are given by

$$E\left[e^{-sd_J}\right] = -\frac{\rho}{\lambda_J}\theta_{tot}(sI-T)^{-1}T(\boldsymbol{e}\otimes\boldsymbol{T}^0_{tot,J});$$

$$E\left[e^{-sw_J}\right] = 1 - s\frac{\rho}{\lambda_J}\theta_{tot}(sI-T)^{-1}R_J(\boldsymbol{e}\otimes\boldsymbol{T}^0_{tot}); \tag{61}$$

$$E\left[e^{-ss_J}\right] = \alpha_J(sI-T_J)^{-1}\boldsymbol{T}^0_J.$$

We first show the following equalities:

$$\begin{aligned} &T(\boldsymbol{e}\otimes T^n_{tot}\boldsymbol{T}^0_{tot,J}) = \boldsymbol{e}\otimes T^{n+1}_{tot}\boldsymbol{T}^0_{tot,J} + R_J\left(\boldsymbol{e}\otimes\boldsymbol{T}^0_{tot}\right)\alpha_J T^n_J\boldsymbol{T}^0_J, \quad n\geq 0;\\ &\beta_J T^{n+1}_J\boldsymbol{T}^0_J = -\left(\beta_J\boldsymbol{T}^0_J\right)\alpha_J T^n_J\boldsymbol{T}^0_J, \quad n\geq 0. \end{aligned} \tag{62}$$

For the first equality in Equation (62), we have

$$\begin{aligned} &T(\boldsymbol{e}\otimes T^n_{tot}\boldsymbol{T}^0_{tot,J})\\ &= \left(I\otimes T_{tot} + \sum_{H\in\aleph}\int_0^\infty \exp\{Tt\}\left(D_{a,H}(dt)\otimes\boldsymbol{T}^0_{tot}\alpha(H)\right)\right)(\boldsymbol{e}\otimes T^n_{tot}\boldsymbol{T}^0_{tot,J})\\ &= \left(I\otimes T_{tot} + \int_0^\infty \exp\{Tt\}\left(D_{a,J}(dt)\otimes\boldsymbol{T}^0_{tot}\alpha(J)\right)\right)(\boldsymbol{e}\otimes T^n_{tot}\boldsymbol{T}^0_{tot,J})\\ &= \boldsymbol{e}\otimes T^{n+1}_{tot}\boldsymbol{T}^0_{tot,J} + \int_0^\infty \exp\{Tt\}\left(D_{a,J}(dt)\otimes I\right)(\boldsymbol{e}\otimes\boldsymbol{T}^0_{tot})\alpha(J)T^n_{tot}\boldsymbol{T}^0_{tot,J}\\ &= \boldsymbol{e}\otimes T^{n+1}_{tot}\boldsymbol{T}^0_{tot,J} + R_J(\boldsymbol{e}\otimes\boldsymbol{T}^0_{tot})\left(\alpha_J T^n_J\boldsymbol{T}^0_J\right). \end{aligned} \tag{63}$$

The second equality in Equation (62) is shown as follows:

$$\begin{aligned} \beta_J(T_J+\boldsymbol{T}^0_J\alpha_J) = 0 \;&\Rightarrow\; \beta_J(T_J+\boldsymbol{T}^0_J\alpha_J)T^n_J\boldsymbol{T}^0_J = 0\\ &\Rightarrow\; \beta_J T^{n+1}_J\boldsymbol{T}^0_J = -\left(\beta_J\boldsymbol{T}^0_J\right)\alpha_J T^n_J\boldsymbol{T}^0_J. \end{aligned} \tag{64}$$

Using equalities in Equation (62), we have

$$
\begin{aligned}
E\left[e^{-sd_J}\right] &= -\frac{\rho}{\lambda_J}\theta_{tot}(sI-T)^{-1}T(\boldsymbol{e}\otimes\boldsymbol{T}^0_{tot,J}) \\
&= -\frac{\rho}{\lambda_J}\theta_{tot}(sI-T)^{-1}\left(\boldsymbol{e}\otimes T_{tot}\boldsymbol{T}^0_{tot,J} + R_J\left(\boldsymbol{e}\otimes\boldsymbol{T}^0_{tot}\right)\alpha_J\boldsymbol{T}^0_J\right) \\
&= -\frac{\rho}{\lambda_J}\theta_{tot}\frac{\left(I+(sI-T)^{-1}T\right)}{s}\left(\boldsymbol{e}\otimes T_{tot}\boldsymbol{T}^0_{tot,J} + R_J\left(\boldsymbol{e}\otimes\boldsymbol{T}^0_{tot}\right)\alpha_J\boldsymbol{T}^0_J\right) \\
&= -\frac{\rho\theta_{tot}}{\lambda_J}\left[\frac{\boldsymbol{e}\otimes T_{tot}\boldsymbol{T}^0_{tot,J}}{s} + (sI-T)^{-1}R_J\left(\boldsymbol{e}\otimes\boldsymbol{T}^0_{tot}\right)\alpha_J\boldsymbol{T}^0_J\right. \\
&\qquad\left. + \frac{(sI-T)^{-1}T\left(\boldsymbol{e}\otimes T_{tot}\boldsymbol{T}^0_{tot,J}\right)}{s}\right] \\
&= -\frac{\rho\theta_{tot}}{\lambda_J}\sum_{n=1}^{\infty}\frac{1}{s^n}\left[\boldsymbol{e}\otimes T^n_{tot}\boldsymbol{T}^0_{tot,J} + s(sI-T)^{-1}R_J\left(\boldsymbol{e}\otimes\boldsymbol{T}^0_{tot}\right)\alpha_J T^{n-1}_J\boldsymbol{T}^0_J\right] \\
&= -\frac{\rho\theta_{tot}}{\lambda_J}\left(\sum_{n=1}^{\infty}\frac{\boldsymbol{e}\otimes T^n_{tot}\boldsymbol{T}^0_{tot,J}}{s^n} + s(sI-T)^{-1}R_J\left(\boldsymbol{e}\otimes\boldsymbol{T}^0_{tot}\right)\alpha_J(sI-T_J)^{-1}\boldsymbol{T}^0_J\right) \\
&= \sum_{n=1}^{\infty}\frac{\alpha_J T^{n-1}_J\boldsymbol{T}^0_J}{s^n} - \frac{\rho\theta_{tot}}{\lambda_J}s(sI-T)^{-1}R_J\left(\boldsymbol{e}\otimes\boldsymbol{T}^0_{tot}\right)\alpha_J(sI-T_J)^{-1}\boldsymbol{T}^0_J \\
&= \left(1 - s\frac{\rho\theta_{tot}}{\lambda_J}(sI-T)^{-1}R_J\left(\boldsymbol{e}\otimes\boldsymbol{T}^0_{tot}\right)\right)\alpha_J(sI-T_J)^{-1}\boldsymbol{T}^0_J \\
&= E\left[e^{-sw_J}\right]E\left[e^{-ss_J}\right]. \qquad (65)
\end{aligned}
$$

This completes the proof of Corollary 4.5. ∎