Workload Process, Waiting Times, And Sojourn Times In A Discrete Time *MMAP*[*K*]/*SM*[*K*]/1/FCFS Queue

Qi-Ming HE

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

Abstract: In this paper, we study the total workload process and waiting times in a queueing system with multiple types of customers and a first-come-first-served service discipline. An *M*/*G*/1 type Markov chain, which is closely related to the total workload in the queueing system, is constructed. A method is developed for computing the steady state distribution of that Markov chain. Using that steady state distribution, the distributions of total workload, batch waiting times, and waiting times of individual types of customers are obtained. Compared to the *GI*/*M*/1 and QBD approaches for waiting times and sojourn times in discrete time queues, the dimension of the matrix blocks involved in the *M*/*G*/1 approach can be significantly smaller.

Key words: Queueing systems, *M*/*G*/1 type Markov chain, waiting times, sojourn times, workload, mean-drift method, matrix analytic methods, ergodicity, and semi-Markov chain.

1. Introduction

Discrete time queueing systems are of special interest to the performance analysis of telecommunications systems and manufacturing systems. For instance, discrete time queues are suitable for analyzing ATM in telecommunications (see Alfa [1], Cortizo, et al. [4], and references therein). Therefore, the study of discrete time queues has attracted great attention from researchers and practitioners.

In this paper, we introduce and study a discrete time queueing model with multiple types of customers. Such a queueing model arises naturally from modern telecommunication networks that are required to handle different types of data from telephone service, video conferences, high volume file transfer service, high definition TV distribution service, etc. B-ISDN is such an example. For such systems, different services may have different requirements in terms of delay times and loss rates. To quantify quality of service (QoS) for individual services, there is a need to differentiate them and to carry out performance analysis at the individual service level.

For queueing models with a last-come-first-served (LCFS) service discipline, the analysis of waiting time, queue strings, and busy periods can be carried out through the use of Markov chains with a tree structure (e.g., HE amd Alfa [13]). However, the analysis of the queue strings in queueing models with a first-come-first-served (FCFS) service discipline is difficult, since we must keep track of the type of each customer in queue. Nonetheless, significant progress has been made for continuous time queueing models with multiple types of customers (e.g., HE [9, 10], HE and Alfa [13], Takine [23, 24], and Takine and Hasegawa [25])). The study of discrete time queues with multiple types of customers is limited, except for Van Houdt and Blondia [26, 27, 28]. In this paper, we analyze a discrete time queueing model with a single server, multiple types of customers, and a FCFS service discipline. The focal point is on the waiting times of individual types of customers. As is the situation in the continuous case, by constructing an *M*/*G*/1 type Markov chain associated with the total workload in the system, we are able to find the steady state distribution of the total workload in the system at an arbitrary time. This enables us to find distributions of the waiting times and sojourn times for batches and for individual types of customers.

The study of waiting times in queues is extensive (e.g., Alfa [1], Cohen [3], Grassmann and Jain [8], Neuts [18], Sengupta [21], Takine [24, 25], and references therein). For discrete time queues, a number of methods have been used to study waiting times. Grassmann and Jian [8] used the Wiener-Hopf factorization method to develop an algorithm for computing the distribution of waiting times. Recently, the matrix analytic methods have been used to study discrete time queues (Alfa [1], Van Houdt and Blondia [26, 27, 28], Yang and Chaudhry [30], etc.). Such methods take advantage of the structures of the Markov chains under consideration and develop efficient algorithms for computing system performance measures. The structures of interest include the *M*/*G*/1 type and *GI*/*M*/1 type (Neuts [17, 19]). In Van Houdt and Blondia [26, 27, 28], *GI*/*M*/1 type Markov chains and QBD processes were constructed for studying the waiting times and sojourn times of discrete time queues with a Markov arrival process with marked transitions and PH-distributed service times. In HE [11], a *GI*/*M*/1 Markov chain is constructed for waiting times and sojourn times of a *discrete* time queue with a semi-Markovian arrival process and PH-distributed service times. In HE [12], a *GI*/*M*/1 Markov process is constructed for waiting times and sojourn times of a *continuous* time queue with a semi-Markovian arrival process and PH-distributed service times. In this paper, the basic tools are matrix analytic methods. We formulate an *M*/*G*/1 type Markov chain to obtain the steady state distributions of waiting times and sojourn times.

The *M*/*G*/1 type Markov chains have been used in the study of various queueing models. For instance, the classical *M*/*G*/1 queue and its variants, the *MAP*/*G*/1 queue and its variants can be analyzed efficiently by using an *M*/*G*/1 type Markov chain. The basic theory concerning *M*/*G*/1 type Markov chains can be found in Gail, et al. [6], Latouche and Ramaswami [15], Neuts [17, 19], and Ramaswami [20]. In Takine [22], a continuous time version of the *M*/*G*/1 type Markov process was introduced and investigated. As an application, the process was used in the study of the total workload and waiting times of a continuous time *MAP*/*G*/1 queue. This paper considers an extension of the idea used in [22] from the continuous time case to the discrete time case. The major differences are 1) we consider a discrete time queueing model; 2) we consider a

queueing system with multiple types of customers; and 3) the $M/G/1$ type Markov chain in this paper has no boundary at the level zero.

The approach developed in this paper – an $M/G/1$ approach – is for queues with general service times, where the methods ($GI/M/1$ or QBD) developed in HE [11] and Van Houdt and Blondia [26, 27, 28] does not apply. Even for the $MMAP[K]/PH[K]/1$ queue, for which all three methods apply, the $M/G/1$ approach has its advantages. As was pointed by Van Houdt and Blondia [28], the dimensions of the matrices involved with the $GI/M/1$ and QBD approaches can be quite large, which increases computation time and space requirement significantly. The size of transition blocks in the $M/G/1$ approach can be significantly smaller.

The rest of the paper is organized as follows. In Section 2, the discrete time $MMAP[K]/SM[K]/1$ queue is introduced. In Section 3, an $M/G/1$ type Markov chain associated with the total workload is defined and its ergodicity condition is found. The steady state distributions of the Markov chain, the total workload, waiting times, sojourn times of batches of customers are given in Section 4. In Section 5, two special cases – the $MMAP[K]/GI[K]/1$ queue and the $MMAP[K]/PH[K]/1$ queue – are considered and more detailed results are obtained. Finally, in Section 6, numerical examples that provide insight into the performance of the queueing models of interest are presented.

2. The Discrete Time *MMAP*[*K*]/*SM*[*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 type, join a single queue and are served by a single server on a first-come-first-served (FCFS) basis. In the reminder of this section, we describe the customer arrival process, the queueing process, and the service process in detail.

The customer arrival process The customer arrival process is a discrete time Markov arrival process with marked transitions ($MMAP[K]$) (Asmussen and Koole [2], HE [9, 10] and HE and Neuts [14]). Customers are distinguished into K types and arrive in batches. To characterize the batches of customers, we introduce a set of strings of integers denoted by

$$\aleph = \{J:\ J = j_1 j_2 \cdots j_n,\ 1 \le j_i \le K,\ 1 \le i \le n,\ n \ge 1\}. \tag{2.1}$$

For the queueing system, the string $J = j_1 j_2 \cdots j_n \in \aleph$ represents a batch with n customers. These n customers are of the types j_1, j_2, …, and j_n, respectively. Within the batch, the n customers are ordered as j_1, j_2, …, and j_n. We call J a string representation of that batch.

The system status is observed at integer epochs $t = 0, 1, 2, \ldots$. Let $I_a(t)$ be the phase of the underlying Markov chain of the arrival process in the time period $[t, t+1)$, which will be called period t, and $X(t)$ be the string representation of the batch associated with the transition in period t if there is an arrival; otherwise, $X(t) = 0$. Then we define

$$P\{I_a(t+1) = j,\ X(t+1) = J \mid I_a(t) = i\} = d_{J,i,j}, \quad 1 \le i, j \le m_a,\ t \ge 0,\ J \in \{0\} \cup \aleph, \tag{2.2}$$

where m_a is the number of phases of the underlying Markov chain. The constant $d_{J,i,j}$ is the probability that a batch J arrives in period t+1 and the phase of the underlying Markov chain is j in period t+1, given that the phase was i in period t. Let D_J be an $m_a \times m_a$ matrix with (i, j)-th element $d_{J,i,j}$. Let D_0 be an $m_a \times m_a$ substochastic matrix for no arrival in a period (i.e., $X(t) = 0$). The set of matrices $\{D_0, D_J, J \in \aleph\}$ provides all information about the Markov arrival process with marked transitions.

Let D be the sum of all matrices in the set $\{D_0, D_J, J \in \aleph\}$. The matrix D is the transition probability matrix of the underlying Markov chain $I_a(t)$. We assume that D is irreducible and $D \neq D_0$. Let θ_a be the (unique) invariant probability vector of the stochastic matrix D, i.e., $\theta_a D = \theta_a$ and $\theta_a \mathbf{e} = 1$, where $\mathbf{e}$ is a column vector with all elements being one. We note that the irreducibility of D implies that the Perron-Frobenius eigenvalue of D_0 (i.e., the eigenvalue with the largest modulus) is less than one since $D \neq D_0$. We refer readers to Gantmacher [7] for more about nonnegative matrices.

In steady state, the total arrival rate of batches of customers is given by $\lambda = 1 - \theta_a D_0 \mathbf{e}$. Let $\lambda_J = \theta_a D_J \mathbf{e}$ for $J \in \aleph$, i.e., the average arrival rate of batches of the type J. The arrival rate of type k customers is given by

$$\lambda_{(k)} = \sum_{J \in \aleph} N(J,k)\lambda_J, \quad 1 \le k \le K, \tag{2.3}$$

where $N(J, k)$ is the number of appearances of the integer k in the string J. Formulas for the arrival rates can be derived by using the classical generating function approach (See HE [10] and Neuts [17, 19]).

The queueing process After a batch of customers has 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 according to their order in the batch. Let $q(t)$ be a string of integers for the queue at period t, which is obtained after possible service completion and arrival in period t–1. Suppose that the services of individual customers in a batch can be distinguished. If $q(t) = j_1 j_2 \cdots j_n$, 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, …, and the last customer waiting in queue is of type j_n. These n customers shall be served in the order $j_1, j_2, \ldots, j_n$. If a batch $J = h_1 h_2 \cdots h_k$ arrives next, the queue becomes $j_1 j_2 \cdots j_n h_1 h_2 \cdots h_k$. If the service of customer j_1 is completed next, the queue becomes $j_2 j_3 \cdots j_n$ and the server starts serving customer j_2. If services are defined for batches only, then the queueing process can be described in terms of batches of customers in a similar manner. According to HE [10], 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 [10] has shown

that the string representation provides great flexibility in modeling with respect to the service order within a batch of customers.

The service process We assume that the service process and the arrival process are independent. The service process (of batches) is governed by a semi-Markov chain with m_s states. Let η_n be the phase of the semi-Markov process after the n-th transition (service). Let X_n the string representation of the n-th batch in service. Then we assume

$$P\{\eta_n = j, s_{X_n} = t \mid X_n = J, \eta_{n-1} = i\} = c_{J,i,j}(t), \quad 1 \le i, j \le m_s, n \ge 1, J \in \aleph, \qquad (2.4)$$

where s_J is the service time of a batch J and t is a positive integer. We assume that the service time of any batch is at least one. In equation (2.4), the current service is in the phase i and the phase becomes j after the service completion, which is the phase of the next service. Note that, if the queueing system is empty, the phase of the underlying semi-Markov chain remains the same. The constant $c_{J,i,j}(t)$ is the probability that the service time is t and the phase becomes j after the completion of the service, given that the current phase is i and the batch in service is of type J. Let $C_J(t)$ be an $m_s \times m_s$ matrix with (i, j)-th element $c_{J,i,j}(t)$. The set of matrices $\{C_J(t), t \ge 1, J \in \aleph\}$ provides all information about the service process. For any $J \in \aleph$, we denote

$$C = \sum_{t=1}^{\infty} C_J(t), \qquad C(t) = \sum_{J \in \aleph} C_J(t), \quad t \ge 1. \qquad (2.5)$$

The matrix C is the transition probability matrix of the underlying Markov chain $\{\eta_n, n \ge 0\}$, which is independent of J (i.e., the type of batch in service has no impact on the environment). We assume that C is irreducible. Let θ_s be the (unique) invariant probability vector of the stochastic matrix C, i.e., $\theta_s C = \theta_s$ and $\theta_s \mathbf{e} = 1$. In steady state, the mean service time of a batch J can be calculated as: $E_{\theta_s, J}(s) = \theta_s \left(\sum_{t=1}^{\infty} t C_J(t) \right) \mathbf{e}$. The service rate of batches of the type J is given as $\mu_J = \left(E_{\theta_s, J}(s) \right)^{-1}$, i.e., the average number of type J batches that can be served per period of time.

The traffic intensity of the queueing system is defined as:

$$\rho = \sum_{J \in \aleph} \lambda_J / \mu_J. \qquad (2.6)$$

According to Loynes [16], the queueing system can reach its steady state if and only if $\rho < 1$. Therefore, throughout this paper, we assume $\rho < 1$.

In the above definition of the queueing model, the service process is general. Later in this paper, we shall consider some special cases such as the case with independent service times (*MMAP*[*K*]/*GI*[*K*]/1) and the case with PH-distributed service times (*MMAP*[*K*]/*PH*[*K*]/1).

3. The Generalized Total Workload Process

The basic idea to analyze the waiting times originates 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

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

where τ_{n+1} is the length of the time between the n-th batch and the (n+1)-st batch, X_n is the string representation of the n-th batch, and s_{X_n} is the service time of the n-th batch. The process $\{w_n, n\geq 0\}$ has a repeating structure, which is difficult to deal with (see Zhao, et al. [29]). To study the waiting times, we construct some processes closely related to $\{w_n, n\geq 0\}$, which can be dealt with under certain conditions on the arrival or the service process. Some examples of such processes are the age process and the total workload (the virtual waiting time) process. The total workload process is used in this paper and the age process is used in HE [11] for the study of waiting times and sojourn times.

Based on equation (3.1), we introduce a process $v_g(t)$ related to the total workload in the queueing system as

$$v_g(t) = w_{n(t)} + s_{X_{n(t)}} - (t - t_{n(t)}). \tag{3.2}$$

where $n(t)$ represents the ordinal number of the last batch arrived in or before period t, and $t_{n(t)}$ is the arrival time of the $n(t)$-th batch. It is readily seen that $n(t)$, $w_{n(t)}$, $X_{n(t)}$, $s_{X_{n(t)}}$, and $t_{n(t)}$ update their values if a batch arrives in period t. The relationship between $n(t)$, $w_{n(t)}$, $t_{n(t)}$, $s_{X_{n(t)}}$, and $v_g(t)$ is shown in Figure 3.1.

Figure 3.1 Variables $v_g(t)$, $n(t)$, $t_{n(t)}$, $w_{n(t)}$, and $s_{X_{n(t)}}$ when $v_g(t)>0$.

The process $\{v_g(t), t\geq 0\}$ evolves as follows. At a batch arrival epoch (say the n-th batch), the total workload in the queueing system is calculated as $w_n + s_{X_n}$, which is the value of $v_g(t)$ in that period. During the arrival of the next batch, the value of $v_g(t)$ decreases by one per period. If $v_g(t) \geq 0$, $v_g(t)$ represents the total workload in the queueing system in period t. If $v_g(t) < 0$, the queueing system is empty in period t. The value $-v_g(t)$ indicates how long the current idle period

has been (or the age of that idle period in period t). When the next batch arrives, the value of $v_g(t)$ is updated as the total workload in that period and the next cycle begins. Note that, if $v_g(t)$ is negative and there is an arrival, then $v_g(t)$ jumps upward to the service time of the arrived batch. Compared to the process $\{w_n, n\geq 0\}$, $\{v_g(t), t\geq 0\}$ provides information about the workload at any time epoch. In fact, $\{w_n, n\geq 0\}$ can be considered as an embedded process of $\{v_g(t), t\geq 0\}$ at batch arrival epochs.

In order to construct a Markov chain related to $\{v_g(t), t\geq 0\}$, we introduce the following auxiliary variables. We define a process $\{I_s(t), t\geq 0\}$ from the Markov chain $\{\eta_n, n\geq 0\}$ as follows: $I_s(t) = \eta_n$ if the n-th batch is the last one arrived in or before period t. Thus, $I_s(t)$ is constant between two consecutive batch arrivals. Recall that $X(t)$ is the batch that arrived in period t. If there is no arrival in period t, then $X(t) = 0$. Then equation (3.2) leads to

$$v_g(t+1) = \begin{cases} v_g(t) - 1, & \text{if} \quad X(t) = 0; \\ \max\{0, v_g(t) - 1\} + s_{X(t)}, & \text{if} \quad X(t) \neq 0. \end{cases} \tag{3.3}$$

Note that $s_0 = 0$ in equation (3.3). Since the arrival process is governed by a Markov chain, the process $\{(v_g(t), I_a(t), I_s(t)), t\geq 0\}$ has the Markovian property during no arrival periods. The phase change in the service process is determined by the Markov chain $\{\eta_n, n\geq 0\}$. The service time of the batches depends on the Markov chain $\{\eta_n, n\geq 0\}$ as well. Thus, the process $\{(v_g(t), I_a(t), I_s(t)), t\geq 0\}$ has the Markovian property in batch arrival periods. Therefore, the process $\{(v_g(t), I_a(t), I_s(t)), t\geq 0\}$ is a Markov chain. Since the value of $v_g(t)$ can decrease at most by one per period, $v_g(t)$ has the skip-free to the left property. Therefore, $\{(v_g(t), I_a(t), I_s(t)), t\geq 0\}$ is a *Markov chain of M/G/1 type with no boundary at the level zero.* The variables $v_g(t)$, $I_a(t)$, and $I_s(t)$ take integer values and their ranges are: $-\infty < v_g(t) < \infty$, $1 \leq I_a(t) \leq m_a$, and $1 \leq I_s(t) \leq m_s$. A typical sample path of $v_g(t)$ is shown in Figure 3.2.

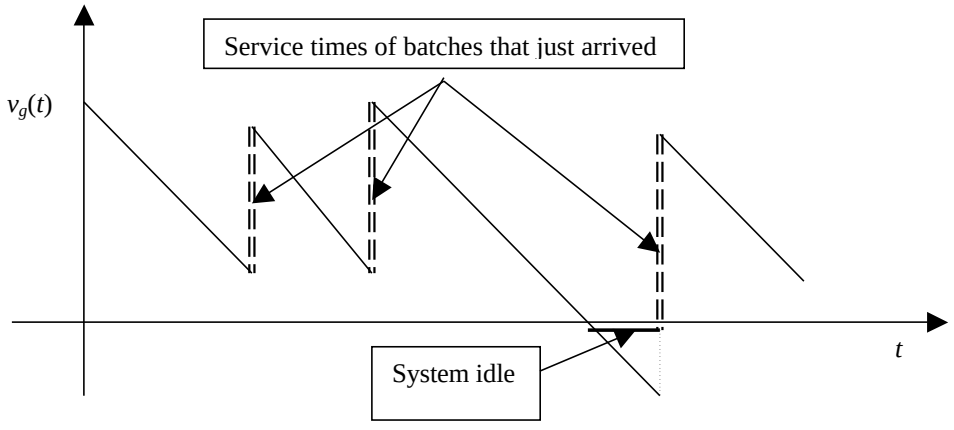

Figure 3.2 A sample path of $v_g(t)$

The transition probability matrix of $\{(v_g(t), I_a(t), I_s(t)), t\geq 0\}$ is given as

$$P_g = \begin{array}{c} \\ \vdots \\ -1 \\ 0 \\ 1 \\ 2 \\ \vdots \end{array} \begin{array}{c} \begin{array}{ccccccc} \cdots & -1 & 0 & 1 & 2 & 3 & \cdots \end{array} \\ \begin{pmatrix} \ddots & \ddots & \vdots & \vdots & \vdots & \vdots & \cdots \\ \ddots & 0 & 0 & A_1 & A_2 & A_3 & \cdots \\ & A_0 & 0 & A_1 & A_2 & A_3 & \cdots \\ & & A_0 & A_1 & A_2 & A_3 & \cdots \\ & & & A_0 & A_1 & A_2 & \cdots \\ & & & & \ddots & \ddots & \ddots \end{pmatrix} \end{array} \leftarrow \text{level } 0, \text{ i.e., } v_g(t) = 0, \tag{3.4}$$

where

$$A_0 = D_0 \otimes I; \quad A_n = \sum_{J \in \aleph} D_J \otimes C_J(n), \ n \geq 1, \tag{3.5}$$

where the notation "⊗" is for Kronecker product of matrix (see Gantmacher [7]). Note that the value of $v_g(t)$ determines the level of the Markov chain. Also note that the Perron-Frobenius eigenvalue of A_0 is equal to the Perron-Frobenius eigenvalue of D_0, which is less than one. The transition probabilities in P_g can be verified as follows:

$$\begin{aligned} &P\{v_g(t+1) = n-1, I_a(t+1) = j', I_s(t+1) = i' \mid v_g(t) = n, I_a(t) = j, I_s(t) = i\} \\ &= \begin{cases} (D_0)_{j,j'}, & \text{if} \quad i = i'; \\ 0, & \text{if} \quad i \neq i'. \end{cases} \qquad (3.6) \\ &P\{v_g(t+1) = n-1+s, I_a(t+1) = j', I_s(t+1) = i' \mid v_g(t) = n, I_a(t) = j, I_s(t) = i\} \\ &= \sum_{J \in \aleph} P\{J(t) = J, I_a(t+1) = j' \mid I_a(t) = j\} P\{s_J = s, I_s(t+1) = i' \mid I_s(t) = i\} \\ &= \sum_{J \in \aleph} (D_J)_{j,j'} (C_J(s))_{i,i'}, \quad s \geq 1, n \geq 1; \\ &P\{v_g(t+1) = s, I_a(t+1) = j', I_s(t+1) = i' \mid v_g(t) = n, I_a(t) = j, I_s(t) = i\} \\ &= \sum_{J \in \aleph} (D_J)_{j,j'} (C_J(s))_{i,i'}, \quad s \geq 1, n \leq 0. \end{aligned}$$

We shall use the steady state distribution of the Markov chain $\{(v_g(t), I_a(t), I_s(t)), t \geq 0\}$ to find the distributions of waiting times. For that purpose, we show that such a steady state distribution exists if the queueing system is stable, i.e., $\rho < 1$. First, we prove the following properties associated with the transition blocks in P_g. Define

$$A^*(z) = \sum_{n=0}^{\infty} z^n A_n, \quad z \geq 0; \qquad A = A^*(1) = \sum_{n=0}^{\infty} A_n, \tag{3.7}$$

if the summations are well-defined. By definition (3.5), we have

$$A = D_0 \otimes I + \sum_{n=1}^{\infty} \sum_{J \in \aleph} D_J \otimes C_J(n) = D_0 \otimes I + \sum_{J \in \aleph} D_J \otimes C. \tag{3.8}$$

Remark 3.1 Before proceeding, we like to discuss the irreducibility of the matrix A and the Markov chain $\{(v_g(t), I_a(t), I_s(t)), t \geq 0\}$, since the steady state analysis of the queueing system depends largely on the assumption that A and $\{(v_g(t), I_a(t), I_s(t)), t \geq 0\}$ are irreducible. Unfortunately, neither A nor $\{(v_g(t), I_a(t), I_s(t)), t \geq 0\}$ is guaranteed to be irreducible under the irreducibility assumptions on the matrices D and C made in Section 2. For instance, let

$$n_{\max} = \max\{t : C_J(t) \neq 0, \text{ for some } J \in \aleph\}. \tag{3.9}$$

It is possible that $n_{\max} = \infty$. If $n_{\max} = 1$, then these states with $v_g(t) \geq 2$ are not reachable from states with $v_g(t) \leq 1$. Furthermore, it is possible for A and $\{(v_g(t), I_a(t), I_s(t)), t \geq 0\}$ to have several closed classes of states. For example, if

$$D_0 = \begin{pmatrix} 0.1 & 0 \\ 0 & 0 \end{pmatrix}, \quad D - D_0 = \begin{pmatrix} 0 & 0.9 \\ 1 & 0 \end{pmatrix}, \quad C = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \tag{3.10}$$

for which the matrix D is irreducible and aperiodic and C is irreducible, the matrix A has two closed subsets $\{(1, 1), (2, 2)\}$ and $\{(1, 2), (2, 1)\}$. To analyze the Markov chain, we must identify the closed classes of states and remove transient states. Then we concentrate on the irreducible subsets. Therefore, in this paper, we shall assume that all transient states have been removed and the Markov chain $\{(v_g(t), I_a(t), I_s(t)), t \geq 0\}$ is irreducible and aperiodic.

Denote by $\chi(z)$ the Perron-Frobenius eigenvalue of the nonnegative matrix $A^*(z)$. Let $\mathbf{u}(z)$ and $\mathbf{v}(z)$ be the left and right eigenvectors corresponding to $\chi(z)$, respectively. 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, then $A^*(z)$ is irreducible and all the elements of the vectors $\mathbf{u}(z)$ and $\mathbf{v}(z)$ are positive for $z>0$. According to Neuts [17], the function $\chi(z)$ and the vectors $\mathbf{u}(z)$ and $\mathbf{v}(z)$ can be chosen as differentiable functions.

Lemma 3.1 Assume that A is irreducible. The vector $\theta_a \otimes \theta_s$ is the invariant probability vector of A. At $z=1$, we have $\chi(1) = 1$ and $(\theta_a \otimes \theta_s)\Sigma_n n A_n \mathbf{e} = \chi^{(1)}(1) = \rho$. 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. From the expression of A in equation (3.8), it is easy to verify that the vector $\theta_a \otimes \theta_s$ is the invariant probability vector of the stochastic matrix A. It is also easy to see $\chi(1) = 1$. Furthermore, we have $\mathbf{u}(1) = \theta_a \otimes \theta_s$ and $\mathbf{v}(1) = \mathbf{e}$. By $\mathbf{u}(z)\mathbf{e} = 1$, we have $\mathbf{u}^{(1)}(z)\mathbf{e} = 0$. 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., $\chi^{(1)}(1) = (\theta_a \otimes \theta_s)\left(\sum_{n=0}^{\infty} nA_n\right)\mathbf{e}$, which can be evaluated as

$$
\begin{aligned}
&(\theta_a \otimes \theta_s)\left(\sum_{n=0}^{\infty} nA_n\right)\mathbf{e} = (\theta_a \otimes \theta_s)\left(\sum_{n=1}^{\infty} n\sum_{J\in\aleph} D_J \otimes C_J(n)\right)\mathbf{e} \\
&= (\theta_a \otimes \theta_s)\left(\sum_{J\in\aleph} D_J \otimes \left(\sum_{n=1}^{\infty} nC_J(n)\right)\right)\mathbf{e} = \sum_{J\in\aleph}(\theta_a D_J\mathbf{e})\left(\theta_s \sum_{n=1}^{\infty} nC_J(n)\mathbf{e}\right) \qquad (3.11)\\
&= \sum_{J\in\aleph} \lambda_J E_{\theta_s,J}(s_J) = \sum_{J\in\aleph} \lambda_J / \mu_J = \rho.
\end{aligned}
$$

Therefore, $\chi^{(1)}(1) = \rho$. Then $\chi^{(1)}(1) < 1$ if and only if $\rho < 1$. This completes the proof of Lemma 3.1.

Denote by G an $(m_a m_s)\times(m_a m_s)$ matrix that is the minimal nonnegative solution to

$$
G = \sum_{n=0}^{\infty} A_n G^n = D_0 \otimes I + \sum_{J\in\aleph}\left(\sum_{n=1}^{\infty}(D_J \otimes C_J(n))G^n\right). \qquad (3.12)
$$

We refer to Latouche and Ramaswami [15] and Neuts [17, 19] for more about the matrix G. Let **g** be the (unique) invariant probability vector of G if G is stochastic.

Theorem 3.2 If the Markov chain $\{(v_g(t), I_a(t), I_s(t)), t\geq 0\}$ is irreducible, it is positive recurrent if and only if $\rho < 1$.

Proof. It is easy to see from the structure of the transition matrix P_g (see equation (3.4)) that A must be irreducible if the Markov chain is irreducible. Therefore, by Lemma 3.1, the vector $\theta_a \otimes \theta_s$ is the invariant probability vector of A.

We use the mean-drift method to prove the theorem. First, we prove the necessity of the condition $\rho < 1$. Note that the level of states of the Markov chain is determined by the value of $v_g(t)$. We consider the levels with positive $v_g(t)$. The transitions among these levels are the same as a classical $M/G/1$ type Markov chain. If the Markov chain $\{(v_g(t), I_a(t), I_s(t)), t\geq 0\}$ is positive recurrent, then the mean first passage time from any positive level to the level zero must be finite. By Neuts [17], we must have $\chi^{(1)}(1) < 1$. Therefore, we must have $\rho < 1$. For completeness, we give the following proof based on the mean drift method.

Let **v** be the (column vector) mean first passage time from the level n to the level n–1, for $n\geq 1$. The vector **v** is independent of the level n (>0) because of the $M/G/1$ structure. Since the Markov chain is positive recurrent, every element of **v** is positive and finite. By conditioning on the first transition, the following equation can be established for **v**:

$$
\mathbf{v} = \mathbf{e} + \sum_{n=1}^{\infty} A_n\left(\sum_{i=0}^{n-1} G^i\mathbf{v}\right) = \mathbf{e} + \left(\sum_{n=1}^{\infty} A_n \sum_{i=0}^{n-1} G^i\right)\mathbf{v}. \qquad (3.13)
$$

By equation (3.13), the Perron-Frobenius eigenvalue of the matrix $\sum_{n=1}^{\infty} A_n \sum_{i=0}^{n-1} G^i$ must be less than one. Since the Markov chain is positive recurrent, G is a stochastic matrix. It can be shown that $I–G+\mathbf{eg}$ is invertible. By routine calculations, we obtain

$$\sum_{n=1}^{\infty} A_n \sum_{i=0}^{n-1} G^i = (A-G)(I-G+\mathbf{eg})^{-1} + \left(\sum_{n=1}^{\infty} nA_n\right)\mathbf{eg}. \tag{3.14}$$

Premultiplying $\theta_a \otimes \theta_s$ on both sides of equation (3.14), we obtain

$$(\theta_a \otimes \theta_s)\left(\sum_{n=1}^{\infty} A_n \sum_{i=0}^{n-1} G^i\right) = \theta_a \otimes \theta_s + \left((\theta_a \otimes \theta_s)\sum_{n=1}^{\infty} nA_n\mathbf{e} - 1\right)\mathbf{g}. \tag{3.15}$$

Postmultiplying the right eigenvector corresponding to the Perron-Frobenius eigenvalue of $\sum_{n=1}^{\infty} A_n \sum_{i=0}^{n-1} G^i$ on both sides of equation (3.15), we find that $(\theta_a \otimes \theta_s)\Sigma_n nA_n\mathbf{e} < 1$, which implies $\chi^{(1)}(1) < 1$. Therefore, by Lemma 3.1, we must have $\rho < 1$.

Now, we assume $\rho < 1$. By Lemma 3.1, we have $\chi^{(1)}(1) < 1$. Thus, we have $\chi(z) < z$ for some $z > 1$ and close to one. Then we have $A^*(z)\mathbf{v}(z) = \chi(z)\mathbf{v}(z) < z\mathbf{v}(z)$. Since the Perron-Frobenius eigenvalue of A_0 is less than one (because the Perron-Frobenius eigenvalue of D_0 is less than one), the matrix $I–A_0$ is invertible. Every element of the vector $(I–A_0)^{-1}\mathbf{v}(z)$ is finite and positive. With z and $\mathbf{v}(z)$, we define a (vector) Lyapunov function (test-function) as follows: for states in the level n,

$$\mathbf{f}(n) = \begin{cases} z^n\mathbf{v}(z), & n \geq 0; \\ -n(I-A_0)^{-1}\mathbf{v}(z), & n < 0. \end{cases} \tag{3.16}$$

(Note: Here we abuse the notation a bit; we use "n" to represent all states in the level n.) Since $A^*(z)$ is irreducible, every element of the vector $\mathbf{v}(z)$ is positive. Therefore, $\mathbf{f}(n) \to \infty$, $|n| \to \infty$. Let $\varepsilon_1 = z–\chi(z)$ (>0) and $\varepsilon_2 = \varepsilon_1 \min\{1, \min\{(\mathbf{v}(z))_i\}\}$. We choose n_0 large enough so that, for $n>n_0$,

$$(A^*(z)-A_0)\mathbf{v}(z) + A_0(I-A_0)^{-1}\mathbf{v}(z) - n\mathbf{v}(z) \leq -\varepsilon_1\mathbf{v}(z). \tag{3.17}$$

The mean-drift to level zero is calculated as follows. For $n\geq 1$,

$$\begin{aligned} &E\left(\mathbf{f}(a_g(t+1)) - \mathbf{f}(a_g(t)) \mid a_g(t) = n, I_a(t), I_s(t)\right) \\ &= z^{n-1}\left(A^*(z)\mathbf{v}(z) - z\mathbf{v}(z)\right) \leq -z^{n-1}\varepsilon_1\mathbf{v}(z) \leq -\varepsilon_1\mathbf{v}(z) \leq -\varepsilon_2\mathbf{e}. \end{aligned} \tag{3.18}$$

For $–n < –n_0$, we have

$$
\begin{aligned}
&E\big(\mathbf{f}(a_g(t+1)) - \mathbf{f}(a_g(t)) \mid a_g(t) = -n,\, I_a(t),\, I_s(t)\big) \\
&= (A^*(z) - A_0)\mathbf{v}(z) + (n+1)A_0(I - A_0)^{-1}\mathbf{v}(z) - n(I - A_0)^{-1}\mathbf{v}(z) \\
&= (A^*(z) - A_0)\mathbf{v}(z) + A_0(I - A_0)^{-1}\mathbf{v}(z) - n\mathbf{v}(z) \\
&\le -\varepsilon_1\mathbf{v}(z) \le -\varepsilon_2\mathbf{e}.
\end{aligned} \tag{3.19}
$$

We also have, for $-n_0 \le n \le 0$,

$$
\begin{aligned}
&E\big(\mathbf{f}(a_g(t+1)) \mid a_g(t) = -n,\, I_a(t),\, I_s(t)\big) \\
&= (A^*(z) - A_0)\mathbf{v}(z) + (n+1)A_0(I - A_0)^{-1}\mathbf{v}(z) \\
&\le A^*(z)\mathbf{v}(z) + (n_0+1)A_0(I - A_0)^{-1}\mathbf{v}(z) < \infty.
\end{aligned} \tag{3.20}
$$

Thus, the mean-drifts away from level zero, with respect to the Lyapunov function given in equation (3.16), are all less than $-\varepsilon_2$, except for a finite number of states. Therefore, the Markov process is positive recurrent by Foster's criterion (see Theorem 2.2.3 in Fayolle, et al. [5]). This completes the proof of Theorem 3.2.

4. Steady State Distributions

In this section, we consider the stationary distribution of $\{(v_g(t), I_a(t), I_s(t)), t\ge 0\}$ under $\rho < 1$ and the irreducibility conditions made in Theorem 3.2. Denote by $\pi = (\ldots, \pi(-1), \pi(0), \pi(1), \ldots)$ the steady state distribution of $\{(v_g(t), I_a(t), I_s(t)), t\ge 0\}$, i.e., $\pi P_g = \pi$, $\pi\mathbf{e} = 1$, where $\pi(n) = (\ldots, \pi_{i,j}(n), \ldots)$ is a row vector of the size $m_a m_s$, and

$$
\pi_{i,j}(n) = \lim_{t\to\infty} P\{v_g(t) = n,\, I_a(t) = i,\, I_s(t) = j \mid v_g(0), I_a(0), I_s(0)\} \tag{4.1}
$$

We further assume that the Markov chain $\{(v_g(t), I_a(t), I_s(t)), t\ge 0\}$ is aperiodic so that the steady state distribution (invariant probability measure) π is unique. First, we find $\{\pi(0), \pi(-1), \pi(-2), \ldots\}$ in terms of $\pi(0)$. By $\pi P_g = \pi$, we have $\pi(-n-1) = \pi(-n)A_0$, for $n\ge 0$, which leads to

$$
\pi(-n) = \pi(0)(A_0)^n \quad \text{for} \quad n\ge 0. \tag{4.2}
$$

Thus, to find π, we only need to concentrate on $\{\pi(0), \pi(1), \pi(2), \ldots\}$. In fact, we can construct another Markov chain, which is related to the (actual) workload process, to find $\{\pi(0), \pi(1), \pi(2), \ldots\}$. The construction of that Markov chain is given as follows. By $\pi P_g = \pi$, we have

$$
\pi(n) = \pi(n+1)A_0 + \pi(n)A_1 + \ldots + \pi(1)A_n + \left(\sum_{i=0}^{\infty}\pi(-i)\right)A_n, \quad n \ge 1. \tag{4.3}
$$

Denote by $\pi(-\infty,0] = \sum_{i=0}^{\infty} \pi(-i) = \pi(0)(I - A_0)^{-1}$. Note that the matrix $(I–A_0)^{-1}$ is finite and nonnegative. Then the probability vector $(\pi(-\infty, 0], \pi(1), \pi(2), \ldots)$ is the steady state distribution of the following transition probability matrix

$$P_v = \begin{pmatrix} A_0 & A_1 & A_2 & A_3 & \cdots \\ A_0 & A_1 & A_2 & A_3 & \cdots \\ & A_0 & A_1 & A_2 & \ddots \\ & & \ddots & \ddots & \ddots \end{pmatrix}. \tag{4.4}$$

For the Markov chain $\{(v_g(t), I_a(t), I_s(t)), t\geq 0\}$, if we only observe the process when $v_g(t) \geq 0$, we obtain a process $\{(v(t), I_a(t), I_s(t)), t\geq 0\}$, where $v(t)$ is the (actual) total workload in the system. It is easy to see that $\{(v(t), I_a(t), I_s(t)), t\geq 0\}$ is an $M/G/1$ type Markov chain and its transition probability matrix is given by equation (4.4). Note that, to obtain the steady state waiting time distributions, we can directly consider the Markov chain $\{(v(t), I_a(t), I_s(t)), t\geq 0\}$. By considering the Markov chain $\{(v_g(t), I_a(t), I_s(t)), t\geq 0\}$, however, we get more information about the queueing process, especially about the idle periods.

Denote by $\pi^*(z) = \pi(-\infty, 0] + \sum_{n=1}^{\infty} z^n \pi(n)$. From equation (4.4), it is easy to obtain the following equation for $\pi^*(z)$:

$$\pi^*(z)(zI - A^*(z)) = \pi(-\infty, 0](z-1)A^*(z). \tag{4.5}$$

Theorem 4.1 If the Markov chain $\{(v_g(t), I_a(t), I_s(t)), t\geq 0\}$ is irreducible and aperiodic and $\rho < 1$, then $\pi^*(1) = \theta_a \otimes \theta_s$ and $\pi(-\infty, 0] = (1–\rho)\mathbf{g}$, $\pi(0) = (1–\rho)\mathbf{g}(I–A_0)$, where $\mathbf{g}$ is the invariant probability vector of G.

Proof. Letting z goes to 1 on both sides of equation (4.5), we obtain $\pi^*(1)(I–A) = 0$. Since $\pi^*(1)$ is a probability vector, it must be the invariant probability vector of A. By Lemma 3.1, $\pi^*(1) = \theta_a \otimes \theta_s$.

Dividing both sides of equation (4.5) by z –1 and multiplying both sides by $\mathbf{e}$, yields

$$\pi(-\infty, 0]A^*(z)\mathbf{e} = \pi^*(z)\frac{(zI - A^*(z))\mathbf{e}}{z-1}. \tag{4.6}$$

Letting z go to 1 on both sides of equation (4.6) and using l'Hôpital's Rule, by Lemma 3.1, we obtain

$$\pi(-\infty, 0]\mathbf{e} = \lim_{z\to 1}\pi^*(z)\frac{(zI - A^*(z))\mathbf{e}}{z-1} = \pi^*(1)\left(1 - \sum_{n=0}^{\infty} nA_n\right)\mathbf{e}$$
$$= (\theta_a \otimes \theta_s)\left(1 - \sum_{n=0}^{\infty} nA_n\right)\mathbf{e} = 1 - \chi^{(1)}(1) = 1 - \rho. \tag{4.7}$$

Consider the embedded Markov chain P_0 at the epochs the Markov chain $\{(v(t), I_a(t), I_s(t)), t\geq 0\}$ enters the level zero. Then we must have $\pi(-\infty, 0] = \pi(-\infty, 0]P_0$. By conditioning on the first transition from the level zero, we obtain

$$P_0 = \sum_{n=0}^{\infty} A_n G^n = G. \tag{4.8}$$

Therefore, we must have $\pi(-\infty, 0] = c\mathbf{g}$, where c is a constant. Since $\pi(-\infty, 0]\mathbf{e} = 1-\rho$, then $c = 1-\rho$. This completes the proof of Theorem 4.1.

Once $\pi(-\infty, 0]$ is found, other vectors $\{\pi(1), \pi(2), \ldots\}$ can be computed by a stable recursion developed in Ramaswami [20]: for $n\geq 1$,

$$\pi(n) = \left(\pi(-\infty, 0]\overline{A}_n + \sum_{j=1}^{n-1}\pi(j)\overline{A}_{n-j+1}\right)(I - \overline{A}_1)^{-1}, \tag{4.9}$$

where $\overline{A}_n = \sum_{j=n}^{\infty} A_j G^{j-n}$ for $n\geq 1$.

Remark 4.1 The results obtained in Theorem 4.1 and the above algorithm may be valid for the more general case. For instance, if $n_{\max} = 1$ (defined in equation (3.9)), the Markov chain $\{(v_g(t), I_a(t), I_s(t)), t\geq 0\}$ is reducible. However, if $\pi(0)$ can be obtained, then $\{\pi(1), \pi(2), \ldots\}$ can be computed by using equation (4.9). For this case, $A_n = 0$ for $n\geq 2$. Equation (4.9) gives $\pi(n) = 0$ for $n\geq 2$.

Now, we return to the total workload and the waiting time of an arbitrary batch. Let v_g be the generic random variable for the total workload in the system in steady state, w be the generic random variable for the waiting time of an arbitrary batch (i.e., the waiting time of the first customer in the batch), w_J be the generic random variable for the waiting time of an arbitrary batch J, d be the generic random variable for the sojourn time of an arbitrary batch (the total time the whole batch is in the system), and d_J be the generic random variable for the sojourn time of an arbitrary batch J. We also denote by I_a the generic random variable of the underlying Markov chain of the arrival process in steady state and by I_s the generic random variable of the service process in steady state.

Theorem 4.2 If the queueing system is stable and the Markov chain $\{(v_g(t), I_a(t), I_s(t)), t\geq 0\}$ is irreducible and aperiodic, the distributions of the total workload, waiting times, and sojourn times are given by:

$$P\{v_g = n\} = \pi(n)\mathbf{e}, \quad -\infty < n < \infty. \tag{4.10}$$

$$P\{w = n\} = \begin{cases} \frac{1}{\lambda}\big((1-\rho)\mathbf{g}((D-D_0)\otimes I)\mathbf{e} + \pi(1)((D-D_0)\otimes I)\mathbf{e}\big), & n = 0; \\ \frac{1}{\lambda}\pi(n+1)((D-D_0)\otimes I)\mathbf{e}, & n \geq 1. \end{cases} \tag{4.11}$$

$$P\{w_J = n\} = \begin{cases} \frac{1}{\lambda_J}\big((1-\rho)\mathbf{g}(D_J\otimes I)\mathbf{e} + \pi(1)(D_J\otimes I)\mathbf{e}\big), & n = 0; \\ \frac{1}{\lambda_J}\pi(n+1)(D_J\otimes I)\mathbf{e}, & n \geq 1. \end{cases} \tag{4.12}$$

$$P\{d = n\} = \frac{1}{\lambda}\left((1-\rho)\mathbf{g}((D-D_0)\otimes C(n)) + \sum_{t=1}^{n}\pi(t)((D-D_0)\otimes C(n+1-t))\right)\mathbf{e}, \; n \geq 1. \tag{4.13}$$

$$P\{d_J = n\} = \frac{1}{\lambda_J}\left((1-\rho)\mathbf{g}(D_J\otimes C_J(n)) + \sum_{t=1}^{n}\pi(t)(D_J\otimes C_J(n+1-t))\right)\mathbf{e}, \quad n \geq 1. \tag{4.14}$$

Proof. Equation (4.10) is obvious. Recall that, if $v_g<0$, the queueing system is empty in that period and the system has been idle for $-v_g$ periods of time.

Equation (3.2) shows that the waiting time of a batch is the (generalized) total workload right before its arrival. Therefore, the distribution of waiting time w_J of an arbitrary batch of type J can be obtained by conditioning on the arrival of a batch J at an arbitrary time, which can be obtained as follows, for $n\geq 1$,

$$\begin{aligned} P\{w_J = n\} &= P\{v_g = n+1 \,|\, \text{Batch } J \text{ arrives next}\} \\ &= \frac{P\{v_g = n+1, \text{ Batch } J \text{ arrives next}\}}{P\{\text{Batch } J \text{ arrives next}\}} \\ &= \frac{1}{\lambda_J}\sum_{i=1}^{m_a}\sum_{j=1}^{m_s} P\{v_g = n+1, I_a = i, I_s = j, \text{ Batch } J \text{ arrives next}\} \\ &= \frac{1}{\lambda_J}\sum_{i=1}^{m_a}\sum_{j=1}^{m_s} P\{v_g = n+1, I_a = i, I_s = j\} \\ &\qquad \cdot P\{\text{Batch } J \text{ arrives next} \,|\, v_g = n+1, I_a = i, I_s = j\} \\ &= \frac{1}{\lambda_J}\sum_{i=1}^{m_a}\sum_{j=1}^{m_s} P\{v_g = n+1, I_a = i, I_s = j\} P\{\text{Batch } J \text{ arrives next} \,|\, I_a = i\} \\ &= \frac{1}{\lambda_J}\sum_{i=1}^{m_a}\sum_{j=1}^{m_s} \pi_{i,j}(n+1)(D_J\mathbf{e})_i, \end{aligned} \tag{4.15}$$

which leads to the second expression in equation (4.12). If n=0, we have $\{w_J = 0\} = \{v_g \le 1 \mid$ Batch J batch arrives next$\}$, which leads to the first expression in equation (4.12).

The distribution of w can be obtained similarly. The distributions of the sojourn times are obtained as the convolution of the distributions of waiting time and service time by taking the phase of the service process into consideration. This completes the proof of Theorem 4.2.

Note that the computation of the moments of the waiting times can be done by using existing formulas for the $M/G/1$ type Markov chains (see Neuts [17, 19]). Details are omitted. The computation of the waiting time distribution of an arbitrary type k customer can be a little bit involved, since the service times of customers in a batch may not be independent. In other words, if the service time is defined at the batch level, the waiting time distribution of a type k customer cannot be found explicitly. Nonetheless, if more information on the service process is available, it might be possible to find the distribution of waiting times of individual types of customers. An example is given in Section 5.

To end this section, we point out that there is a close relation between the age process of the batch in service and the total workload process. We define the *age* of a batch in 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 $\{a_g(t), t \ge 0\}$ of the batch in service or to be served next (if the system is empty) is defined as

$$a_g(t) = w_{n(t)} + s_{X_{n(t)}} - \tau_{n(t)+1} + t - t_{n(t)}, \tag{4.16}$$

where $n(t)$ is the ordinal number of the last batch served in or before period t and $t_{n(t)}$ is the departure period of the $n(t)$-th batch. The values of $n(t)$, $w_{n(t)}$, and $t_{n(t)}$ are updated if a departure occurs in period t, where $w_{n(t)}$ can be computed by using equation (3.1). According to Corollary 4.2 in HE [11], $a_g(t)$ and $v_g(t)$ have the same steady-state distribution. Thus, equation (4.10) can be used for computing the steady state distribution of the age of the batch in service at an arbitrary time as well.

5. The Discrete Time *MMAP*[*K*]/*GI*[*K*]/1 Queue

For the discrete time *MMAP*[*K*]/*GI*[*K*]/1 queue, the service times of individual customers are independent with general distributions, i.e., $m_s = 1$. Consequently, the service times of batches are independent as well. Denote by s_k the service time of a type k customer. Then the service rate of type k customers is given by $\mu_{(k)} = 1/(E(s_k))$. The service time of a batch $J = j_1 j_2 \cdots j_n$ is $s_J = \sum_{i=1}^{n} s_{j_i}$ with mean $E(s_J) = \sum_{i=1}^{n} E(s_{j_i})$. By definition (2.6), we have

$$\begin{aligned}\rho &= \sum_{J\in\aleph}\lambda_J/\mu_J = \sum_{J\in\aleph}\lambda_J\sum_{i=1}^{|J|}E(s_{j_i}) = \sum_{J\in\aleph}\lambda_J\sum_{k=1}^{K}N(J,k)E(s_k)\\ &= \sum_{k=1}^{K}\sum_{J\in\aleph}\lambda_J N(J,k)E(s_k) = \sum_{k=1}^{K}\lambda_{(k)}/\mu_{(k)}.\end{aligned} \tag{5.1}$$

Thus, for the *MMAP*[*K*]/*GI*[*K*]/1 queue, definition (2.6) for the traffic intensity is consistent with the classical definition.

The distribution of the total workload and the waiting time of batches can be calculated by using formulas (4.10) and (4.11). For this special case, we are able to compute the waiting times of individual types of customers. Let $w_{(k)}$ be the generic random variable for the waiting time of an arbitrary type k customer. Based on equation (4.11), we have, for $n\geq 0$ and $1\leq k\leq K$,

$$\begin{aligned}P\{w_{(k)}=n\} &= \frac{1}{\lambda_{(k)}}\big((1-\rho)\mathbf{g}+\pi(1)\big)\sum_{J\in\aleph}\sum_{t=1}^{|J|}\big(D_J\otimes I\big)\mathbf{e}P\{s_{j_1}+s_{j_2}+\ldots+s_{j_{t-1}}=n\}\delta_{\{j_t=k\}}\\ &\quad+\frac{1}{\lambda_{(k)}}\sum_{l=2}^{n}\sum_{J\in\aleph}\sum_{t=1}^{|J|}\pi(l)\big(D_J\otimes I\big)\mathbf{e}P\{s_{j_1}+s_{j_2}+\ldots+s_{j_{t-1}}=n+1-l\}\delta_{\{j_t=k\}}\\ &= \sum_{J\in\aleph}\frac{\lambda_J}{\lambda_{(k)}}\sum_{t=1}^{|J|}P\{w_J+s_{j_1}+s_{j_2}+\ldots+s_{j_{t-1}}=n\}\delta_{\{j_t=k\}},\end{aligned} \tag{5.2}$$

where $\delta_{\{.\}}$ is the indicator function. Let $d_{(k)}$ be the generic random variable for the sojourn time of an arbitrary type k customer. Note that, for the definition of $d_{(k)}$, we assume that a customer leaves the system immediately after its service. We have, for $n\geq 1$ and $1\leq k\leq K$,

$$P\{d_{(k)}=n\} = \sum_{J\in\aleph}\frac{\lambda_J}{\lambda_{(k)}}\sum_{t=1}^{|J|}P\{w_J+s_{j_1}+s_{j_2}+\ldots+s_{j_t}=n\}\delta_{\{j_t=k\}}. \tag{5.3}$$

An important special case is the discrete time *MMAP*[*K*]/*PH*[*K*]/1 queue, where the service times of individual customers are independent and have PH-distributions with matrix representations $\{(m_k, \alpha_k, T_k), 1\leq k\leq K\}$, where m_k is the number of phases of the PH-distribution, α_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., $\alpha_k\mathbf{e} = 1$, $1\leq k\leq K$). See Neuts [17] for more about PH-distribution. Denote by $\mathbf{T}_k^0 = (I-T_k)\mathbf{e}$. Since PH-distributions are closed under convolution, the service time of a batch J also has a discrete time PH-distribution with a matrix representation $\{m_J, \alpha_J, T_J\}$, where, for $J = j_1 j_2 \ldots j_n$,

$$
\begin{aligned}
&m_J = \sum_{i=1}^{n} m_{j_i};\\
&\alpha_J = (\alpha_{j_1}, 0, ...,0);\\
&T_J = \begin{pmatrix}
T_{j_1} & \mathbf{T}^0_{j_1}\alpha_{j_2} & & & \\
& T_{j_2} & \mathbf{T}^0_{j_2}\alpha_{j_3} & & \\
& & \ddots & \ddots & \\
& & & T_{j_{n-1}} & \mathbf{T}^0_{j_{n-1}}\alpha_{j_n}\\
& & & & T_{j_n}
\end{pmatrix}, \qquad \mathbf{T}^0_J = \begin{pmatrix} 0\\ \vdots\\ \vdots\\ 0\\ \mathbf{T}^0_{j_n}\end{pmatrix}.
\end{aligned} \tag{5.4}
$$

According to Neuts [17], the distributions of service times are given by

$$C_J(t) = \alpha_J T_J^{t-1}\mathbf{T}^0_J, \quad t \geq 1,\ J \in \aleph. \tag{5.5}$$

The matrix blocks of the transition matrix of $\{(v_g(t), I_a(t), I_s(t)), t\geq 0\}$ (see equation (3.5)) can be computed easily by

$$A_0 = D_0, \quad A_n = \sum_{J\in\aleph}\left(\alpha_J T_J^{n-1}\mathbf{T}^0_J\right)D_J, \ \ n \geq 1. \tag{5.6}$$

The steady state distributions of waiting times can be computed accordingly by using formulas provided in Sections 3 and 4. Note that for this case, $A = D$. So A is irreducible if and only if D is irreducible.

We point out that for the discrete time *MMAP*[*K*]/*PH*[*K*]/1 queue, since both the arrival process and the service process are governed by Markov chains, a QBD process can be introduced to describe the total workload process and the age process of the batch in service. That approach was used in HE [11] and Van Houdt and Blondia [28]. The advantage of the QBD approach is that the steady state distribution of the QBD process has a matrix geometric distribution, which is easy to compute. On the other hand, as was pointed out by Van Houdt and Blondia [28], the dimension of the matrix blocks involved can be quite large for the QBD approach, which is

$$m_a\left(1+\sum_{J\in\aleph}\sum_{i=1}^{|J|} m_{j_i}\right). \tag{5.7}$$

That number can be significantly larger than m_a - the dimension of the blocks for the *M*/*G*/1 type Markov chain developed in this paper. Even if we consider the computation of PH-distributions used in constructing matrices $\{A_n, n\geq 0\}$ (see equation (5.6)), the largest dimension of matrices involved with the *M*/*G*/1 type Markov chain is $\max\{m_a, m_J, J\in\aleph\}$, which can be significantly smaller than the number in equation (5.7). Therefore, the *M*/*G*/1 approach can play an important role in the analysis of the waiting times and sojourn times for the *MMAP*[*K*]/*PH*[*K*]/1 case. More discussion on this issue is given in Section 6.

6. Numerical Examples

In this section, we outline a scheme for computing distributions of waiting times and sojourn times and briefly discuss three numerical examples. First, we summarize the steps for computing performance measures in the following examples.

1) Compute vectors θ_a and θ_s, the arrival rates $\{\lambda_J, J\in\aleph\}$, arrival rates $\{\mu_J, J\in\aleph\}$, and ρ (by equation (2.6)).
2) Compute transition blocks $\{A_0, A_1, A_2, \ldots\}$ by equation (3.5).
3) Compute the matrix G by equation (3.12) and the vector $\mathbf{g}$.
4) Compute the vector $\pi(-\infty, 0]$ and $\pi(0)$ by Theorem 4.1.
5) Compute vectors $\{\pi(0), \pi(-1), \pi(-2), \ldots\}$ by equation (4.2).
6) Compute vectors $\{\pi(1), \pi(2), \ldots\}$ by the recursion given in equation (4.9).
7) Determine distributions of the total workload, waiting times, and sojourn times by formulas given in Corollary 4.2 or equations (5.2) and (5.3).

Most of the above steps can be done in a straightforward manner except the computation of the matrix G and vectors $\{\pi(1), \pi(2), \ldots\}$. Nonetheless, for the computation of G and $\{\pi(1), \pi(2), \ldots\}$, efficient algorithms are available in the literature (Neuts [19] and Ramaswami [20]).

Example 6.1 Consider an *MMAP*[2]/*PH*[2]/1 queueing with following system parameters:

$$\begin{aligned} &K = 2, \quad m_a = 2, \quad D_0 = \begin{pmatrix} 0.6 & 0.0 \\ 0.4 & 0.1 \end{pmatrix}, \quad D_1 = \begin{pmatrix} 0.0 & 0.1 \\ 0.0 & 0.0 \end{pmatrix}, \quad D_2 = \begin{pmatrix} 0.3 & 0.0 \\ 0.4 & 0.1 \end{pmatrix}; \\ &m_1 = 1, \quad \alpha_1 = (1), \quad T_1 = (0.4); \quad m_2 = 2, \quad \alpha_2 = (0.1, 0.9), \quad T_2 = \begin{pmatrix} 0.1 & 0.1 \\ 0.1 & 0.0 \end{pmatrix}. \end{aligned} \tag{6.1}$$

This queueing system has a light traffic with $\rho = 0.5139$. For this queueing system, a type 1 customer is mostly likely to be followed by type 2 customers. Since type 1 customers have a slower service rate, it is expected that type 2 customers face longer waiting time than type 1 customers, which is shown in Table 6.1. The distributions of waiting times and sojourn times are given in Table 6.1.

Table 6.1 Distributions of waiting times and sojourn times

n	0	1	2	3	4	5	6	7	…
$P\{v=n\}$	0.4861	0.3340	0.1025	0.0431	0.0186	0.0083	0.0038	0.0017	…
$P\{w=n\}$	0.8125	0.1063	0.0451	0.0196	0.0088	0.0040	0.0018	0.0008	…
$P\{w_1=n\}$	0.8557	0.0849	0.0338	0.0140	0.0062	0.0028	0.0012	0.0006	…
$P\{w_2=n\}$	0.8006	0.1122	0.0482	0.0211	0.0095	0.0043	0.0020	0.0009	…
$P\{d=n\}$	0	0.6695	0.1895	0.0792	0.0338	0.0151	0.0068	0.0031	…
$P\{d_1=n\}$	0	0.5134	0.2563	0.1228	0.0575	0.0267	0.0123	0.0057	…
$P\{d_2=n\}$	0	0.7125	0.1711	0.0671	0.0272	0.0119	0.0053	0.0024	…

It is interesting to see that $P\{v=0\} = 0.4861 = 1 - \rho$, which is much smaller than $P\{w=0\}=0.8125$. In fact, for many numerical examples, $P\{v=0\}+P\{v=1\} \approx P\{w=0\}$. The reason is that, for discrete queues, an arrival and a departure can occur simultaneously. Therefore, if the total workload is one, the waiting time of the next arrival is actually zero. Next, we examine a queueing system in heavy traffic.

Example 6.2 Consider an *MMAP*[5]/*PH*[5]/1 queue with following system parameters: $K = 5$, $m_a = 3$,

$$
\begin{aligned}
&D_0 = \begin{pmatrix} 0.1 & 0.1 & 0 \\ 0.2 & 0.3 & 0.1 \\ 0.1 & 0 & 0.3 \end{pmatrix}, \quad D_1 = \begin{pmatrix} 0.1 & 0.4 & 0 \\ 0.1 & 0.2 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad D_2 = \begin{pmatrix} 0 & 0.1 & 0 \\ 0 & 0.1 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \\
&D_3 = \begin{pmatrix} 0.1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0.2 \end{pmatrix}, \quad D_4 = \begin{pmatrix} 0.1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0.1 \end{pmatrix}, \quad D_5 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0.3 \end{pmatrix},
\end{aligned} \tag{6.2}
$$

$$
\begin{aligned}
&m_i = 1, \quad \alpha_i = (1), \quad T_i = (0.1), \quad i = 1, 2; \\
&m_3 = 2, \quad \alpha_3 = (0.1, 0.9), \quad T_3 = \begin{pmatrix} 0.7 & 0.1 \\ 0.4 & 0 \end{pmatrix}; \\
&m_i = 2, \quad \alpha_i = (0, 1), \quad T_i = \begin{pmatrix} 0.7 & 0.1 \\ 0.2 & 0 \end{pmatrix}, \quad i = 4, 5.
\end{aligned} \tag{6.3}
$$

The traffic intensity of this queueing system is $\rho = 0.9304$. The distributions of waiting times are given in Table 6.2.

Table 6.2 Distributions of waiting times

n	0	1	2	3	4	5	6	7	…
$P\{w=n\}$	0.1585	0.0348	0.0286	0.0263	0.0250	0.0240	0.0231	0.0222	…
$P\{w_1=n\}$	0.1852	0.0404	0.0294	0.0262	0.0246	0.0235	0.0224	0.0215	…
$P\{w_2=n\}$	0.1864	0.0414	0.0294	0.0261	0.0245	0.0234	0.0224	0.0214	…
$P\{w_3=n\}$	0.1333	0.0293	0.0278	0.0264	0.0254	0.0245	0.0237	0.0229	…
$P\{w_4=n\}$	0.1427	0.0308	0.0280	0.0264	0.0253	0.0244	0.0235	0.0227	…
$P\{w_5=n\}$	0.1177	0.0268	0.0273	0.0263	0.0255	0.0247	0.0240	0.0232	…

Some observations can be made from Table 6.2. First, compared to Example 6.1, the distributions have a heavier tail. That is, the waiting times are much longer than that of Example 6.1. Second, the waiting times of the five types of customers are significantly different. For example, $P\{w_5=0\}=0.1177$, which significantly smaller than $P\{w_1=0\}=0.1852$. Thus, in a queueing system, the waiting times of different types of customers can be dramatically different. Similar conclusions can be drawn for the sojourn times. For this example, the size of the matrix blocks in computation is 3. If the QBD approach or the *GI*/*M*/1 approach (HE [11] and Van Houdt and Blondia [28]) is used, the size of the matrix blocks in computation is 27 or 24, which is significantly larger.

Example 6.3 Consider an *MMAP*[2]/*PH*[2]/1 queueing with three types of batches $J_1 = 1$, $J_2 = 2$, and $J_3 = 12$ and following system parameters: $K = 2$, $m_a = 2$,

$$D_0 = \begin{pmatrix} 0.5 & 0 \\ 0.2 & 0.5 \end{pmatrix}, \quad D_1 = \begin{pmatrix} 0 & 0.3 \\ 0 & 0.1 \end{pmatrix}, \quad D_2 = \begin{pmatrix} 0.2 & 0 \\ 0 & 0 \end{pmatrix}, \quad D_{12} = \begin{pmatrix} 0 & 0 \\ 0 & 0.2 \end{pmatrix},$$
$$m_1 = 1, \quad \alpha_1 = (1), \quad T_1 = (0.4); \quad m_2 = 2, \quad \alpha_2 = (0.9, 0.1), \quad T_2 = \begin{pmatrix} 0.1 & 0 \\ 0.1 & 0.7 \end{pmatrix}. \tag{6.4}$$

The traffic intensity of this queueing system is $\rho = 0.774$. By using equation (5.2), the distributions of waiting times of the two types of customers can be computed and the results are given in Table 6.3.

Table 6.3 Distributions of waiting times

n	0	1	2	3	4	5	6	7	…
$P\{w_1=n\}$	0.4422	0.1051	0.0804	0.0643	0.0522	0.0428	0.0352	0.0292	…
$P\{w_2=n\}$	0.5176	0.0915	0.0682	0.0548	0.0448	0.0369	0.0306	0.0254	…
$P\{w_{12}=n\}$	0.2914	0.1322	0.1048	0.0833	0.0670	0.0544	0.0446	0.0367	…
$P\{w_{(1)}=n\}$	0.3819	0.1159	0.0902	0.0719	0.0581	0.0474	0.0390	0.0322	…
$P\{w_{(2)}=n\}$	0.2070	0.1415	0.1168	0.0955	0.0773	0.0626	0.0510	0.0417	…

Note that the w_J is the waiting time of an arbitrary batch J and $w_{(i)}$ is the waiting time of an arbitrary type i customer, $i = 1, 2$. Table 6.3 shows that the waiting times of type 2 customers

are significantly longer than that of type 1 customers, because many type 2 customers are served after a type 1 customer. Thus, the order of service in a batch can have significant effect on the queueing processes of different types of customers.

Acknowledgement: The author would like to thank K-C Wang Foundation and Chinese Academy of Science for their support on this research project. The author would like to thank Dr. Blake for proofreading the paper. The author would also like to thank two anonymous referees for their valuable comments and suggestions. This research was partially supported by a NSERC research grant.

References

[1] Alfa, S.A. (2002), Discrete time queues and matrix-analytic methods, *TOP*, **Vol 10**, No. 2, 147-210.

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

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

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

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

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

[7] Gantmacher, F.R. (1959), *The theory of matrices*, Chelsea, New York.

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

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

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

[11] HE, Qi-Ming (2003), Age process, total workload, sojourn times, and waiting times in a discrete time *SM*[*K*]/*PH*[*K*]/1/FCFS queue (submitted).

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

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

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

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

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

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

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

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

[20] Ramaswami, V. (1988), Stable recursion for the steady state vector in Markov chains of *M*/*G*/1 type, *Stochastic Models*, **Vol 4**, 183-188.

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

[22] Takine, T. (1996), A continuous version of matrix-analytic methods with the skip-free to the left property, *Stochastic Models*, **Vol 12**, No. 4, 673-682.

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

[24] Takine, T. (2001), A recent progress in algorithmic analysis of FIFO queues with Markovian arrival streams, *J. Korean Math. Soc*., **Vol 38**, No. 4, 807-842.

[25] Takine, T. and T. Hasegawa (1994), The workload in a *MAP*/*G*/1 queue with state-dependent services: its applications to a queue with preemptive resume priority, *Stochastic Models*, **Vol 10**, No. 1, 183-204.

[26] Van Houdt, B. and C. Blondia (2002), The delay distribution of a type *k* customer in a FCFS *MMAP*[*K*]/*PH*[*K*]/1 queue, *Journal of Applied Probability*, **Vol 39**, No. 1, 213-222.

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

[28] Van Houdt, B. and C. Blondia (2004), 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*, **Vol 20**, No. 1, 55-69.

[29] Zhao, Y.Q, W. Li, and W. J. braun (2001), 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.

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