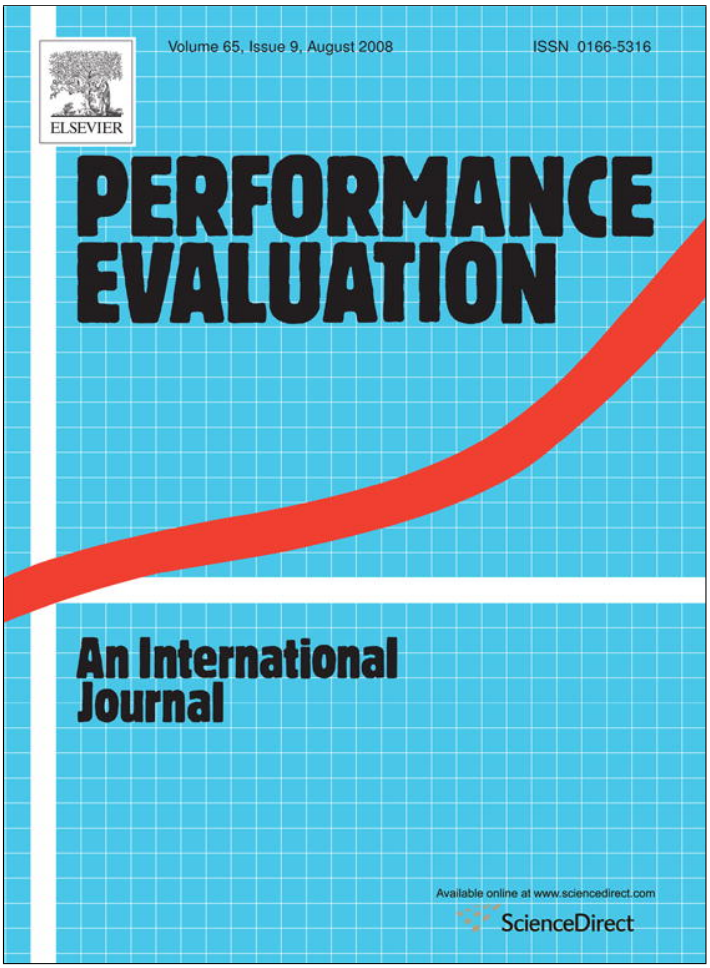

Available online at www.sciencedirect.com
ScienceDirect
Performance Evaluation 65 (2008) 623–640
PERFORMANCE EVALUATION
An International Journal
www.elsevier.com/locate/peva

Algorithmic analysis of the discrete time $GI^X/G^Y/1$ queueing system

Attahiru S. Alfa[a,*], Qi-Ming He[b]

[a] *Department of Electrical & Computer Engineering, University of Manitoba, Winnipeg, Manitoba, Canada R3T 5V6*
[b] *Department of Industrial Engineering, Dalhousie University, Halifax, Nova Scotia, Canada B3J 2X4*

Received 1 April 2006; received in revised form 27 August 2007; accepted 8 February 2008
Available online 14 February 2008

Abstract

In this paper, we study the discrete time $GI^X/G^Y/1$ queueing system. First, some general results are obtained for the stability condition, stationary distributions of the queue lengths and waiting times. Then we show that, for some practical situations, the queueing system of interest has the properties of both the $M/G/1$ and $GI/M/1$ non-skip-free Markov chains described by [H.R. Gail, S.L. Hantler, B.A. Taylor, Non-skip-free $M/G/1$ and $GI/M/1$ types of Markov chains, Advances in Applied Probability, 29 (3) (1997) 733–758]. For such cases, the queueing system can be analyzed as a QBD process, a $M/G/1$ type Markov chain, or a $GI/M/1$ type Markov chain with finite blocks after re-blocking. Computational procedures are developed for computing a number of performance measures for the queueing system of interest. In addition, we study a $GI/M/1$ type Markov chain associated with the age process of the customers in service.

Keywords: Queueing theory; Matrix-analytic methods; Markov chain

1. Introduction

Single server queues with batch arrivals and batch services are general and have several applications in telecommunications, manufacturing, and computer systems. In modern wireless communication systems, especially when dealing with multimedia type of data, requests arrive in batches of varying sizes and services are provided in varying batches. As an example, in the wireless communication systems of third generation (3G) and beyond, that use CDMA (Code Division Multiple Access) technology, there is continued interest to maximize the throughput of the system by implementing adaptive modulation and coding techniques. If, for CDMA technology, we consider the forward link transmission where data for a mobile usually arrive at the base station in batches (e.g. from a website), the data are usually sent from the base station to the appropriate mobile station in batches whenever that station is scheduled to receive data. The arriving batches of data are random and the batch sizes transmitted are random depending on the modulation and coding structure selected which both depend on the channel conditions and several other physical layer parameters. Such a situation can be seen as a typical example of the $GI^X/G^Y/1$ system. In some

* Corresponding author.
E-mail addresses: alfa@ee.umanitoba.ca, alfa@cc.umanitoba.ca (A.S. Alfa), Qi-Ming.He@DAL.CA (Q.-M. He).

0166-5316/$ - see front matter
doi:10.1016/j.peva.2008.02.001

manufacturing systems jobs are usually processed in batches, and these batches are forwarded to another station where they are re-batched for another batch processing. In an intermediate station one may consider the queueing system as batch arrival and batch service types. One may argue that such a system is more of the $G^X/G^Y/1$ type. Nevertheless the $GI^X/G^Y/1$ is a special case of this. In the area of transportation we have situations in which less-than-truck-loads arrive at a major trucking terminal to be consolidated for shipment to another intermediate main terminal where this new load is now broken into smaller packages for delivery to individual destinations. Arrivals of loads or packages at the terminals are in batches and service is provided in batches of different sizes.

There are several examples of the $GI^X/G^Y/1$ system in most real life situations and yet limited results are available except when interarrival or service times are exponential. It is surprising that little attention has been paid to this class of queues. This is perhaps because it is difficult to analyze. The difficulty lies in the fact that customers arrive in batches and are served in batches, and the problem of the analysis is further compounded especially when it comes to analyzing the waiting times of customers. A customer, selected at random, arrives in a batch which is of random size and his service may be carried out in a different batch from its arriving batch, with the service batch also being random. We believe this aspect to be the difficulty associated with analyzing this class of queues. An earlier work that closely relates to the $GI^X/G^Y/1$ is the work of [9], in which he studied the $GI^X/G^D/1$, where D implies a fixed batch size for service. He also included in this model the case in which the first service times after an idle period is of different duration from the other service times — usually longer. For this system he was able to obtain the bounds for the expected waiting times and expected number in the queue. There has been extensive work carried out for $GI/G/1$ systems. Most of the earlier results are approximations for the waiting times and bounds for the queue lengths. For a detailed list of references on this see [2] and [1]. The known results for queues with random size bulk arrivals and random size batch service are mainly for the cases with Poisson arrival processes and general service times ([3,10, 7,13] and [16] all for unlimited size buffer; [4] studied the finite buffer case and [5] studied the finite buffer case also including both $M^X/G^Y/1$ and the $GI^X/M^Y/1$ types). In this current paper we study the $GI^X/G^Y/1$ system in discrete time. We present an algorithmic method, based on the matrix-analytic approach [8], and [11,12] for obtaining the distributions of the queue length, waiting times in the system and the age of the leading customer in the system.

The rest of the paper is organized as follows. In Section 2, the queueing model of interest is defined. A Markov chain that is associated with the number of customers waiting in queue, the number of customers in service, and the phases of the underlying Markov chains of the arrival process and the service time is introduced. A simple condition for the queueing system to be stable is given. In Section 3, the stationary distributions of the queue length at an arbitrary epoch and at the arrival epoch of an arbitrary customer are derived, respectively. Based on the results on the queue lengths, the stationary distribution of the waiting time of an arbitrary customer is obtained. In Section 4, some practical situations are considered. Computational issues related to the stationary distributions are addressed. A Markov chain associated with the age process of the customers in service is introduced and investigated in Section 5. A number of numerical examples are presented in Section 6. Finally, Section 7 discusses future research in this area.

2. The discrete time $GI^X/G^Y/1$ queueing system

The queueing system of interest has a simple structure: a single type of customers join a single queue and are served by a single server on a first-come-first-served (FCFS) basis. Customers arrive at the queueing system in batches and are served by the server in batches. We observe the system status at equally spaced time epochs sequentially numbered as 0, 1, 2, ... Observations are carried out only at the beginning of time periods. Therefore, all events occur at time epochs 0, 1, 2, ... Next, we define the arrival process and the service process explicitly.

Batches of customers arrive at the queueing system according to a discrete time PH-renewal process, where the interarrival times between any two consecutive batches have a common discrete time PH-distribution with possibly infinite number of phases. Denote by $\mathbf{A}$ a generic random variable of the interarrival time of batches.

- Then $\mathbf{A}$ has a discrete time PH-distribution with a matrix representation $(\boldsymbol{\alpha}, T)$ of order K_a, where $\boldsymbol{\alpha}$ is a substochastic measure, T is a substochastic matrix, and K_a is a positive integer that can be finite or infinite. Let $a = 1/E[\mathbf{A}]$, $\mathbf{T}^0 = \mathbf{e} - T\mathbf{e}$, and $Q_a = T + \mathbf{T}^0\boldsymbol{\alpha}$, where $\mathbf{e}$ is a column vector with all elements being one. Denote by $\boldsymbol{\theta}_a$ the left invariant vector of Q_a, i.e., $\boldsymbol{\theta}_a = \boldsymbol{\theta}_a Q_a$, and $\boldsymbol{\theta}_a\mathbf{e} = 1$. Then we have $a = \boldsymbol{\theta}_a\mathbf{T}^0$.
- Customers arrive in batches. Let δ_n be the probability that an arbitrary batch has n customers, $n = 1, 2, 3, \ldots$. Denote by $\delta = \sum_{n=1}^{\infty} n\delta_n < \infty$, the mean size of an arriving batch. If a batch has n customers, then the n customers are ordered as number 1, 2, 3, . . ., and n customer (called the 1st customer, 2nd customer, 3rd customer,

..., and the nth customer). They join the queue in that order as well. Consequently, if $i > j$, then the jth customer is served no later than the ith customer.

Customers are served in batches. The service batch size is determined randomly. If the server is available and the number of customers waiting in the queue is greater than or equal to the service batch size r, then r customers enter the server for service. If the number of customers waiting in the queue is less than the service batch size, then all customers enter the server. Once a service is started, no more customers can enter the server until the current service is complete. Denote by $\mathbf{S}$ a generic random variable of the batch service time.

- Then $\mathbf{S}$ has a discrete time PH-distribution with a matrix representation $(\boldsymbol{\beta}, S)$ of order K_s, where $\boldsymbol{\beta}$ is a substochastic measure, S is a substochastic matrix, and K_s is a positive integer that can be finite or infinite. Let $s = 1/E[\mathbf{S}]$, $\mathbf{S}^0 = \mathbf{e} - S\mathbf{e}$, and $Q_s = S + \mathbf{S}^0\boldsymbol{\beta}$. Denote by $\boldsymbol{\theta}_s$ the left invariant vector of Q_s, i.e., $\boldsymbol{\theta}_s = \boldsymbol{\theta}_s Q_s$, and $\boldsymbol{\theta}_s\mathbf{e} = 1$. Then we have $s = \boldsymbol{\theta}_s\mathbf{S}^0$.
- Let γ_n be the probability that the service batch size is n, for $n = 1, 2, 3, \ldots$. Denote by $\gamma = \sum_{n=1}^{\infty} n\gamma_n < \infty$, the mean size of a service batch. Let $\mathbf{r} = (\gamma_1, \gamma_2, \ldots)$.

We assume that the sizes of arriving batches, the sizes of service batches, the interarrival times, and the service times are random variables independent of each other.

Example 2.1. Consider a special case of the $GI^X/G^Y/1$ queue for which the interarrival times and service times have general distributions. Assume that the generic interarrival time has a general distribution $P\{\mathbf{A} = n\} = a_n$, for $n \geq 1$; and the generic service time has a general distribution $P\{\mathbf{S} = n\} = s_n$, $n \geq 1$. Both the distribution of the interarrival time and the distribution of the service time are PH-distributions with matrix representations $(\boldsymbol{\alpha}, T)$ and $(\boldsymbol{\beta}, S)$, respectively, where

$$\boldsymbol{\alpha} = (a_1, a_2, \cdots), \quad \boldsymbol{\beta} = (s_1; s_2, \ldots), \quad T = S = \begin{pmatrix} 0 & & & \\ 1 & 0 & & \\ & 1 & 0 & \\ & & \ddots & \ddots \end{pmatrix}. \tag{2.1}$$

It is easy to verify that $\boldsymbol{\theta}_a = a(1, 1 - a_1, 1 - a_1 - a_2, \ldots)$ and $\boldsymbol{\theta}_s = s(1, 1 - s_1, 1 - s_1 - s_2, \ldots)$ for this case.

Example 2.1 demonstrates that the $GI^X/G^Y/1$ queueing system is a quite general queueing model.

In order to study the queue length and waiting times as well as to obtain information on some other performance measures of the queueing model, we introduce the following Markov chain for the state of the queueing system. Let

q_n be the number of customers waiting in the queue at time n;

M_n be the number of customers receiving service at time n;

$J_{a,n}$ be the phase of the interarrival time at time n;

$J_{s,n}$ be the phase of the service time of the on-going service at time n.

Note that $q_n = -1$ if the queueing system is empty. It is easy to see that $\{(q_n, M_n, J_{a,n}, J_{s,n}), n \geq 0\}$ is a Markov chain with a state space $\{(-1, k), 1 \leq k \leq K_a\} \cup \{(q, m, k, j), q \geq 0, m \geq 1, 1 \leq k \leq K_a, 1 \leq j \leq K_s\}$. We consider q_n as the level variable and others as auxiliary variables. Then the transition probability matrix of $\{(q_n, M_n, J_{a,n}, J_{s,n}), n \geq 0\}$ is of the $GI/G/1$ type with blocks of finite or infinite size and can be written as

$$P = \begin{pmatrix} T & B_0 & B_1 & B_2 & B_3 & \cdots \\ \mathbf{e} \otimes T \otimes \mathbf{S}^0 & \bar{A}_0 & A_1 & A_2 & A_3 & \cdots \\ 0 & \bar{A}_{-1} & A_0 & A_1 & A_2 & \ddots \\ 0 & \bar{A}_{-2} & A_{-1} & A_0 & A_1 & \ddots \\ 0 & \bar{A}_{-3} & A_{-2} & A_{-1} & A_0 & \ddots \\ \vdots & \vdots & \vdots & \ddots & \ddots & \ddots \end{pmatrix}, \tag{2.2}$$

where

$$
\begin{aligned}
B_0 &= \left(\delta_1 \sum_{m=1}^{\infty} \gamma_m, \delta_2 \sum_{m=2}^{\infty} \gamma_m, \ldots\right) \otimes \left(\mathbf{T}^0\boldsymbol{\alpha}\right) \otimes \boldsymbol{\beta}; \\
B_n &= (\delta_{1+n}\gamma_1, \delta_{2+n}\gamma_2, \cdots) \otimes \left(\mathbf{T}^0\boldsymbol{\alpha}\right) \otimes \boldsymbol{\beta}, \quad n \geq 1;
\end{aligned} \tag{2.3}
$$

$$
\begin{aligned}
A_n &= (\mathbf{e} \otimes (\delta_{1+n}\gamma_1, \delta_{2+n}\gamma_2, \ldots)) \otimes \left(\mathbf{T}^0\boldsymbol{\alpha}\right) \otimes \left(\mathbf{S}^0\boldsymbol{\beta}\right) + \delta_n I \otimes \left(\mathbf{T}^0\boldsymbol{\alpha}\right) \otimes S, \quad n \geq 1; \\
A_0 &= (\mathbf{e} \otimes (\delta_1\gamma_1, \delta_2\gamma_2, \ldots)) \otimes \left(\mathbf{T}^0\boldsymbol{\alpha}\right) \otimes \left(\mathbf{S}^0\boldsymbol{\beta}\right) + I \otimes T \otimes S; \\
A_{-n} &= \left(\mathbf{e} \otimes \left(\sum_{m=n+1}^{\infty} \delta_{m-n}\gamma_m \mathbf{e}(m)\right)\right) \otimes \left(\mathbf{T}^0\boldsymbol{\alpha}\right) \otimes \left(\mathbf{S}^0\boldsymbol{\beta}\right) + (\gamma_n \mathbf{e} \otimes \mathbf{e}(n)) \otimes T \otimes \left(\mathbf{S}^0\boldsymbol{\beta}\right), \quad n \geq 1; \\
\bar{A}_0 &= \left(\mathbf{e} \otimes \left(\sum_{m=1}^{\infty} \delta_m \left(\sum_{j=m}^{\infty} \gamma_j\right) \mathbf{e}(m)\right)\right) \otimes \left(\mathbf{T}^0\boldsymbol{\alpha}\right) \otimes \left(\mathbf{S}^0\boldsymbol{\beta}\right) + I \otimes T \otimes S; \\
\bar{A}_{-n} &= \left(\mathbf{e} \otimes \left(\sum_{m=1}^{\infty} \delta_m \left(\sum_{j=m+n}^{\infty} \gamma_j\right) \mathbf{e}(m+n)\right)\right) \otimes \left(\mathbf{T}^0\boldsymbol{\alpha}\right) \otimes \left(\mathbf{S}^0\boldsymbol{\beta}\right) \\
&\quad + \left(\left(\sum_{j=n}^{\infty} \gamma_j\right) \mathbf{e} \otimes \mathbf{e}(n)\right) \otimes T \otimes \left(\mathbf{S}^0\boldsymbol{\beta}\right), \quad n \geq 1;
\end{aligned}
$$

I is the identity matrix, and $\mathbf{e}(m)$ is a row vector with the mth element being one and all other elements being zero, for $m \geq 0$. In Eq. (2.3), "$\otimes$" is for the Kronecker product of matrices.

We say that the queueing system is stable if the Markov chain $\{(q_n, M_n, J_{a,n}, J_{s,n}), n \geq 0\}$ is positive recurrent. The next theorem gives a condition to ensure the stability of the queueing system.

Theorem 1. *Assume that the Markov chain* $\{(q_n, M_n, J_{a,n}, J_{s,n}), n \geq 0\}$ *is irreducible. Then the Markov chain* $\{(q_n, M_n, J_{a,n}, J_{s,n}), n \geq 0\}$ *is positive recurrent if and only if* $a\delta < s\gamma$.

Proof. By definitions, $a\delta$ can be interpreted as the mean number of customers arriving per unit time and $s\gamma$ can be interpreted as the mean number of customers that can be served per unit time. Thus, for a stable queue, we must have $a\delta < s\gamma$.

Next, we show the sufficiency of the condition $a\delta < s\gamma$. By definition, we have

$$
A = \sum_{n=-\infty}^{\infty} A_n = I \otimes Q_a \otimes S + (\mathbf{e} \otimes (\gamma_1, \gamma_2, \ldots)) \otimes Q_a \otimes (\mathbf{S}^0\boldsymbol{\beta}). \tag{2.4}
$$

It is easy to verify that A is a stochastic matrix. Let $\boldsymbol{\theta} = \mathbf{r} \otimes \boldsymbol{\theta}_a \otimes \boldsymbol{\theta}_s$. Then

$$
\begin{aligned}
\boldsymbol{\theta}A &= \mathbf{r} \otimes (\boldsymbol{\theta}_a Q_a) \otimes (\boldsymbol{\theta}_s S) + (\gamma_1, \gamma_2, \ldots) \otimes (\boldsymbol{\theta}_a Q_a) \otimes \left(\boldsymbol{\theta}_s \mathbf{S}^0\boldsymbol{\beta}\right) \\
&= \mathbf{r} \otimes \boldsymbol{\theta}_a \otimes \left(\boldsymbol{\theta}_s (S + \mathbf{S}^0\boldsymbol{\beta})\right) \\
&= \boldsymbol{\theta}.
\end{aligned} \tag{2.5}
$$

Thus, $\boldsymbol{\theta}$ is an invariant measure of A. Since the Markov chain $\{(q_n, M_n, J_{a,n}, J_{s,n}), n \geq 0\}$ is irreducible, and the mean service time and mean arrival batch are finite, it is positive recurrent if and only if

$$
\boldsymbol{\theta} \sum_{n=1}^{\infty} n A_{-n} \mathbf{e} > \boldsymbol{\theta} \sum_{n=1}^{\infty} n A_n \mathbf{e}. \tag{2.6}
$$

Note that this ergodicity condition can be shown by a method similar to that of [11] for the case where all the distributions have a finite support (e.g., using the mean-drift method). The right-hand side of Eq. (2.6) can be evaluated

as follows

$$\begin{aligned}\boldsymbol{\theta}\sum_{n=1}^{\infty} nA_n\mathbf{e} &= \boldsymbol{\theta}\left[\mathbf{e}\otimes\left(\sum_{n=1}^{\infty} n\delta_{1+n}\gamma_1, \sum_{n=1}^{\infty} n\delta_{2+n}\gamma_2, \ldots\right)\otimes\left(\mathbf{T}^0\boldsymbol{\alpha}\right)\otimes\left(\mathbf{S}^0\boldsymbol{\beta}\right)\right]\mathbf{e}\\ &\quad+\boldsymbol{\theta}\left[\sum_{n=1}^{\infty} n\delta_n I\otimes\left(\mathbf{T}^0\boldsymbol{\alpha}\right)\otimes S\right]\mathbf{e}\\ &= \left(\sum_{m=1}^{\infty}\sum_{n=1}^{\infty} n\delta_{m+n}\gamma_m\right)\left(\boldsymbol{\theta}_a\mathbf{T}^0\right)\left(\boldsymbol{\theta}_s\mathbf{S}^0\right)+\delta\left(\boldsymbol{\theta}_a\mathbf{T}^0\right)\left(\boldsymbol{\theta}_s S\mathbf{e}\right)\\ &= as\left(\sum_{m=1}^{\infty}\sum_{n=1}^{\infty} n\delta_{m+n}\gamma_m\right)+\delta a\,(1-s)\\ &= as\left(2\sum_{m=1}^{\infty} m\delta_m\gamma_m-\sum_{m=1}^{\infty}\left(\sum_{n=1}^{m} n\delta_n\right)\gamma_m-\sum_{m=1}^{\infty}\left(\sum_{n=1}^{m} n\gamma_n\right)\delta_m\right)+a\delta. \end{aligned}\tag{2.7}$$

The left-hand side of Eq. (2.6) can be evaluated as follows:

$$\begin{aligned}\boldsymbol{\theta}\sum_{n=1}^{\infty} nA_{-n}\mathbf{e} &= \boldsymbol{\theta}\left[\sum_{n=1}^{\infty} n\mathbf{e}\otimes\left(\sum_{m=n+1}^{\infty}\delta_{m-n}\gamma_m\mathbf{e}(m)\right)\otimes\left(\mathbf{T}^0\boldsymbol{\alpha}\right)\otimes\left(\mathbf{S}^0\boldsymbol{\beta}\right)\right]\mathbf{e}\\ &\quad+\boldsymbol{\theta}[\mathbf{e}\otimes(\gamma_1, 2\gamma_2, \ldots)\otimes T\otimes(\mathbf{S}^0\boldsymbol{\beta})]\mathbf{e}\\ &= \left(\sum_{m=1}^{\infty} n\left(\sum_{n=1}^{\infty}\delta_m\gamma_{m+n}\right)\right)\left(\boldsymbol{\theta}_a\mathbf{T}^0\right)\left(\boldsymbol{\theta}_s\mathbf{S}^0\right)+\gamma\left(\boldsymbol{\theta}_a T\right)\left(\boldsymbol{\theta}_s\mathbf{S}^0\right)\\ &= as\left(\sum_{m=1}^{\infty} n\left(\sum_{n=1}^{\infty}\delta_m\gamma_{m+n}\right)\right)+\gamma\,(1-a)\,s\\ &= as\left(2\sum_{m=1}^{\infty} m\delta_m\gamma_m-\sum_{m=1}^{\infty}\left(\sum_{n=1}^{m} n\delta_n\right)\gamma_m-\sum_{m=1}^{\infty}\left(\sum_{n=1}^{m} n\gamma_n\right)\delta_m\right)+s\gamma. \end{aligned}\tag{2.8}$$

By Eqs. (2.7) and (2.8), Eq. (2.6) is equivalent to $a\delta < s\gamma$. This completes the proof of Theorem 1. ■

In the rest of the paper, we shall assume that $a\delta < s\gamma$, i.e., the queueing system is stable.

3. Stationary distributions of queue lengths and waiting times

Denote by $\mathbf{x}$ the stationary distribution of the Markov chain $\{(q_n, M_n, J_{a,n}, J_{s,n}), n \geq 0\}$. Then $\mathbf{x}$ can be obtained by solving the equations

$$\mathbf{x} = \mathbf{x}P, \qquad \mathbf{xe} = 1, \tag{3.1}$$

with $\mathbf{x} = (\mathbf{x}_{-1}, \mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2, \ldots)$, $\mathbf{x}_q = (\mathbf{x}_{q,1}, \mathbf{x}_{q,2}, \mathbf{x}_{q,3}, \ldots)$, for $q \geq 0$, and

$$\begin{aligned}&(\mathbf{x}_{-1})_k = \lim_{n\to\infty} P\{q_n = -1, J_{a,n} = k | q_0, M_0, J_{a,0}, J_{s,0}\}, \quad 1 \leq k \leq K_a;\\ &(\mathbf{x}_{q,m})_{k,j} = \lim_{n\to\infty} P\{q_n = q, M_n = m, J_{a,n} = k, J_{s,n} = j | q_0, M_0, J_{a,0}, J_{s,0}\},\\ &\quad 1 \leq k \leq K_a, 1 \leq j \leq K_s, m \geq 1, q \geq 0.\end{aligned}\tag{3.2}$$

Since the Markov chain $\{(q_n, M_n, J_{a,n}, J_{s,n}), n \geq 0\}$ is of $GI/G/1$ type, solving Eq. (3.1) is, in general, not straightforward. Usually, a truncation method must be applied. On the other hand, for real applications, the sizes of the arriving batches and the service batches are always bounded from above. The distributions of the interarrival times and service times always have a finite support as well. For these cases, the Markov chain $\{(q_n, M_n, J_{a,n}, J_{s,n}), n \geq 0\}$ has a special structure such as the $GI/M/1$, $M/G/1$, or QBD type. Then some existing algorithms can be used for

computing the distribution $\mathbf{x}$ accordingly ([6], [8], and [11]). More discussion on the computation of $\mathbf{x}$ can be found in Section 4.

Once the stationary distribution $\mathbf{x}$ can be found, the stationary distribution of the queue length at a customer arrival epoch can be found. Denote by $\mathbf{y}_{q,m}$ the vector whose elements are the probabilities that an arbitrary customer finds q customers waiting ahead of it (this takes account of the customers that arrive in the same batch with that customer) and m customers currently in service right after its arriving epoch. We define $\mathbf{y}_{-1}$ for the server idle situation. Then $\{\mathbf{y}_{-1}, \mathbf{y}_{q,m}, q \geq 0, m \geq 1\}$ represents the stationary distribution of the system state as seen by an arbitrary arriving customer. Denote by

$$\delta_n^* = \sum_{j=n}^{\infty} \delta_j \quad \text{and} \quad \gamma_n^* = \sum_{j=n}^{\infty} \gamma_j, \quad \text{for } n \geq 1. \tag{3.3}$$

By the renewal theory, δ_n^*/δ is the probability that an arbitrary customer is the nth customer within its arriving batch (i.e., there are $n-1$ customers in its batch who should be served no later than the customer under consideration), for $n \geq 1$.

By conditioning on the system state at an arbitrary arriving epoch, a relationship between the queue length distributions at an arbitrary epoch and at the arriving epoch can be established as follows:

$$\mathbf{y}_{-1} = \frac{\mathbf{x}_{-1}(\mathbf{T}^0\boldsymbol{\alpha})}{a}\frac{\delta_1^*}{\delta} + \sum_{w=1}^{\infty}\left(\frac{\mathbf{x}_{0,w}\left[(\mathbf{T}^0\boldsymbol{\alpha})\otimes\mathbf{S}^0\right]}{a}\frac{\delta_1^*}{\delta}\right), \tag{3.4}$$

$$\begin{aligned}\mathbf{y}_{0,m} &= \frac{\mathbf{x}_{-1}\left[(\mathbf{T}^0\boldsymbol{\alpha})\otimes\boldsymbol{\beta}\right]}{a}\left(\frac{\delta_{m+1}^*}{\delta}\gamma_m^*\right) + \frac{\mathbf{x}_{0,m}\left[(\mathbf{T}^0\boldsymbol{\alpha})\otimes S\right]}{a}\left(\frac{\delta_1^*}{\delta}\right)\\ &\quad + \sum_{j=0}^{m}\sum_{w=1}^{\infty}\left(\frac{\mathbf{x}_{j,w}\left[(\mathbf{T}^0\boldsymbol{\alpha})\otimes\left(\mathbf{S}^0\boldsymbol{\beta}\right)\right]}{a}\left(\frac{\delta_{1+m-j}^*}{\delta}\gamma_m^*\right)\right), \quad m \geq 1;\end{aligned} \tag{3.5}$$

$$\begin{aligned}\mathbf{y}_{n,m} &= \frac{\mathbf{x}_{-1}\left[(\mathbf{T}^0\boldsymbol{\alpha})\otimes\boldsymbol{\beta}\right]}{a}\left(\frac{\delta_{1+n+m}^*}{\delta}\gamma_m\right) + \sum_{j=0}^{n}\left(\frac{\mathbf{x}_{j,m}\left[(\mathbf{T}^0\boldsymbol{\alpha})\otimes S\right]}{a}\left(\frac{\delta_{1+n-j}^*}{\delta}\right)\right)\\ &\quad + \sum_{j=0}^{n+m}\sum_{w=1}^{\infty}\left(\frac{\mathbf{x}_{j,w}\left[(\mathbf{T}^0\boldsymbol{\alpha})\otimes\left(\mathbf{S}^0\boldsymbol{\beta}\right)\right]}{a}\left(\frac{\delta_{1+n+m-j}^*}{\delta}\gamma_m\right)\right), \quad n \geq 1, m \geq 1.\end{aligned} \tag{3.6}$$

Apparently, we must have $\mathbf{y}_{-1}\mathbf{e} + \sum_{n=0}^{\infty}\sum_{m=1}^{\infty}\mathbf{y}_{n,m}\mathbf{e} = 1$, which is useful in debugging programs and checking accuracy for computations.

Once the stationary distribution $\mathbf{x}$ is obtained, the stationary distribution $\{\mathbf{y}_{-1}, \mathbf{y}_{q,m}, q \geq 0, m \geq 1\}$ can be computed as well. We present an interpretation to equations in (3.4)–(3.6). First note that

(a) $\frac{\mathbf{x}_{-1}(\mathbf{T}^0\boldsymbol{\alpha})}{a}$ and $\frac{\mathbf{x}_{-1}(\mathbf{T}^0\boldsymbol{\alpha})\otimes\boldsymbol{\beta}}{a}$ are the (vector) probability that a batch of customers arrives, and the queueing system was empty;

(b) $\frac{\mathbf{x}_{0,m}\left[(\mathbf{T}^0\boldsymbol{\alpha})\otimes S\right]}{a}$ is the (vector) probability that a batch of customers arrives, no service completion at the arriving epoch, the queue was empty, and m customers were in service,

(c) $\frac{\mathbf{x}_{j,w}\left[(\mathbf{T}^0\boldsymbol{\alpha})\otimes\left(\mathbf{S}^0\boldsymbol{\beta}\right)\right]}{a}$ is the (vector) probability that a batch of customer arrives, a service completes, there were j customers waiting in the queue, and w customers were in service.

(d) $\frac{\delta_j^*}{\delta}\gamma_k$ can be interpreted as the joint probability that an arriving customer is the jth customer in its batch and the size of the next service batch is k.

To show Eq. (3.4), note that there are only two cases for an arbitrary customer to enter the server upon its arrival. The first case is that the queueing system is empty at the arriving epoch. The second case is that a service is completed at the arriving epoch. For the first case, the arbitrary customer enters the server if the size of the next service batch is larger than or equal to the ordinal number of the arbitrary customer in its batch. For the second case, the arbitrary customer enters the server if the size of the next service batch is larger than or equal to the ordinal number of the

arbitrary customer in its batch plus the number of customers already waiting in the queue. Combining the two cases, Eq. (3.4) is proved. Eqs. (3.5) and (3.6) can be proved similarly.

Now, we are ready to develop a method for computing the stationary distribution of the actual waiting time of an arbitrary customer. For that purpose, we consider a Markov chain given as

$$P_w = \begin{pmatrix} 1 & & & & \\ \gamma_1^* \mathbf{S}^0 & S & & & \\ \gamma_2^* \mathbf{S}^0 & \gamma_1 \mathbf{S}^0 \boldsymbol{\beta} & S & & \\ \gamma_3^* \mathbf{S}^0 & \gamma_2 \mathbf{S}^0 \boldsymbol{\beta} & \gamma_1 \mathbf{S}^0 \boldsymbol{\beta} & S & \\ \vdots & \vdots & \ddots & \ddots & \ddots \end{pmatrix}. \tag{3.7}$$

The Markov chain P_w describes the service process, which has a state space $\{-1\} \cup \{(q, j), q \geq 0, 1 \leq j \leq K_s\}$. The absorption time into the state -1 (level -1) of P_w corresponds to the waiting time of a customer, if an initial distribution is given.

We define a probability vector $\mathbf{z}^{(0)}$ for the joint distribution of the number of customers in queue who are ahead of an arbitrary customer and the service phase right after the arriving epoch of the arbitrary customer. Decompose $\mathbf{z}^{(0)}$ as $\mathbf{z}^{(0)} = (z_{-1}^{(0)}, \mathbf{z}_0^{(0)}, \mathbf{z}_1^{(0)}, \ldots)$. By the definitions of $\mathbf{z}^{(0)}$ and $\{\mathbf{y}_{-1}, \mathbf{y}_{q,m}, q \geq 0, m \geq 1\}$, it can be shown that

$$\begin{aligned} z_{-1}^{(0)} &= \mathbf{y}_{-1}\mathbf{e}; \\ \mathbf{z}_q^{(0)} &= \sum_{m=1}^{\infty} \mathbf{y}_{q,m}(\mathbf{e} \otimes I), \quad q \geq 0. \end{aligned} \tag{3.8}$$

Let

$$\mathbf{z}^{(n)} = \mathbf{z}^{(n-1)} P_w = \mathbf{z}^{(0)} (P_w)^n, \quad n \geq 1. \tag{3.9}$$

Let w_n be the probability that the actual waiting time of an arbitrary customer in the queue is less than or equal to n. It is easy to see that the actual waiting time of an arbitrary customer is the absorption time to the state -1 of the Markov chain P_w, given that the initial distribution is $\mathbf{z}^{(0)}$. Then we have

$$w_n = z_{-1}^{(n)}, \quad n \geq 0. \tag{3.10}$$

Using the distribution $\{w_n, n \geq 0\}$ and the distribution of the batch service time, the distribution of the sojourn time of an arbitrary customer can be found.

In summary, if the stationary distribution $\mathbf{x}$ of the Markov chain $\{(q_n, M_n, J_{a,n}, J_{s,n}), n \geq 0\}$ can be found, then Eqs. (3.4)–(3.6) can be used for computing the stationary distribution of the system state seen by an arbitrary customer right after its arrival, and Eq. (3.10) can be used for computing the stationary distribution of the actual waiting time of an arbitrary customer.

We point out that Eq. (3.9) can be approximated by using the rectangle-iterative approach suggested by [15]. For a given $\varepsilon > 0$, choose a positive integer N_0 such that

$$\sum_{j=N_0+1}^{\infty} \mathbf{z}_j^{(0)} \mathbf{e} < \varepsilon. \tag{3.11}$$

Let $\mathbf{z}^{(0)}(N_0) = (\mathbf{z}_0^{(0)}, \mathbf{z}_1^{(0)}, \ldots, \mathbf{z}_{N_0}^{(0)})$. Consider the matrix $P_w(m-1, m)$, written as

$$P_w(m-1, m) = \begin{pmatrix} S & 0 & & & & \\ \gamma_1 \mathbf{S}^0 \boldsymbol{\beta} & S & 0 & & & \\ \gamma_2 \mathbf{S}^0 \boldsymbol{\beta} & \gamma_1 \mathbf{S}^0 \boldsymbol{\beta} & S & 0 & & \\ \vdots & \ddots & \ddots & \ddots & \ddots & \\ \gamma_{N_0+m-1} \mathbf{S}^0 \boldsymbol{\beta} & \cdots & \gamma_2 \mathbf{S}^0 \boldsymbol{\beta} & \gamma_1 \mathbf{S}0 \boldsymbol{\beta} & S & 0 \end{pmatrix}, \quad m \geq 1. \tag{3.12}$$

According to Lemma 4.1 in [15], the probability that the Markov chain has not been absorbed into level -1 at time n is approximated by

$$1 - w_n \approx \mathbf{z}^{(0)}(N_0) P_w(0,1) P_w(1,2) \cdots P_w(n-1,n)\mathbf{e}, \quad n \geq 1. \tag{3.13}$$

There is a good discussion on the error bounds for this method in [15].

4. Practical situations

In Sections 2 and 3, some theoretical results on the $GI^X/G^Y/1$ queue are obtained. However, some of the matrices involved may be of infinite size, which may cause problems in numerical computation. Nonetheless, as was pointed out in Section 3, for practical situations, distributions of the sizes of arriving batches and the sizes of service batches always have finite supports. The supports of the interarrival times and service times are finite as well. For such cases, the transition matrix P of the Markov chain $\{(q_n, M_n, J_{a,n}, J_{s,n}), n \geq 0\}$ exhibits the properties of both the non-skip-free $M/G/1$ and the non-skip-free $GI/M/1$ types of Markov chains presented in [6], which is based on re-blocking the transition matrix P. In this section, we show how to find the stationary distribution $\mathbf{x}$ for such cases. Define

$$\begin{aligned} N &= \max\{n : \delta_n \neq 0, n \geq 1\}; \\ M &= \max\{n : \gamma_n \neq 0, n \geq 1\}. \end{aligned} \tag{4.1}$$

Next, we consider three cases based on the values of N and M.

The QBD case: We assume $N < \infty$, $M < \infty$, $K_a < \infty$, and $K_s < \infty$. For this case, the state space of Markov chain $\{(q_n, M_n, J_{a,n}, J_{s,n}), n \geq 0\}$ is $\{(-1,k), 1 \leq k \leq K_a\} \cup \{(q,m,k,j), q \geq 0, 1 \leq m \leq M, 1 \leq k \leq K_a, 1 \leq j \leq K_s\}$. Thus, the matrix blocks in Eq. (2.2), which are defined in Eq. (2.3), are matrices of finite size. The transition probability matrix P can be re-blocked into a QBD type transition matrix. For example, if $M \leq N < \infty$, the matrix P can be re-blocked as follows

$$P = \begin{pmatrix} T & B_0 & B & & & \\ \mathbf{e} \otimes T \otimes \mathbf{S}^0 & \bar{A}_0 & \bar{U} & & & \\ & \bar{D} & F & U & & \\ & & D & F & U & \\ & & & D & F & \ddots \\ & & & & \ddots & \ddots \end{pmatrix}, \tag{4.2}$$

where $B = (B_1, B_2, \ldots, B_N)$, $\bar{U} = (A_1, A_2, \ldots, A_N)$,

$$\bar{D} = \begin{pmatrix} \bar{A}_{-1} \\ \vdots \\ \bar{A}_{-M} \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \qquad F = \begin{pmatrix} A_0 & A_1 & \cdots & A_M & \cdots & A_{N-1} \\ A_{-1} & A_0 & \cdots & A_{M-1} & \cdots & A_{N-2} \\ \vdots & \vdots & \ddots & \vdots & \cdots & \vdots \\ A_{-M} & A_{-(M-1)} & \cdots & A_0 & \cdots & A_{N-1-M} \\ & \ddots & \ddots & \cdots & \ddots & \vdots \\ & & A_{-M} & A_{-(M-1)} & \cdots & A_0 \end{pmatrix}, \tag{4.3}$$

$$D = \begin{pmatrix} 0 & \cdots & 0 & A_{-M} & \cdots & A_{-1} \\ & \ddots & \ddots & \ddots & \ddots & \vdots \\ & & \ddots & \ddots & \ddots & A_{-M} \\ & & & \ddots & \ddots & 0 \\ & & & & \ddots & \vdots \\ & & & & & 0 \end{pmatrix}, \qquad U = \begin{pmatrix} A_N & & & \\ A_{N-1} & A_N & & \\ \vdots & \vdots & \ddots & \\ A_1 & A_2 & \cdots & A_N \end{pmatrix}.$$

Similarly, the transition matrix P can be re-blocked into a QBD type matrix for the other case with $N < M$. For both cases, the stationary distribution $\hat{\mathbf{x}}$ has a matrix geometric form and can be found by solving equations:

$$\begin{aligned}
&\hat{\mathbf{x}}_n = \hat{\mathbf{x}}_1 R^{n-1}, \quad n \geq 1;\\
&(\hat{\mathbf{x}}_{-1}, \hat{\mathbf{x}}_0, \hat{\mathbf{x}}_1) = (\hat{\mathbf{x}}_{-1}, \hat{\mathbf{x}}_0, \hat{\mathbf{x}}_1) \begin{pmatrix} T & B_0 & B \\ \mathbf{e} \otimes T \otimes \mathbf{S}^0 & \bar{A}_0 & \bar{U} \\ 0 & \bar{D} & F + RD \end{pmatrix};\\
&\hat{\mathbf{x}}_{-1}\mathbf{e} + \hat{\mathbf{x}}_0\mathbf{e} + \hat{\mathbf{x}}_1(I - R)^{-1}\mathbf{e} = 1,
\end{aligned} \tag{4.4}$$

where the matrix R is the minimal non-negative solution to

$$R = U + RF + R^2 D. \tag{4.5}$$

Existing algorithms can be applied for computing R using Eq. (4.5) (see [8] and [11]). For instance, R can be obtained as the limit of the following non-decreasing sequence $\{R[k], k \geq 0\}$ generated as $R[0] = 0$ and $R[k+1] = (U + R^2[k]D)(I - F)^{-1}$, for $k \geq 0$. Then $\hat{\mathbf{x}}$ can be computed from Eq. (4.5).

The relationship between $\mathbf{x}$ (defined in Section 3) and $\hat{\mathbf{x}}$ is

$$\begin{aligned}
&\hat{\mathbf{x}}_{-1} = \mathbf{x}_{-1};\\
&\hat{\mathbf{x}}_0 = \mathbf{x}_0;\\
&\hat{\mathbf{x}}_{k+1} = (\mathbf{x}_{kN+1}, \mathbf{x}_{kN+2}, \ldots, \mathbf{x}_{kN+N}), \quad 1 \leq j \leq N,\ k \geq 0.
\end{aligned}$$

We also keep in mind that the probability that the queue is empty is actually $\mathbf{x}_{-1}\mathbf{e} + \mathbf{x}_0\mathbf{e}$.

The $M/G/1$ case: $N = \infty$, $M < \infty$, $K_a < \infty$, and $K_s < \infty$. For this case, the state space of the Markov chain $\{(q_n, M_n, J_{a,n}, J_{s,n}), n \geq 0\}$ is $\{(-1, k), 1 \leq k \leq K_a\} \cup \{(q, m, k, j), q \geq 0, 1 \leq m \leq M, 1 \leq k \leq K_a, 1 \leq j \leq K_s\}$. The transition matrix P can be re-blocked into non-skip-free $M/G/1$ type with matrix blocks of finite size as follows

$$P = \begin{pmatrix}
T & B_0 & \bar{B}_1 & \bar{B}_2 & \bar{B}_3 & \cdots \\
\mathbf{e} \otimes T \otimes \mathbf{S}^0 & \bar{A}_0 & \bar{A}_1 & \bar{A}_2 & \bar{A}_3 & \cdots \\
 & \bar{E}_{-1} & E_0 & E_1 & E_2 & \cdots \\
 & & E_{-1} & E_0 & E_1 & \cdots \\
 & & & E_{-1} & E_0 & \ddots \\
 & & & & \ddots & \ddots
\end{pmatrix}, \tag{4.6}$$

where

$$\begin{aligned}
&\bar{B}_n = (B_{1+(n-1)M}, B_{2+(n-1)M}, \ldots, B_{M+(n-1)M}), \quad n \geq 1;\\
&\bar{A}_n = (A_{1+(n-1)M}, A_{2+(n-1)M}, \ldots, A_{M+(n-1)M}), \quad n \geq 1;\\
&\bar{E}_{-1} = \begin{pmatrix} \bar{A}_{-1} \\ \vdots \\ \bar{A}_{-M} \end{pmatrix}, \qquad E_{-1} = \begin{pmatrix} A_{-M} & A_{-(M-1)} & \cdots & A_{-1} \\ & A_{-M} & \cdots & A_{M-2} \\ & & \ddots & \vdots \\ & & & A_{-M} \end{pmatrix},\\
&E_n = \begin{pmatrix} A_{nM} & A_{nM+1} & \cdots & A_{nM+M-1} \\ A_{nM-1} & A_{nM} & \cdots & A_{nM+M-2} \\ \vdots & \vdots & \cdots & \vdots \\ A_{(n-1)M+1} & A_{(n-1)M+2} & \cdots & A_{nM} \end{pmatrix}, \quad n \geq 0.
\end{aligned} \tag{4.7}$$

Existing algorithms can be applied for computing stationary distribution $\mathbf{x}$ ([6], [8], and [11]). For example, a stable algorithm developed in [14] can be applied for computing $\mathbf{x}$. Details are omitted.

The $GI/M/1$ case: $M = \infty$, $N < \infty$, $K_a < \infty$, and $K_s < \infty$. For this case, the transition matrix P can be re-blocked into non-skip-free $GI/M/1$ type. That is true since $B_n = 0$, for $n > N$, and $A_n = 0$, for $n > N$. However,

the matrix blocks are of infinite size. Thus, we consider a new Markov chain $\{(q_n, J_{a,n}, J_{s,n}), n \geq 0\}$, which is also of $GI/M/1$ type. The state space of the Markov chain $\{(q_n, J_{a,n}, J_{s,n}), n \geq 0\}$ is $\{(-1, k), 1 \leq k \leq K_a\} \cup \{(q, k, j), q \geq 0, 1 \leq k \leq K_a, 1 \leq j \leq K_s\}$ and its transition probability matrix is given as

$$\hat{P} = \begin{pmatrix} T & \hat{B}_0 & \hat{B}_1 & \hat{B}_2 & \hat{B}_3 & \cdots \\ T \otimes \mathbf{S}^0 & \tilde{A}_0 & \hat{A}_1 & \hat{A}_2 & \hat{A}_3 & \cdots \\ 0 & \tilde{A}_{-1} & \hat{A}_0 & \hat{A}_1 & \hat{A}_2 & \ddots \\ 0 & \tilde{A}_{-2} & \hat{A}_{-1} & \hat{A}_0 & \hat{A}_1 & \ddots \\ 0 & \tilde{A}_{-3} & \hat{A}_{-2} & \hat{A}_{-1} & \hat{A}_0 & \ddots \\ \vdots & \cdots & \ddots & \ddots & \ddots & \ddots \end{pmatrix}, \tag{4.8}$$

where

$$\begin{aligned}
\hat{B}_0 &= \left(\sum_{k=1}^{\infty} \delta_k \gamma_k^*\right) \left[\left(\mathbf{T}^0 \boldsymbol{\alpha}\right) \otimes \boldsymbol{\beta}\right]; \\
\hat{B}_n &= \left(\sum_{k=1}^{\infty} \delta_{k+n} \gamma_k\right) \left[\left(\mathbf{T}^0 \boldsymbol{\alpha}\right) \otimes \boldsymbol{\beta}\right], \quad n \geq 1; \\
\hat{A}_n &= \left(\sum_{k=1}^{\infty} \delta_{k+n} \gamma_k\right) \left[\left(\mathbf{T}^0 \boldsymbol{\alpha}\right) \otimes \left(\mathbf{S}^0 \boldsymbol{\beta}\right)\right] + \delta_n \left(\mathbf{T}^0 \boldsymbol{\alpha}\right) \otimes S, \quad n \geq 1; \\
\hat{A}_0 &= \left(\sum_{k=1}^{\infty} \delta_k \gamma_k\right) \left[\left(\mathbf{T}^0 \boldsymbol{\alpha}\right) \otimes \left(\mathbf{S}^0 \boldsymbol{\beta}\right)\right] + T \otimes S; \\
\hat{A}_{-n} &= \left(\sum_{k=1}^{\infty} \delta_k \gamma_{k+n}\right) \left[\left(\mathbf{T}^0 \boldsymbol{\alpha}\right) \otimes \left(\mathbf{S}^0 \boldsymbol{\beta}\right)\right] + \gamma_n \left[T \otimes \left(\mathbf{S}^0 \boldsymbol{\beta}\right)\right], \quad n \geq 1; \\
\tilde{A}_0 &= \left(\sum_{k=1}^{\infty} \delta_k \gamma_k^*\right) \left(\mathbf{T}^0 \boldsymbol{\alpha}\right) \otimes \left(\mathbf{S}^0 \boldsymbol{\beta}\right) + T \otimes S; \\
\tilde{A}_{-n} &= \left(\sum_{k=1}^{\infty} \delta_k \gamma_{k+n}^*\right) \left(\mathbf{T}^0 \boldsymbol{\alpha}\right) \otimes \left(\mathbf{S}^0 \boldsymbol{\beta}\right) + \gamma_n^* T \otimes \left(\mathbf{S}^0 \boldsymbol{\beta}\right), \quad n \geq 1.
\end{aligned} \tag{4.9}$$

Similarly to Theorem 1, it can be shown that the Markov chain $\{(q_n, J_{a,n}, J_{s,n}), n \geq 0\}$ is positive recurrent if and only if $a\delta < s\gamma$. Then the Markov chain $\{(q_n, J_{a,n}, J_{s,n}), n \geq 0\}$ has a stationary distribution.

Since $N < \infty$, we have $\hat{B}_n = 0$, for $n > N$, and $\hat{A}_n = 0$, for $n > N$. Then we can re-block the transition matrix in Eq. (4.8) into a non-skip-free $GI/M/1$ type Markov chain as

$$\hat{P} = \begin{pmatrix} T & \hat{B}_0 & \bar{B} & & & \\ T \otimes \mathbf{S}^0 & \tilde{A}_0 & \bar{H}_1 & & & \\ & \bar{H}_{-1} & H_0 & H_1 & & \\ & \bar{H}_{-2} & H_{-1} & H_0 & H_1 & \\ & \bar{H}_{-3} & H_{-2} & H_{-1} & H_0 & H_1 \\ & \vdots & \vdots & \ddots & \ddots & \ddots \end{pmatrix}, \tag{4.10}$$

where

$$\bar{B} = (\hat{B}_1, \hat{B}_2, \ldots, \hat{B}_N), \qquad \bar{H}_1 = (\hat{A}_1, \hat{A}_2, \ldots, \hat{A}_N), \quad n \geq 1; \tag{4.11}$$

$$\bar{H}_{-n} = \begin{pmatrix} \tilde{A}_{-(1+(n-1)N)} \\ \vdots \\ \tilde{A}_{-(N+(n-1)N)} \end{pmatrix}, \quad n \geq 1; \qquad H_1 = \begin{pmatrix} \hat{A}_N & & & \\ \hat{A}_{N-1} & \hat{A}_N & & \\ \vdots & \ddots & \ddots & \\ \hat{A}_1 & \cdots & \hat{A}_{N-1} & \hat{A}_N \end{pmatrix},$$

$$H_{-n} = \begin{pmatrix} \hat{A}_{-nN} & \hat{A}_{1-nN} & \cdots & \hat{A}_{N-1-nN} \\ \hat{A}_{-nN-1} & \hat{A}_{-nN} & \cdots & \hat{A}_{N-2-nN} \\ \vdots & \vdots & \cdots & \vdots \\ \hat{A}_{-nN+1-N} & \hat{A}_{-nN+2-N} & \cdots & \hat{A}_{-nN} \end{pmatrix}, \quad n \geq 0.$$

Thus, the stationary distribution $\boldsymbol{\pi}$ of the Markov chain $\{(q_n, J_{a,n}, J_{s,n}), n \geq 0\}$ has a matrix geometric form and can be found by solving the equations $\boldsymbol{\pi}\hat{P} = \boldsymbol{\pi}$ and $\boldsymbol{\pi}\boldsymbol{e} = 1$. Because of the $GI/M/1$ structure, $\boldsymbol{\pi}$ has a matrix geometric form given as

$$\begin{aligned} &\boldsymbol{\pi}_n = \boldsymbol{\pi}_1 R^{n-1}, \quad n \geq 1; \\ &(\boldsymbol{\pi}_{-1}, \boldsymbol{\pi}_0, \boldsymbol{\pi}_1) = (\boldsymbol{\pi}_{-1}, \boldsymbol{\pi}_0, \boldsymbol{\pi}_1) \begin{pmatrix} T & \hat{B}_0 & \bar{B} \\ T \otimes \mathbf{S}^0 & \tilde{A}_0 & \bar{H}_1 \\ 0 & \sum_{n=1}^{\infty} R^{n-1} \bar{H}_{-n} & \sum_{n=0}^{\infty} R^n H_{-n} \end{pmatrix}; \\ &\boldsymbol{\pi}_{-1}\mathbf{e} + \boldsymbol{\pi}_0\mathbf{e} + \boldsymbol{\pi}_1(I - R)^{-1}\mathbf{e} = 1, \end{aligned} \tag{4.12}$$

where the matrix R is the minimal non-negative solution to

$$R = \hat{H}_1 + \sum_{n=0}^{\infty} R^{n+1} \hat{H}_{-n}. \tag{4.13}$$

Based on Eqs. (4.12) and (4.13), existing algorithms can be applied for computing the stationary distribution $\boldsymbol{\pi}$ of the Markov chain $\{(q_n, J_{a,n}, J_{s,n}), n \geq 0\}$ ([6,8], and [11]).

Based on the above results and results obtained in Sections 2 and 3, for the practical situations, the following performance measures can be obtained.

(1) Distribution of the queue length at an arbitrary time. Consequently, the mean queue length at an arbitrary time can be obtained.

(2) Distribution of the number of customers in service at an arbitrary time.

(3) The total number of customers in the queueing system.

(4) Distribution of the queue length ahead of an arbitrary customer right after its arrival epoch.

(5) Distribution of the actual waiting time of an arbitrary customer.

Numerical examples are presented in Section 6.

5. The age process and sojourn times

In this section, we introduce and study an age process associated with the sojourn times of the batches of customers and the sojourn times of customers. Let

a_n be the total time that the last batch currently in service has been in the queueing system; (Note: Some customers in the last batch may not be in service. Also note that all customers in previous batches are either in service or have left the queueing system.)

R_n be the number of customers from the last batch in service that are not in service;

$I_{a,n}$ be the phase of the arrival process right after the arrival of the last batch in service;

$I_{s,n}$ be the phase of the current service.

It is easy to see that $\{(a_n, R_n, I_{a,n}, I_{s,n}), n \geq 0\}$ is a Markov chain with a state space $\{(q, r, k, j), q \geq 0, r \geq 0, 1 \leq k \leq K_a, 1 \leq j \leq K_s\}$. Furthermore, it is readily seen that the value of a_n can increase by at most one in each transition. Therefore, the Markov chain $\{(a_n, R_n, I_{a,n}, I_{s,n}), n \geq 0\}$ is of $GI/M/1$ type. However, if the batch sizes of

service are larger than one, the Markov chain $\{(a_n, R_n, I_{a,n}, I_{s,n}), n \geq 0\}$ is a *level dependent* $GI/M/1$ type Markov chain. The computation of the stationary distribution (if it exists) of such a Markov chain is not straightforward.

Thus, in the rest of this section, we concentrate on a special case for which the batch sizes of services are one, i.e., $\gamma_1 = 1$, and $\gamma_n = 0$, for $n \geq 2$, or $M = 1$. For such a case, the transition matrix of the Markov chain $\{(a_n, R_n, I_{a,n}, I_{s,n}), n \geq 0\}$ is of level independent $GI/M/1$ type and can be found as follows. If the service continues in a period, the value of a_n increases by one. If the service is completed in a period and the customer who enters service is in the same batch as the customer who just departed, then a_n increases by one. Thus, the conditional probabilities for a_n to increase (by one) are given by

$$F_0 = I \otimes I \otimes S + \begin{pmatrix} 0 & & & \\ 1 & 0 & & \\ & 1 & 0 & \\ & & \ddots & \ddots \end{pmatrix} \otimes I \otimes \left(\mathbf{S}_0\boldsymbol{\beta}\right). \tag{5.1}$$

If a service is complete and a customer from a new batch enters service, then the value of a_n depends on the interarrival time of the two consecutive batches. The probabilities for the value of a_n decreasing by $m-1$ is given by, for $m \geq 1$,

$$F_m = \begin{pmatrix} \delta_1 & \delta_2 & \cdots \\ 0 & 0 & \cdots \\ \vdots & \vdots & \vdots \end{pmatrix} \otimes (T^{m-1}\mathbf{T}^0\boldsymbol{\alpha}) \otimes \left(\mathbf{S}^0\boldsymbol{\beta}\right) \equiv \Delta \otimes (T^{m-1}\mathbf{T}^0\boldsymbol{\alpha}) \otimes \left(\mathbf{S}^0\boldsymbol{\beta}\right). \tag{5.2}$$

In summary, the transition matrix of the Markov chain $\{(a_n, R_n, I_{a,n}, I_{s,n}), n \geq 0\}$ is given by

$$Q = \begin{pmatrix} \bar{F}_0 & F_0 & & & \\ \bar{F}_1 & F_1 & F_0 & & \\ \bar{F}_2 & F_2 & F_1 & F_0 & \\ \vdots & \vdots & \ddots & \ddots & \ddots \end{pmatrix}, \tag{5.3}$$

where

$$\bar{F}_m = \Delta \otimes \left(T^m\mathbf{e}\boldsymbol{\alpha}\right) \otimes \left(\mathbf{S}^0\boldsymbol{\beta}\right), \quad m \geq 0. \tag{5.4}$$

Assume that Q is irreducible. It can be shown that, under the system stability condition given in Theorem 1 (i.e., $a\delta < s\gamma$ or $a\delta < s$ since $\gamma = 1$), the Markov chain Q is positive recurrent. In fact, let

$$\hat{F} = \sum_{m=0}^{\infty} F_m = F_0 + \Delta \otimes (\mathbf{e}\boldsymbol{\alpha}) \otimes \left(\mathbf{S}^0\boldsymbol{\beta}\right). \tag{5.5}$$

It is easy to verify that the probability vector $(\delta_1^*/\delta, \delta_2^*/\delta, \ldots) \otimes \boldsymbol{\alpha} \otimes \boldsymbol{\theta}_s$ is the invariant measure of the matrix $\hat{F}$. By [11], the Markov chain Q is positive recurrent if and only if

$$\left(\frac{1}{\delta}(\delta_1^*, \delta_2^*, \ldots) \otimes \boldsymbol{\alpha} \otimes \boldsymbol{\theta}_s\right)\left(\sum_{m=0}^{\infty} mF_m\right)\mathbf{e} > 1, \tag{5.6}$$

which is equivalent to

$$\begin{aligned}\left(\frac{1}{\delta}(\delta_1^*, \delta_2^*, \ldots) \otimes \boldsymbol{\alpha} \otimes \boldsymbol{\theta}_s\right)\left(\sum_{m=0}^{\infty} mF_m\right)\mathbf{e} &= \left(\frac{1}{\delta}(\delta_1^*, \delta_2^*, \ldots) \otimes \boldsymbol{\alpha} \otimes \boldsymbol{\theta}_s\right)\left(\Delta \otimes ((I-T)^{-2}\mathbf{T}^0\boldsymbol{\alpha}) \otimes \left(\mathbf{S}^0\boldsymbol{\beta}\right)\right)\mathbf{e} \\ &= \left(\frac{1}{\delta}(\delta_1, \delta_2, \ldots) \otimes \left(\boldsymbol{\alpha}(I-T)^{-2}\mathbf{T}^0\boldsymbol{\alpha}\right) \otimes \left(\boldsymbol{\theta}_s\mathbf{S}^0\boldsymbol{\beta}\right)\right)\mathbf{e} \\ &= \frac{s}{\delta a} > 1. \end{aligned} \tag{5.7}$$

In Eq. (5.7), note that $\delta_1^* = 1$. The last inequality in Eq. (5.7) is guaranteed by $a\delta < s$.

For practical situations, we have $N < \infty$, $K_a < \infty$, and $K_s < \infty$. Thus, the matrix blocks in Eq. (5.3) are finite in size. Thus, existing algorithms can be used for computing the stationary distribution of Q. Denote by $\mathbf{u} = (\mathbf{u}_0, \mathbf{u}_1, \ldots, \mathbf{u}_m, \ldots)$ the stationary distribution of Q. We can find $\mathbf{u}$ by solving the linear system $\mathbf{u}Q = \mathbf{u}$ and $\mathbf{ue} = 1$. Let R be the minimal non-negative solution to

$$R = \sum_{k=0}^{\infty} R^k F_k = F_0 + \left(\sum_{k=1}^{\infty} R^k (I \otimes T^{k-1} \otimes I)\right)\left(\Delta \otimes (\mathbf{T}^0\boldsymbol{\alpha}) \otimes (\mathbf{S}^0\boldsymbol{\beta})\right). \tag{5.8}$$

By [11], the stationary distribution $\mathbf{u}$ can be obtained by

$$\begin{aligned} &\mathbf{u}_m = \mathbf{u}_0 R^m, \quad m \geq 0; \\ &\mathbf{u}_0 = \mathbf{u}_0 \sum_{m=0}^{\infty} R^m \bar{F}_m; \\ &\mathbf{u}_0 (I - R)^{-1}\mathbf{e} = 1. \end{aligned} \tag{5.9}$$

In summary, for the case with $M = 1, N < \infty$, $K_a < \infty$, and $K_s < \infty$, we can compute the stationary distribution $\mathbf{u}$ as well as the following performance measures. We decompose the stationary distribution $\mathbf{u}$ further into $\mathbf{u}_m = (\mathbf{u}_{m,0}, \mathbf{u}_{m,1}, \ldots)$, for $m \geq 0$, where the second index is for the number of customers from the last batch in service who are still in the queue.

$\mathbf{u}_m\mathbf{e}$: the probability that the age of the customer currently in service is m, which is also the probability that the age of the batch whose customer is currently in service is m, for $m \geq 0$.

- $\mathbf{u}_m(\mathbf{e} \otimes \mathbf{S}^0)(1 - \mathbf{x}_{-1}\mathbf{e})/(a\delta)$: the probability that the sojourn time of an arbitrary departing customer is $m + 1$, for $m \geq 0$. Then we must have $\left(\sum_{m=0}^{\infty} \mathbf{u}_m\right)(\mathbf{e} \otimes \mathbf{S}^0)(1 - \mathbf{x}_{-1}\mathbf{e}) = a\delta$. That equation establishes a relationship between the stationary distributions of the Markov chain $\{(q_n, M_n, J_{a,n}, J_{s,n}), n \geq 0\}$ defined in Section 2 and the Markov chain $\{(a_n, R_n, I_{a,n}, I_{s,n}), n \geq 0\}$ defined in Section 5. Such a relationship is useful for debugging programs and checking accuracy for computations.
- $\mathbf{u}_{m,r}\mathbf{e}$: the probability that the age of the customer currently in service is m and that customer is the $(r + 1)$th customer in the reversed order in its batch, for $m \geq 0$ and $r \geq 0$. Then $\mathbf{u}_{m,0}\mathbf{e}$ is the probability that the age of the customer currently in service is m and that customer is the last customer in its batch.
- $\mathbf{u}_{m,r}(\mathbf{e} \otimes \mathbf{S}^0)(1 - \mathbf{x}_{-1}\mathbf{e})/(a\delta)$: the probability that the sojourn time of an arbitrary departing customer, which is the $(r+1)$th customer in the reversed order in a batch, is $m+1$, for $m \geq 0$. Note that $\{\mathbf{u}_{m,0}(\mathbf{e} \otimes \mathbf{S}^0)(1 - \mathbf{x}_{-1}\mathbf{e})/(a\delta), m \geq 0\}$ is the distribution of the sojourn time of an arbitrary departing customer, who is the last customer in a batch, which is also the distribution of the sojourn time of an arbitrary batch.

6. Numerical examples

In this section, we present a number of examples that demonstrate some interesting behaviours of the $GI^X/G^Y/1$ queue. In particular, we discuss the relationship between the mean queue length at an arbitrary time, the variances of the arrival batch size and the service batch size, and the distributions of the arrival batch size and the service batch size.

For all the examples to be presented, we consider a $GI^X/G^Y/1$ queue with $K_a = 4$, $\boldsymbol{\alpha} = (0.1, 0.3, 0.2, 0.4)$, and $K_s = 3$, $\boldsymbol{\beta} = (0.2, 0.7, 0.1)$.

Example 6.1. Let $N = 2$, $\delta_1 = 0.8$, $\delta_2 = 0.2$, and $M = 2$, $\gamma_1 = 0.3$, $\gamma_2 = 0.7$. The distributions of the queue length at an arbitrary time and the waiting time of an arbitrary customer are shown in Figs. 1 and 2, respectively. Fig. 1 and Fig. 2 show that, after a certain point, both distributions decrease in a way similar to that of the geometric distribution (exponential decay). However, for small q in Fig. 1 or small n in Fig. 2, the shapes of the distributions are affected by the distributions of the interarrival times, service times, arrival batch sizes, and service batch sizes.

It is interesting to see that the probability that the system is empty is 0.32, while the probability of no waiting is 0.67. The reason is that, in each time period, there can be both arrivals and departures. Thus, it is possible that a

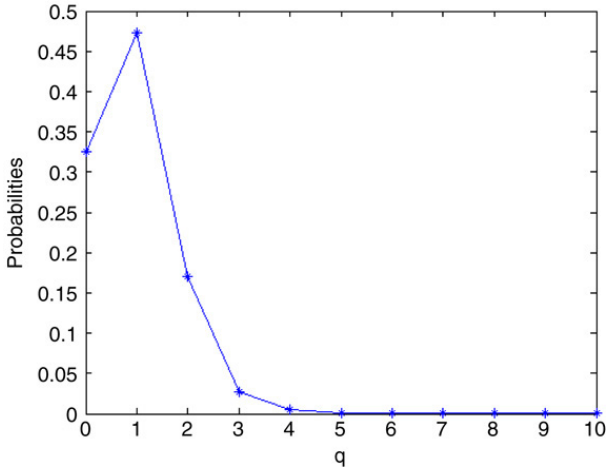

Fig. 1. Distribution of the number of customers in the system.

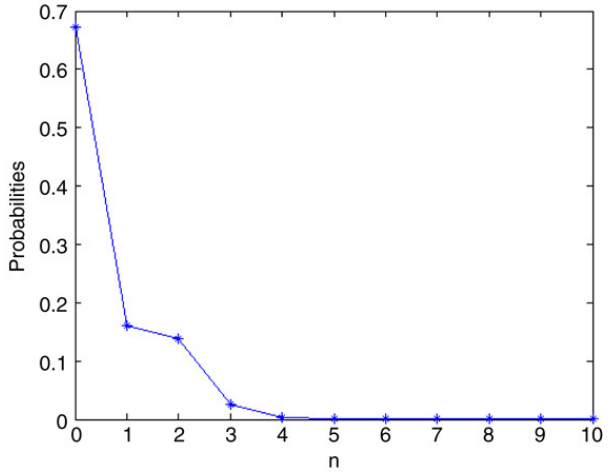

Fig. 2. Distribution of waiting time of an arbitrary customer.

customer (or a group of customers) enters the server without waiting, while the server gets no break between two consecutive services.

Next, we have a look at how the distributions and the variances of batch sizes affect the queueing process. We focus on the mean queue length.

Example 6.2. In this example, we consider the following two cases.

(1) For case 1, let $M = 3$, $\gamma_1 = 0.4$, $\gamma_2 = 0.2$, $\gamma_3 = 0.4$, and $N = 6$, $\delta = 1.7$. For the variance of the arrival batch size between 0.24 and 3, the corresponding mean queue length at an arbitrary time is plotted in Fig. 3.

(2) For case 2, let $N = 9$, $\delta_1 = 0.6$, $\delta_2 = 0.1$, $\delta_3 = 0.1$, $\delta_4 = 0.05$, $\delta_5 = 0.04$, $\delta_6 = 0.05$, $\delta_7 = 0.05$, $\delta_8 = 0.005$, $\delta_9 = 0.005$, and $M = 8$, $\gamma = 1.98$. For the variance of the service batch size between 3.9 and 5.2, the corresponding mean queue length at an arbitrary time is plotted in Fig. 4.

The results from this set of experiments seem to produce results that are counter-intuitive. Intuitively, it would mean that the mean queue length, as a function of the variance, is a monotonic type, or at least a smooth type of function. To a certain extent, the mean queue length is an increasing function of the variance of the arrival batch size or the service batch size, which is demonstrated by the above two examples and many other examples we have tested. However, Fig. 4 also shows that the relationship between the mean queue length and the variance of the arrival/service batch size can be complicated. The reason is that, when the variances of batch sizes are fixed, the actual distributions of batch sizes still have an effect on the mean queue length. The examples demonstrate that the mean queue length is affected not only by the variance of the batch sizes, but also by their actual distributions.

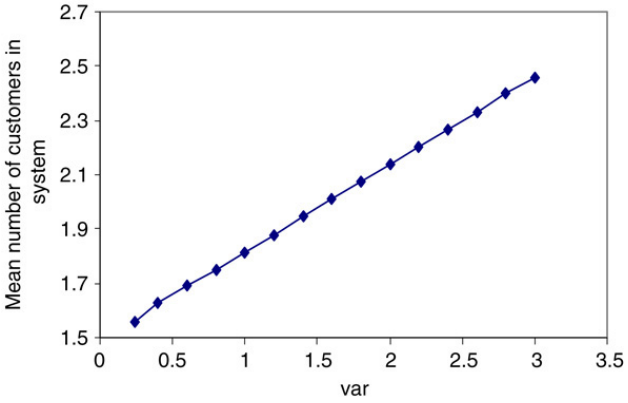

Fig. 3. Mean queue length for case 1.

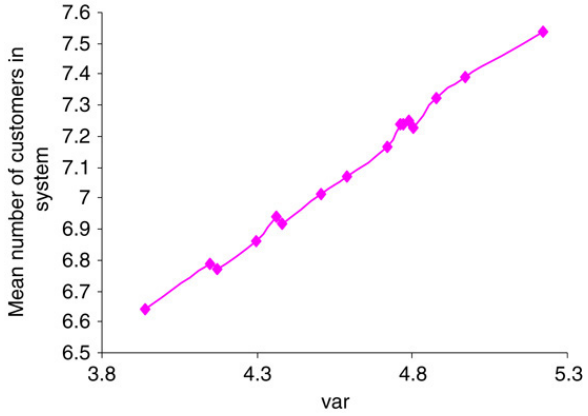

Fig. 4. Mean queue length for case 2.

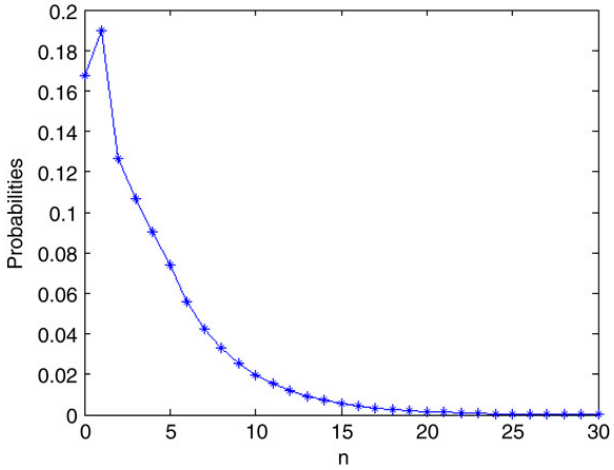

Fig. 5. Probability distribution of the age of a customer in service.

Example 6.3. In this example, results related to the age process are presented. Let $N = 3$, $\delta_1 = 0.8, \delta_2 = 0.1$, $\delta_3 = 0.1$, and $M = 1$, $\gamma_1 = 1$. The distribution of the age of an arbitrary customer in service is plotted in Fig. 5. The distribution of the sojourn time of an arbitrary customer is plotted in Fig. 6. Fig. 7 shows the distribution of the age of the customer in service who is the $(3 - j)$th customer in its arrival batch, for $j = 0, 1$, and 2. Fig. 8 shows the distribution of the sojourn time of the $(3 - j)$th customer in its arrival batch, for $j = 0, 1$, and 2.

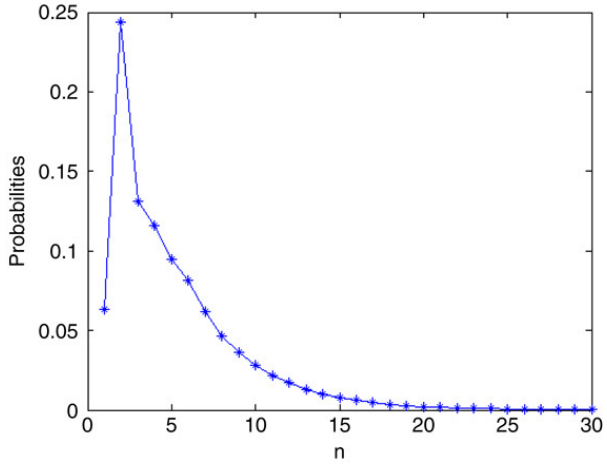

Fig. 6. Probability distribution of the sojourn time of an arbitrary customer.

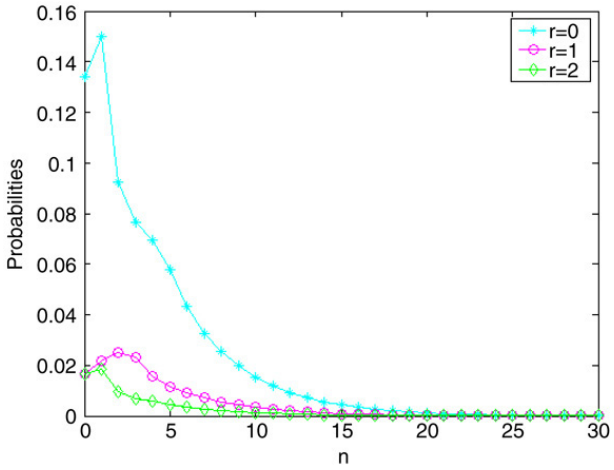

Fig. 7. Probability distributions of the age of customers in service.

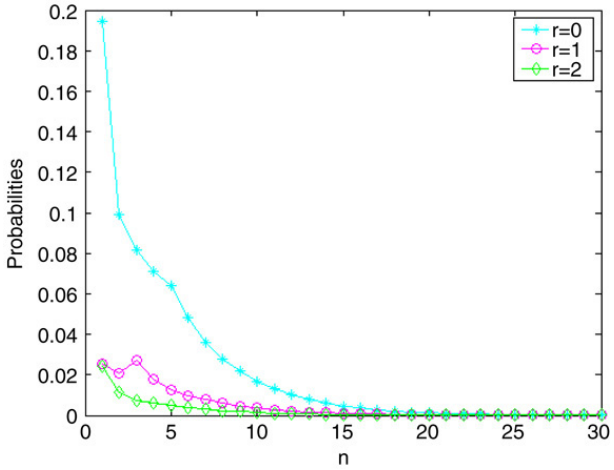

Fig. 8. Probability distributions of the sojourn times of customers in a batch.

Note that the sojourn time of a customer is always positive, while the age can be zero. The probability distributions seem, in most cases, to have the expected forms, i.e. decreasing in a kind of geometric form after a certain point (exponential decay).

In Figs. 7 and 8, the distributions for $r = 1$ and $r = 2$ are close to each other, and they are drastically different from that of $r = 0$. Thus, Figs. 7 and 8 show that the distributions of customers in different positions in an arrival batch may or may not have drastically different waiting times. This demonstrates the necessity to understand the waiting processes of individual customers in different positions.

In summary, our examples demonstrate that the mean queue length and the queueing process are affected not only by the variance of the batch sizes, but also by the actual distributions of the batch sizes.

7. Discussions and conclusions

In this paper, we presented, for the first time, a discrete time model and associated algorithmic aspects for a $GI^X/G^Y/1$ system. We showed that the system can be studied as QBD in a practical situation by simply re-blocking the matrices. We then presented some numerical examples that seem to follow what we expected, except when the traffic intensity is high in the case of the effect of variances of batch sizes on the mean queue length. The case of high traffic intensity needs to be studied further. We also studied the age process of a special case of this model and derived results for obtaining the probability distributions of the age of a leading customer in service and the sojourn times of different customers.

The results obtained in this paper can be used to study several single server systems in telecommunications and manufacturing systems. In the past the $GI^X/G^Y/1$ systems have been approximated by simpler models because of the difficulties involved in developing appropriate algorithms. The algorithm presented is easy to implement. Hence the results obtained in this paper have potentials for several applications in real life problems.

Even though the algorithm is easy to implement, for real practical problems the size of the block matrices could be huge and thus lead to a dimensionality problem. In such situations one may capitalize on the special structures of the matrices that represent the interarrival time and service time distributions and use that to reduce the computational efforts. One interesting aspect for further research is how to improve on computational efficiency in the case of large size problems.

Acknowledgements

The authors acknowledge Qian Wang for her assistance with writing the Matlab code for this problem and plotting the graphs. The research of the authors was partially supported by grants from the Natural Sciences and Engineering Research Council (NSERC) of Canada.

References

[1] A.S. Alfa, The combined elapsed time and matrix-analytic approach for the $GI/G/1$ and the $GI^X/G/1$ systems, Queueing System 45 (2003) 5–25.
[2] A.S. Alfa, W. Li, Matrix-geometric solution of the discrete time $GI/G/1$ system, Stochastic Models 17 (4) (2001) 541–554.
[3] A.S. Alfa, Time-inhomogeneous bulk server queue in discrete time: A transportation type problem, Operations Research 30 (4) (1982) 650–658.
[4] T.P. Bagchi, J.G.C. Templeton, A note on the $M^X/G^Y/1$, K bulk queueing system, Journal of Applied Probability 10 (1973) 901–906.
[5] U.N. Bhat, Imbedded Markov chain analysis of single server bulk queues, Journal of Australian Mathematical Society 4 (1964) 244–263.
[6] H.R. Gail, S.L. Hantler, B.A. Taylor, Non-skip-free $M/G/1$ and $GI/M/1$ types of Markov chains, Advances in Applied Probability 29 (3) (1997) 733–758.
[7] J. Keilson, The general bulk queue as a Hilbert problem, Journal of Royal Statistical Society Series B 24 (1962) 344–358.
[8] G. Latouche, V. Ramaswami, Introduction to Matrix Analytic Methods in Stochastic Modelling, in: ASA-SIAM Series on Statistics and Applied Probability, SIAM, Philadelphia, PA, 1999.
[9] K.T. Marshall, Bounds for some generalizations of the $GI/G/1$ queue, Operations Research 16 (4) (1968) 841–848.
[10] R.G. Miller, A contribution to the theory of bulk queues, Journal of Royal Statistical Society Series B 21 (1959) 320–337.
[11] M.F. Neuts, Matrix-Geometric Solutions in Stochastic Models, The John Hopkins University Press, 1981.
[12] M.F. Neuts, Structured Stochastic Matrices of M/G/1 Type and their Applications, Marcel Dekker, New York, 1989.
[13] W.B. Powell, P. Humblet, The bulk service queue with a general control strategy: Theoretical analysis and a new computational procedure, Operations Research 34 (1986) 27–275.
[14] V. Ramaswami, Stable recursion for the steady state vector in Markov chains of $M/G/1$ type, Stochastic Models 4 (1988) 183–188.
[15] D.H. Shi, J. Guo, L. Liu, SPH-distributions and rectangle-iterative algorithm, in: S. Chakravarthy, A.S. Alfa (Eds.), Matrix-Analytic Methods in Stochastic Models, Marcel Dekker, New York, 1996, pp. 207–224.

[16] H.P. Simao, W.B. Powell, Waiting time distributions for transient bulk queues with general vehicle dispatching strategies, Naval Research Logistics 35 (1988) 285–206.

Attahiru S. Alfa is NSERC Industrial Research Chair of Telecommunications and Professor, Department of Electrical and Computer Engineering at the University of Manitoba, Winnipeg, Manitoba, Canada. He obtained his B. Eng. from Ahmadu Bello University in Nigeria, MSc. from the University of Manitoba and Ph.D. from the University of New South Wales in Australia. Dr. Alfa carries out research in the areas of queueing and network theories with applications mostly to telecommunication systems. He has also applied these theories to manufacturing and transportation and traffic systems in the past. His current research interests are in the area of wireless communication networks, mobility, Internet traffic, stochastic models, performance analysis, network restoration, and teletraffic forecasting models. He has contributed significantly in the area of matrix-analytic methods for stochastic models. He has published in several journals, and most recently in Stochastic Models, Queueing Systems - Theory and Applications, Naval Research Logistics, IEEE Journal on Selected Areas in Communications, IEEE Transactions on Vehicular Technology, IEEE Transaction on Wireless, IEEE Transaction on Mobile Computing, IEEE Transaction on Communications, IEEE Transaction on Parallel and Distributed Systems, Performance Evaluation, Journal of Applied Probability, Advances in Applied Proabability, Mathematics of Computations and Numerische Mathematik. He belongs to the following organizations: APEGM, IEEE, and INFORMS.

Qi-Ming He is a professor in the Industrial Engineering Department of Dalhousie University. His main research areas are algorithmic methods in applied probability, queueing theory, inventory control, stochastic modelling, and supply chain management. Recently, he is working on queueing systems with multiple types of customers, inventory systems with multiple types of demands, and the characterization of phase-type distributions.