Analysis of finite-buffer discrete-time batch-service queue with batch-size-dependent service

A. Banerjee, U.C. Gupta, V. Goswami

A. Banerjee, U.C. Gupta, V. Goswami

Department of Mathematical Sciences, Indian Institute of Technology (BHU), Varanasi 221005, India
Department of Mathematics, Indian Institute of Technology Kharagpur, Kharagpur 721302, India
School of Computer Application, KIIT University, Bhubaneswar 751024, India

Abstract

Over the last two decades there has been considerable growth in digital communication systems which operate on a slotted system. In several applications, transmission of packets over the network takes place in batches of varying size, and transmission time depends upon the size of the batch. Performance modelling of these systems is usually done using discrete-time queues. In view of this, we consider a single-server queue with finite-buffer in a discrete-time domain where the packets are transmitted in batches (of varying size) according to minimum and maximum threshold limit, usually known as general batch service rule. The transmission time (in number of slots) of these batches depends on the number of packets within the batch under transmission, and is arbitrarily distributed. We obtain, in steady-state, distribution of the number of packets waiting in the queue and in service (those being transmitted in batches). In addition, we also obtain average number of packets waiting in queue, in the system, with the server, rejection probabilities, etc. Finally, computational experiences with a variety of numerical results have been discussed by introducing a cost model which gives optimum value of the lower threshold limit.

1. Introduction

The study of queueing models in discrete-time environment have become popular in recent years due to its wide application in computer and telecommunication systems, Broadband Integrated Services Digital Network (B-ISDN), Asynchronous Transfer Mode (ATM). A detailed discussion and application of discrete-time queues can be found in books by Takagi (1993), Bruneel and Kim (1993), Woodward (1994) and Alfa (2010). For a review and recent development on discrete time queue, see e.g. (Artalejo, Atencia, & Moreno, 2005; Bruneel, Mélangé, Steyaert, Claeyts, & Walraevens, 2012; Chang & Choi, 2005; Chaudhry & Chang, 2004; Claeyts, Walraevens, Laevens, & Bruneel, 2010a; Claeyts, Walraevens, Laevens, & Bruneel, 2010b, 2011; Clercq, Laevens, Steyaert, & Bruneel, 2013; Feyaerts, Vuyts, Bruneel, & Wittevrongel, 2012; Fiems & Bruneel, 2013; Goswami & Mund, 2011; Goswami & Samanta, 2009; Gupta & Goswami, 2002; Hong Li, 2013; Hong Li & Shuo Tian, 2007; Samanta, Chaudhry, & Gupta, 2007; Samanta, Gupta, & Chaudhry, 2009; Tao, Zhang, Xu, & Gao, 2013; Yi, Kim, Yoon, & Chae, 2007).

The modern telecommunication networks have been designed to transfer voice, video, data and various other types of information services, which we usually refer as messages. These messages are first broken into packets (cells) and then they are transmitted over the network consisting of several nodes by establishing virtual connection. Each node behaves like a single/multiple server queue with finite/infinite buffer. It is seen at times that, an individual node operates either as a multi-server or sometimes as if a server offering batch service (which is due to simultaneous transfer of several packets) in queueing analogy. For example, point-to-multipoint communication is used most frequently in wireless Internet and IP telephony via gigahertz radio frequencies. The key components of 802.16 systems are a Base Station (BS) and a Subscriber Station (SS). A cell with point-to-multipoint communication is used most frequently in wireless Internet and IP telephony via gigahertz radio frequencies. The key components of 802.16 systems are a Base Station (BS) and a Subscriber Station (SS). A cell with point-to-multipoint structure can be constructed using the BS and one or more SSs. According to the bandwidth demand BS allocates variable number of physical slots to each SS. The application must initiate and establish a connection between a BS and SS before data is transmitted. In such systems, the data transfer takes place in uplink (SS to BS) and downlink (BS to SS) directions with the help of Time Division Multiple Access (TDMA). The time is divided between frames and each frame is broken into multiple time slots. The BS can dynamically allocate...
different time slots for downlink and uplink. In case of internet traffic, downlink gets higher preference over the uplink. However, in certain applications, like VOIP, the same amount of slots is allocated for both uplink and downlink. As the BS transmits data in batches having certain threshold limits on the size of batches and makes transmission depending on the size of the transmitted batches to a SS (using TDMA) the system can be modeled and analyzed as if a batch service queue. This scheme of transmission of packets (e.g., from BS to SS) in batches enhances the overall efficiency of the system as well as improves the quality of service (QoS). The virtual diagram of this information transmission system is displayed in Fig. 1.

In view of this, several researchers have carried out analysis of batch service queues in discrete time domain, see e.g., Chaudhry and Chang (2004), Samanta et al. (2007), Yi et al. (2007) and Samanta et al. (2009). In particular, Gupta and Goswami (2002) analyzed a discrete time finite-buffer batch-service queue with “general bulk service (a, b)” (GBS) rule and obtained queue length distribution at various epochs under both arrival-first (AF) and departure-first (DF) management policies (Gravey, Louvion, & Boyer, 1990). In all these studies, it is assumed that service time of batches remains the same irrespective of the size of the batch under service.

Another interesting service discipline in batch-service queue which has received considerable attention in recent years, is batch-service queue with batch-size-dependent service. Here service times of the batches depend on the size of the batch. The analysis of such queue requires a fresh analysis because the queue length analysis of batch-service queues carried out by all other researchers mentioned above and in particular by Gupta and Goswami (2002) will not yield the desired information. For example, from their analysis one cannot get the distribution of number of customers being served in a batch at arbitrary slot boundary. Unless and until we have such information one cannot use their model to apply batch-size-dependent service. However, some work in this direction under continuous time setup has been carried out recently by Bar-Lev, Parlar, Perry, Stadje, and der Duyn Schouten (2007), Chaudhry and Gai (2012), Banerjee and Gupta (2012) and Banerjee, Gupta, and Sikdar (2013). In a series of papers, Claey’s, Steyaert, Walraevens, Laeens, and Bruneel (2013a, 2013b) carried out analysis of similar queues under discrete time setup by assuming infinite buffer space and obtained various performance measures. To the best of authors’ knowledge no such queueing model involving discrete-time batch-service queue with finite-buffer under batch-size-dependent service has been discussed so far in the literature.

In this paper, we analyze a single-server finite-buffer discrete-time batch-service queue where server serves customers in batches according to the GBS rule. The interarrival times of customers are independent and geometrically distributed, and the service times of the batches are arbitrarily distributed and depend on the size of the batch undergoing service. We denote this model by Geo/C,a,b/1/N queue, where N is the queue capacity. The focus of this paper is to present the theoretical as well as computational aspects of this queue. More specifically, we obtain the joint distribution of the number of customers in the queue and the number with the server at an arbitrary slot. Several performance measures of interest such as average number of customers waiting in queue, in the system, with the server, probability of blocking, and average waiting time of a customer in the queue as well as in the system have been obtained. Finally, computational experiences with a variety of numerical results are discussed along with a cost model which gives the optimum value of lower threshold value of batch size u.’

The rest of the paper is organized as follows. In Section 2, description of the model and its solution procedure is given which includes evaluation of departure epoch probabilities, and the relations between the state probabilities at departure- and arbitrary-epochs. System performance measures and numerical results are given in Sections 3 and 4, respectively. The cost model is discussed in Section 4.1.1. The paper ends with conclusions and future scope.

2. The model description

Let us assume that the time is slotted in intervals of equal length with length of a slot being unity. Further, let the time axis be marked by 0, 1, ..., m, ..., and a potential arrival and a departure takes place around a slot boundary. More specifically we assume that a potential customer arrives in the interval (m−, m) and a potential departure occurs in the interval (m, m+). It is equivalent to the arrival first (AF) policy or late arrival system with delayed access (LAS-DA) (Gravey et al., 1990; Hunter, 1983).

We consider a single-server queue under the above setup and assume that the interarrival times (A) of customers (packets) are independent and geometrically distributed with probability mass function (pmf) \( a_n = P(A = n) = \lambda^{n-1}e^{-\lambda}, \ 0 < \lambda < 1, \ n \geq 1, \) and \( \lambda = 1 - \overline{\lambda}. \) The customers (packets) are served (transmitted) by a

---

**Fig. 1.** Point to multipoint based IEEE 802.16 system architecture.
single-server in batches of maximum size ‘b’ with a minimum threshold value ‘a’, however, if number of customers (packets) in the queue at the beginning of the transmission (service) is less than the minimum threshold value ‘a’, the server waits until the number in the queue hits ‘a’. If at the beginning of the service ‘b’ or more customers are present in the queue, only ‘b’ of them are taken into for service. The service times of the batches depend on the size of the batches, i.e., the service times of the batches of size \( r \) for \( a \leq r \leq b \), are independent random variables with pmf \( h_i(k) = P(S_k = k), \) \( k = 0, 1, 2, \ldots, \) with \( h_0(k) \equiv 0 \), and \( \sum_{k=0}^{a} h_i(k) \equiv 1 \) for \( a \leq r \leq b \). Further, let \( h_i^{\prime}(z) = \sum_{k=0}^{a} h_i(k) z^k \), is probability generating function (pgf) of \( h_i(k) \) and \( h_i^{\prime}(1) \) is the mean service time of a batch of size \( r \), where \( h_i^{\prime\prime}(z) = \left( h_i^{\prime}(z) \right) \) is the derivative of \( h_i(z) \) evaluated at \( z = 1 \). The system has finite waiting (queue) space of size \( N(> b) \).

In the following subsections we obtain the distributions of the number of customers in the queue and the number with the server at various time epochs.

### 2.1. Joint distribution of the number of customers in the queue and number with the departing batch at departure-epoch

In this section, we obtain the joint distribution of the number of customers in the queue and number with the departing batch at departure-epoch. Towards this end let us consider \( p_{n,x} \) \((0 \leq n \leq N, a \leq r \leq b) \) be the probability that there are \( n \) customers in the queue immediately after the departure of a batch of size \( r \). The unknown quantities \((p_{n,x})\) can be obtained by solving the system of equations \( p = \pi \), where \( \pi = (\pi_0, \pi_1, \ldots, \pi_n) \) and each \( \pi_n \) \((0 \leq n \leq N) \) is a row vector of order \((b-a+1)\), i.e., \( \pi_n = (p_{n0}, p_{n1}, \ldots, p_{nb}) \), \( 0 \leq n \leq N \), and \( P \) is the one-step transition probability matrix (TPM) given by

\[
P =  \begin{pmatrix}
0 & D_0 & D_1 & D_2 & \cdots & D_{b-1} & D_b & \cdots & D_{b-1} & D_b & \cdots & D_{b-1} \\
1 & D_0 & D_1 & D_2 & \cdots & D_{b-1} & D_b & \cdots & D_{b-1} & D_b & \cdots & D_{b-1} \\
\vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
0 & D_0 & D_1 & D_2 & \cdots & D_{b-1} & D_b & \cdots & D_{b-1} & D_b & \cdots & D_{b-1} \\
0 & 0 & \cdots & 0 & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & 0 & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
\end{pmatrix}
\]

where each \( 0 \) and \( D_0 \) are matrices of dimension \((b-a+1)\) and are given as:

\[
D_0^i = e_i^T \otimes \mathcal{K}, \quad 1 \leq i \leq b-a+1, \quad 0 \leq j \leq N-1, \quad i \leq j
\]

\[
D_0^i = e_i^T \otimes \mathcal{K} + e_i^\prime \otimes \mathcal{K}, \quad 1 \leq i \leq b-a, \quad i \leq j
\]

\[
D_0^{i-j} = e_i^\prime \otimes \mathcal{K}, \quad b < j \leq N
\]

with

- \( e_i \) is a column vector of dimension \((b-a+1)\) with 1 at ith position and 0 elsewhere.
- \( \mathcal{K} \) is a column vector of dimension \((b-a+1)\) consisting of \( e_i^T \)'s and \( e_i^\prime \)'s are the probabilities of \( j \) arrivals during the service period of a batch of size \( r \) \((a \leq r \leq b)\) and is given by,

\[
e_j^\prime = \frac{h_i(k)}{j} x^{j-1}, \quad j \geq 0.
\]

- \( \mathcal{K} \) is a column vector of dimension \((b-a+1)\) consisting of \((1 - \sum_j e_j^T)\)’s.

### 2.2. Joint distribution of the number of customers in the queue and number with the server at an arbitrary slot

As we observe the system just before the beginning of a slot boundary, let us define the state of the system just before a potential arrival (at \( m \)) by the following random variables:

- \( N_q(m) \equiv \) Number of customers in the queue waiting for service,
- \( S(m) \equiv \) Number of customers with the server, and
- \( U(m) \equiv \) Remaining service time of the batch in service (if any).

Further, we define the joint probabilities

\[
p_{n,0}(m) = \Pr\{N_q(m) = n, S(m) = 0\}, \quad 0 \leq n \leq a-1,
\]

\[
p_{n,i}(u, m) = \Pr\{N_q(m) = n, S(m) = r, U(m) = u\}, \quad u \geq 1, \quad 0 \leq n \leq N, \quad a \leq r \leq b.
\]

In limiting case we define

\[
p_{n,0} = \lim_{m \rightarrow \infty} p_{n,0}(u) \quad \text{and} \quad p_{n,i} = \lim_{m \rightarrow \infty} p_{n,i}(u).
\]

Now relating the state of the system at \( m \) and \((m+1)\), and using probabilistic argument, we have in steady-state following difference equations for \( u \geq 1 \):

\[
p_{n,0} = \lambda p_{n,0} + \lambda \sum_{i=a}^{b} p_{n,i}(1),
\]

\[
p_{n,i}(u) = \lambda p_{n,i}(u+1) + \lambda p_{n-1,i}(u+1) + \lambda p_{n-1,i}(u), \quad 1 \leq n \leq a-1,
\]

\[
p_{n,0}(u) = \lambda p_{n,0}(u+1) + \lambda p_{n-1,0}(u+1) + \lambda p_{n-1,0}(u), \quad a \leq r \leq b,
\]

\[
p_{n,i}(u) = \lambda p_{n,i}(u+1) + \lambda p_{n-1,i}(u+1), \quad a \leq r \leq b-1,
\]

\[
p_{n,0}(u) = \lambda p_{n,0}(u+1) + \lambda p_{n-1,0}(u+1) + \lambda p_{n-1,0}(u), \quad 1 \leq n \leq N-1,
\]

\[
p_{n,i}(u) = \lambda p_{n,i}(u+1) + \lambda p_{n-1,i}(u+1) + \lambda p_{n-1,i}(u), \quad a \leq r \leq b,
\]

\[
p_{n,0}(u) = \lambda p_{n,0}(u+1) + \lambda p_{n-1,0}(u+1) + \lambda p_{n-1,0}(u), \quad 1 \leq n \leq N-1,
\]

\[
p_{n,i}(u) = \lambda p_{n,i}(u+1) + \lambda p_{n-1,i}(u+1) + \lambda p_{n-1,i}(u), \quad 1 \leq n \leq N-1.
\]
\[ p_{n;b}(u) = \frac{n}{2} p_{n,1}(u + 1) + \lambda p_{n-1;1}(u + 1), \quad N - b + 1 \leq n \leq N - 1, \]  
\[ p_{n;r}(u) = p_{n,1}(u + 1) + \lambda p_{n-1;1}(u + 1), \quad a \leq r \leq b. \]  
\[ \sum_{n=0}^{a-1} p_{n,0}(u) z^n. \]  
Let us further define the probability generating function of \( p_{n;r}(u) \) as  
\[ \hat{p}_{n;r}(z) = \sum_{n=0}^{\infty} p_{n;r}(u) z^n, \]  
so that, \( p_{n;r} = \hat{p}_{n;r}(1) \).  
The normalizing condition is given by  
\[ \sum_{n=0}^{a-1} p_{n,0} + \sum_{n=a}^{b} p_{n,r} = 1. \]  
Now our main objective is to obtain \( p_{n,0} \) \((0 \leq n \leq a - 1)\) and \( p_{n;r} \) \((a \leq r \leq b, 0 \leq n \leq N)\). As we have already obtained \( p_{n,0}^b \), we need to establish relation between \( p_{n,0} \), \( p_{n;r} \), and \( p_{n;r}^b \). Towards this end, multiplying \((7)-(13)\) by \( z^n \) and summing over \( u \) from 1 to \( \infty \), we obtain  
\[ \left( \frac{z - 1}{z} \right) \hat{p}_{n,0}(z) = \lambda \hat{p}_{n-1;1}(z) + \lambda \hat{p}_{n-1;r}(z) \sum_{i=0}^{b} p_{n;i}(1) \]  
\[ + \lambda \hat{p}_{n-1;r}(2)(1) - \hat{p}_{n,0}(1), \]  
\[ \left( \frac{z - 1}{z} \right) \hat{p}_{n;r}(z) = \lambda \hat{p}_{n-1;1}(z) + \lambda \hat{p}_{n-1;r}(z) \sum_{i=0}^{b} p_{n-1;i}(1) - \hat{p}_{n;r}(1), \]  
\[ a + 1 \leq r \leq b, \]  
\[ a \leq r \leq b - 1, 1 \leq n \leq N - 1. \]  
Summing \((23)\) and \((16)-(21)\), we obtain  
\[ 1 - \sum_{n=0}^{a-1} p_{n,0} = \frac{h_n^b(z) - 1}{z - 1} - \sum_{n=0}^{a-1} \sum_{r=0}^{b} p_{n;r}(1) \]  
\[ + \sum_{n=0}^{b-1} \sum_{r=0}^{b} \frac{\lambda}{z - 1} \hat{p}_{n;1}(z) + \hat{p}_{n;r}(z) - 1 \]  
\[ \hat{p}_{n;r}(1), \]  
\[ \frac{h_n^b(z) - 1}{z - 1} - \sum_{n=0}^{b-1} \sum_{r=0}^{b} p_{n;r}(1). \]  
Taking limit as \( z \to 1 \) in the above expression, we get  
\[ 1 - \sum_{n=0}^{a-1} p_{n,0} = \frac{h_n^b(1)}{a} - \sum_{n=0}^{a-1} \sum_{r=0}^{b} p_{n;r}(1) \]  
\[ + h_b^b \sum_{n=0}^{b} p_{n;r}(1). \]  
Now we establish relation between \( p_{n;r}^b \) and \( p_{n;r}(1) \). One may note here that \( p_{n;r}(1) \) is the joint probability that there are \( n \) customers in the queue, \( r \) with the server, and remaining service time is just one slot. Therefore \( p_{n;r}^b \) and \( p_{n;r}(1) \) are connected by the following relations  
\[ p_{n;r}(1) = \frac{d}{dt} p_{n;r}(1), \]  
\[ a \leq r \leq b, \]  
\[ p_{n;r} = d(p_{n;r}(1) + p_{n;r-1}(1)), \]  
\[ a \leq r \leq b, 1 \leq n \leq N - 1, \]  
\[ p_{n;r} = d(p_{n;r}(1) + p_{n;r+1}(1)), \]  
\[ a \leq r \leq b, \]  
\[ \text{where } d^{-1} = \sum_{n=0}^{N} \sum_{r=0}^{b} p_{n;r}(1). \]  
Further, we denote the probability of \( n \) customers in the queue at departure-epoch by  
\[ p_n^a = \sum_{r=0}^{b} p_{n;r}(1), \]  
\( 0 \leq n \leq N. \]  
The following lemma gives an alternative expression for \( d \) which will used in \( \text{Theorem 1}. \)  

**Lemma 1.** The value of \( d \), as appearing in \((26)-(28)\), is given by  
\[ d^{-1} = g^{-1} \left( 1 - \sum_{n=a}^{a-1} p_{n,0} \right), \]  
\[ \text{where } g = h_a \sum_{n=0}^{a-1} p_{n,1} + \sum_{n=a}^{b} h_n p_n^b + h_b \sum_{n=0}^{a-1} p_{n,1}^b. \]

**Proof.** On dividing \((25)\) by \( \sum_{n=a}^{N} \sum_{r=0}^{b} p_{n;r}(1) \) and after some simple algebraic manipulations with the help of \((26)-(29)\), we get  
\[ d(1 - \sum_{n=0}^{a-1} p_{n,0}) = g. \]

**Theorem 1.** The steady-state probabilities \( \{p_{n,0}; p_{n;r}\} \) and \( \{p_{n;r}; p_{n}^b\} \) are related by  
\[ p_{n,0} = E^{-1} \left( \sum_{j=0}^{a} p_{j}^b \right), \]  
\( 0 \leq n \leq a - 1, \]  
\[ p_{n,a} = E^{-1} \left( \sum_{j=0}^{a} p_{j}^b - \sum_{j=0}^{a} p_{j} \right), \]  
\( 0 \leq n \leq N - 1, \]  
\[ p_{n;r} = E^{-1} \left( \sum_{j=0}^{r-1} p_{j}^b \right), \]  
\( a + 1 \leq r \leq b - 1, 0 \leq n \leq N - 1, \]  
\[ p_{n;b} = E^{-1} \left( \sum_{j=0}^{b} p_{j} - \sum_{j=0}^{a} p_{j}^b \right), \]  
\( 0 \leq n \leq N - 1, \]  
\[ \text{where } E = i g + \sum_{n=0}^{a-1} (a - i) p_{i}^b \]  
and \( g \) is defined in **Lemma 1**.  

**Proof.** On dividing Eqs. \((5)\) and \((6)\) by \( \sum_{n=a}^{N} \sum_{r=0}^{b} p_{n;r}(1) \) and using \((26)-(30)\) we obtain the desired result \((31)\) after certain algebraic manipulations.
Now setting $z = 1$ in (23), and (16)–(20), and dividing them by $\sum_{n=0}^{N-b} p_n(1)$ and using (26)–(30) we obtain the desired result (32)–(34) after certain algebraic manipulation. □

Now the only unknown term is $p_{nr}$ ($a \leq r < b$) which can be obtained as follows:

Differentiate Eq. (21) with respect to $z$ and then setting $z = 1$, we get

$$p_{nr} = \frac{\partial}{\partial z} p_{nr}(1), \quad a \leq r < b,$$

(35)

where $p_{nr}(1)$ is the derivative of $p_{nr}(z)$ with respect to $z$ at $z = 1$. Now to get $p_{nr}(1)$, differentiate Eqs. (15)–(20) with respect to $z$ and set $z = 1$. So that

$$\frac{\partial}{\partial z} p_{nr}(1) = \lambda h_a p_{nr} + \lambda h_r \sum_{i=1}^{b} p_{nr}(1) + \lambda h_b \sum_{i=1}^{a} p_{nr-1}(1) - p_{nr}, \quad a + 1 \leq n \leq b,$$

(36)

$$\frac{\partial}{\partial z} p_{nr}(1) = \lambda h_n \sum_{i=1}^{b} p_{nr}(1) + \lambda h_r p_{nr}(1) - p_{nr}, \quad a \leq r < b - 1, 1 \leq n < N - 1,$$

(37)

$$\frac{\partial}{\partial z} p_{nr}(1) = \lambda p_{nr}(1) - p_{nr}, \quad a \leq r < b,$$

(38)

$$\frac{\partial}{\partial z} p_{nr}(1) = \lambda h_n \sum_{i=1}^{b} p_{nr-1}(1) - p_{nr-b}, \quad 1 \leq n \leq N - b - 1,$$

(39)

$$\frac{\partial}{\partial z} p_{nr}(1) = \lambda p_{nr-1}(1) + \lambda h_b \sum_{i=1}^{a} p_{nr}(1) - p_{nr-b}, \quad N - b + 1 \leq n \leq N - 1,$$

(40)

As all $p_{nr}$'s are known completely, some other distributions of interest can be easily obtained and are listed as follows:

- Distribution of the number of customers in the system (including number of customers with the server) is given by

$$p_n^{\text{stem}} = \begin{cases} p_{n0} + \sum_{r=0}^{\min(n,N-b)} p_{nr} & 0 \leq n \leq a - 1, \\ \sum_{r=\max(n-N,a)}^{b} p_{nr} & a \leq n \leq N + b. \end{cases}$$

(42)

- Distribution of the number of customers in the queue is given by

$$p_n^{\text{queue}} = \begin{cases} p_{n0} + \sum_{i=1}^{b} p_{nr} & 0 \leq n \leq a - 1, \\ \sum_{i=\max(n-N,a)}^{b} p_{nr} & a \leq n \leq N. \end{cases}$$

(43)

- Distribution of the number of customers in service given server is busy is given by

$$p_n^{\text{service}} = \sum_{r=0}^{N-b} p_{nr} \quad a \leq r < b, \quad \text{where} \quad \alpha = a - 1 + \sum_{n=0}^{a-1} p_{n0}.$$

- Distribution of the number of customers in the queue when server is busy is given by

$$p_n^{\text{bus}} = \sum_{i=0}^{b} p_{nr} \quad 0 \leq n \leq N.$$

(44)

### 2.3. Computational procedure for evaluating the joint distributions

In this section we provide the stepwise procedure for evaluating the joint distributions $p_{n,r}$ and $p_{n,r}$ [$0 \leq n \leq N, a \leq r < b$].

#### Step 1. Input data:

(a) Buffer-size ($N$), lower threshold limit ($a$), upper threshold limit ($b$), arrival-rate ($\lambda$).

(b) Specify the pmf of service-time, $h_1(k), h_2(k), \ldots, h_b(k)$. One may also specify the same service time distribution for all $a \leq r < b$.

#### Step 2. Compute $\pi_{0}^{n}(a \leq r < b, j > 0)$ using Eq. (4), and generate the column vectors $\pi_k$ and $\pi_\Lambda$.

#### Step 3. Construct the matrices $D_{nj}^{\text{sys}}$'s using (1)–(3).

#### Step 4. Calculate one-step transition probability matrix $\mathbf{P}$ as given in Section 2.1.

#### Step 5. Solve the system of linear equation $\pi \mathbf{P} = \pi$ to obtain $p_{nr}$ using software packages, e.g., MAPLE or MATLAB, etc. ($\pi$ is defined in Section 2.1.).

#### Step 6. Calculate $p_{n0} = \sum_{a=0}^{b} p_{n,a}$, where $a \leq n \leq N - 1$ using (31)–(34).

#### Step 7. Calculate $p_{nr}$ using Eqs. (35)–(41), recursively.

#### Step 8. Calculate $p_{n,\text{stem}}$, $p_{n,\text{queue}}$, $p_{n,\text{service}}$ and $p_{n,\text{bus}}$ using (42)–(44).

### 3. Performance measures

The performance measures of a queueing system usually reflect both the qualitative and quantitative aspects of the concerned model. It provides the system designer and analyst a powerful tool in making decisions and judging the efficacy of the concerned system in terms of quality (QoS). The major performance measures for the present model are: average buffer content in the queue ($L_q$), average buffer content in the system ($L_s$), average number of customers with the server when server is busy ($L_{server}$), average delay of a customer in the queue ($W_q$), average delay of a customer in the system ($W_s$) and the probability of blocking ($PBL$). These measures may also facilitate in making decisions, in setting trade off, among the costs associated with the service and the waiting time in the system. As the state probabilities are known, these can be easily obtained and are given below:

$$L = \sum_{n=0}^{N+b} n p_n^{\text{stem}}, \quad L_q = \sum_{n=0}^{N-b} n p_n^{\text{queue}}, \quad L_{server} = \sum_{r=a}^{b} r p_r^{\text{service}},$$

$$PBL = p_{n,\text{bus}}, \quad W_s = \frac{L_s}{Z}, \quad W_q = \frac{L_q}{Z},$$

where $Z = (1 - PBL) \times$ the effective arrival rate.

### 4. Numerical results

We have carried out extensive numerical work by considering variety of service time distributions and using the analytic procedure described in previous sections. However, we present here only a few of them. In Table 1 we present the joint distribution of the number of customers in the queue and number with the departing batch at departure epoch ($p_{n,r}$) for Geo/$D^{(2)}/1/125$ queue with $h_{s,30} = 1, h_{s,31} = 1, h_{s,34} = 1, h_{s,36} = 1, h_{s,10.38} = 1, h_{i,11.40} = 1, h_{i,12.42} = 1$ and $Z = 0.25$. In Table 2 we present the joint distribution of the number of customers in the queue and the number with the server at an arbitrary slot ($p_{n,0}$, $p_{n,r}$). The last row of Table 2 displays useful performance measures. The purpose of these results is to show as how the analytic results developed in previous sections can be implemented in order to get various distributions and performance measures.

Table 3 displays all performance measures of Geo/Geo/$^{(2)}/1/N$ queue for varying buffer-size ($N$). The last row of this table displays
the correlations (COR) between these performance measures. For example, COR(L, W) indicates the correlation between L and W. It can be observed from the table that L and W are positively correlated which validates the well known Little's law, concerning these two measures. This is also due to the fact that L and W are linearly related as $L = \lambda (1 - PBL) W$. The factor $\lambda (1 - PBL)$ is the effective arrival rate which is always positive. So this is the case of perfect linear relationship between L and W and slope being positive, hence COR(L, W) will be approximately one. This is also true for $L_{server}$ and $L_q$ as bigger queue length will always keep server busy, and in case of batch service, server will always have more number of customers to serve. Hence COR($L_{server}$, $L_q$) will be positive. However, when we consider average number of customers in the system (L) and probability of blocking (PBL), they are negatively correlated. This result is also on the expected line in the sense that on an average more number of customers in the queue (when buffer size increases) will result in reduction in probability of blocking.

4.1. Optimal control problem

In this section we consider an optimal control problem for the queueing model discussed in this paper. In many applications it is important to know the optimal values of the key parameters $a$ and/or $b$ at the pre-implement stage of the model. We consider an optimal control problem which is dependent on the cost model discussed below.

4.1.1. Cost model

Here we first derive the expression for the total system cost (TSC) under the following assumptions:

- $C_a$ be the holding cost per customer in the queue per unit time. Thus in long run the average holding cost is $C_a \cdot L_q$.
- $C_L$ be the cost of idleness of the server per unit time. Thus the associated long run idleness cost is $(C_L \cdot \sum_{n=0}^{\infty} P_{n,0})$, since the server will be idle for the fraction $\sum_{n=0}^{\infty} P_{n,0}$ of time.
- $C_b$ be the operating (serving) cost per unit time. Thus in long run the average operating cost is $C_b \cdot L_{server}$.
- $C_r$ be the rejection cost per unit time. Thus in long run the average rejection cost is $C_r \cdot PBL$.

Then in long run the total system cost per cycle is given by

$$\text{TSC} = \lambda C_a + C_L \cdot L_q + C_b \cdot \sum_{n=0}^{\infty} P_{n,0} + C_c \cdot \sum_{n=0}^{\infty} P_{n,b} + C_r \cdot PBL.$$ 

As we do not have explicit analytic expressions for $L_q$, $L_{server}$ and $PBL$, the optimal value of $a$, for which TSC is minimum, can not be obtained analytically by using usual optimization techniques. Therefore we use numerical approach to obtain optimal value of $a$ such that TSC is minimum. For this we first fix $b$, $N$ and then we chose the value of $a$ so that TSC is minimum.

Using the procedure described above, a numerical example is considered below:

<table>
<thead>
<tr>
<th></th>
<th>Start-up cost ($C_a$)</th>
<th>$4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>Holding cost per customer in the queue ($C_b$)</td>
<td>$0.50$</td>
</tr>
<tr>
<td>3</td>
<td>Idleness cost ($C_L$)</td>
<td>$1$</td>
</tr>
<tr>
<td>4</td>
<td>Operating cost per unit time ($C_c$)</td>
<td>$5$</td>
</tr>
<tr>
<td>5</td>
<td>Rejection cost per unit time ($C_r$)</td>
<td>$3$</td>
</tr>
</tbody>
</table>

Table 4 displays the values of TSC for Geo/D/\(c^{18}\)/1/20 queueing model. Here we consider batch-size-dependent service policy with deterministic service time distribution, i.e., the number of slots required to serve a batch of size $r$ ($a \leq r \leq b$) is $(k + r - a)$.
Table 2
Joint distribution of the number of customers in the queue and the number with the server at an arbitrary slot ($p_{k|0}$, $p_{k|1}$).

<table>
<thead>
<tr>
<th>n</th>
<th>Idle</th>
<th>Busy</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p_{0</td>
<td>0}$</td>
<td>$p_{0</td>
</tr>
<tr>
<td>0</td>
<td>0.00000588</td>
<td>0.01792250</td>
</tr>
<tr>
<td>1</td>
<td>0.00068400</td>
<td>0.01798049</td>
</tr>
<tr>
<td>2</td>
<td>0.00039070</td>
<td>0.01773576</td>
</tr>
<tr>
<td>3</td>
<td>0.00000000</td>
<td>0.00000000</td>
</tr>
<tr>
<td>4</td>
<td>0.00000000</td>
<td>0.00000000</td>
</tr>
<tr>
<td>5</td>
<td>0.00000000</td>
<td>0.00000000</td>
</tr>
</tbody>
</table>

Fig. 2 displays the effect of a on TSC for fixed b = 15, k = 30. N = 25 under Cases 1 and 2. It is clear from the figure that as a increases TSC also increases, but it is lower in case 1 as compared to the case 2 when a ≤ 8. However, when a varies from 9 to 14, the values of TSC coincides for both the cases. Since our main objective is to minimize TSC, so from the figure one can conclude that it is better to use batch-size-dependent service.

5. Conclusions and future scope

In this paper, we have analyzed a discrete-time finite-buffer single server queue where the packets are transmitted in batches of varying size with a minimum and maximum threshold limit. The transmission times of the batches are assumed to be arbitrarily distributed and are contingent on the number of packets within the batch being transmitted. We have obtained the steady-state distribution of the number of packets waiting in the queue and

(k ∈ N). In column 2 to 6 the values of TSC have been displayed for five different values of k, viz., k = 20, 23, 26, 30, 35. From Table 4 one can find the optimal value of a for each for which TSC is minimum. These optimal values of a are indicated by bold numerals in the table. For example when k = 23 the optimal value of a is 5 and the minimum value of TSC is 35.7508.

Next we compare batch-size-dependent service with batch-size-independent service in order to examine the effect of these service policies on the cost function. Towards this end we consider deterministic service time distribution and following two cases.

Case 1. In this case the number of slots required to complete the service of a batch of size $r$ ($a ≤ r ≤ b$) is $(r - k - b)$, $k ∈ N$, i.e., batch-size-dependent service.

Case 2. In this case the number of slots required to complete the service of a batch of size $r$ ($a ≤ r ≤ b$) is $k$, $k ∈ N$, i.e., batch-size-independent service.
in service. Additionally, various performance measures, such as, average number of packets waiting in queue, in the system, with the server, and rejection probabilities are also obtained. We have also introduced a cost model associated with our model, which will also introduce a cost model associated with our model, which will

...characterisation of the customer delay in queuing system with batch arrivals and service discipline. Mathematical Methods of Operations Research, 72, 1–23.

...delay analysis of two batch-service queueing models with batch arrivals: Geo/G/1. 4OR, 8, 255–269.

...multi-class discrete-time queueing system under the FCFS discipline service. Annals of OP, 202, 59–73.

...delay analysis of two batch-service queuing models with batch arrivals: Geo/G/1. 4OR, 45, 528–549.

...Performance Evaluation, 68, 528–549.

...delay analysis of two batch-service queuing models with batch arrivals: Geo/G/1. 4OR, 45, 528–549.

...Performance Evaluation, 68, 528–549.

...characterisation of the customer delay in queuing system with batch arrivals and service discipline. Mathematical Methods of Operations Research, 72, 1–23.

...delay analysis of two batch-service queuing models with batch arrivals: Geo/G/1. 4OR, 8, 255–269.

...multi-class discrete-time queueing system under the FCFS discipline service. Annals of OP, 202, 59–73.

...delay analysis of two batch-service queuing models with batch arrivals: Geo/G/1. 4OR, 45, 528–549.

...Performance Evaluation, 68, 528–549.

...characterisation of the customer delay in queuing system with batch arrivals and service discipline. Mathematical Methods of Operations Research, 72, 1–23.

...delay analysis of two batch-service queuing models with batch arrivals: Geo/G/1. 4OR, 8, 255–269.

...multi-class discrete-time queueing system under the FCFS discipline service. Annals of OP, 202, 59–73.

...delay analysis of two batch-service queuing models with batch arrivals: Geo/G/1. 4OR, 45, 528–549.

...Performance Evaluation, 68, 528–549.

...characterisation of the customer delay in queuing system with batch arrivals and service discipline. Mathematical Methods of Operations Research, 72, 1–23.

...delay analysis of two batch-service queuing models with batch arrivals: Geo/G/1. 4OR, 8, 255–269.

...multi-class discrete-time queueing system under the FCFS discipline service. Annals of OP, 202, 59–73.

...delay analysis of two batch-service queuing models with batch arrivals: Geo/G/1. 4OR, 45, 528–549.

...Performance Evaluation, 68, 528–549.

...characterisation of the customer delay in queuing system with batch arrivals and service discipline. Mathematical Methods of Operations Research, 72, 1–23.

...delay analysis of two batch-service queuing models with batch arrivals: Geo/G/1. 4OR, 8, 255–269.

...multi-class discrete-time queueing system under the FCFS discipline service. Annals of OP, 202, 59–73.

...delay analysis of two batch-service queuing models with batch arrivals: Geo/G/1. 4OR, 45, 528–549.

...Performance Evaluation, 68, 528–549.

...characterisation of the customer delay in queuing system with batch arrivals and service discipline. Mathematical Methods of Operations Research, 72, 1–23.

...delay analysis of two batch-service queuing models with batch arrivals: Geo/G/1. 4OR, 8, 255–269.

...multi-class discrete-time queueing system under the FCFS discipline service. Annals of OP, 202, 59–73.

...delay analysis of two batch-service queuing models with batch arrivals: Geo/G/1. 4OR, 45, 528–549.

...Performance Evaluation, 68, 528–549.

...characterisation of the customer delay in queuing system with batch arrivals and service discipline. Mathematical Methods of Operations Research, 72, 1–23.

...delay analysis of two batch-service queuing models with batch arrivals: Geo/G/1. 4OR, 8, 255–269.

...multi-class discrete-time queueing system under the FCFS discipline service. Annals of OP, 202, 59–73.

...delay analysis of two batch-service queuing models with batch arrivals: Geo/G/1. 4OR, 45, 528–549.

...Performance Evaluation, 68, 528–549.

...characterisation of the customer delay in queuing system with batch arrivals and service discipline. Mathematical Methods of Operations Research, 72, 1–23.

...delay analysis of two batch-service queuing models with batch arrivals: Geo/G/1. 4OR, 8, 255–269.

...multi-class discrete-time queueing system under the FCFS discipline service. Annals of OP, 202, 59–73.

...delay analysis of two batch-service queuing models with batch arrivals: Geo/G/1. 4OR, 45, 528–549.

...Performance Evaluation, 68, 528–549.

...characterisation of the customer delay in queuing system with batch arrivals and service discipline. Mathematical Methods of Operations Research, 72, 1–23.

...delay analysis of two batch-service queuing models with batch arrivals: Geo/G/1. 4OR, 8, 255–269.

...multi-class discrete-time queueing system under the FCFS discipline service. Annals of OP, 202, 59–73.

...delay analysis of two batch-service queuing models with batch arrivals: Geo/G/1. 4OR, 45, 528–549.

...Performance Evaluation, 68, 528–549.

...characterisation of the customer delay in queuing system with batch arrivals and service discipline. Mathematical Methods of Operations Research, 72, 1–23.

...delay analysis of two batch-service queuing models with batch arrivals: Geo/G/1. 4OR, 8, 255–269.

...multi-class discrete-time queueing system under the FCFS discipline service. Annals of OP, 202, 59–73.

...delay analysis of two batch-service queuing models with batch arrivals: Geo/G/1. 4OR, 45, 528–549.

...Performance Evaluation, 68, 528–549.

...characterisation of the customer delay in queuing system with batch arrivals and service discipline. Mathematical Methods of Operations Research, 72, 1–23.

...delay analysis of two batch-service queuing models with batch arrivals: Geo/G/1. 4OR, 8, 255–269.

...multi-class discrete-time queueing system under the FCFS discipline service. Annals of OP, 202, 59–73.

...delay analysis of two batch-service queuing models with batch arrivals: Geo/G/1. 4OR, 45, 528–549.

...Performance Evaluation, 68, 528–549.