Computers & Industrial Engineering 75 (2014) 121–128

Contents lists available at ScienceDirect

Computers & Industrial Engineering

journal homepage: www.elsevier.com/locate/caie

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

A. Banerjee [a], U.C. Gupta [b,*], V. Goswami [c]

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

ARTICLE INFO

Article history:
Received 16 August 2013
Received in revised form 9 June 2014
Accepted 10 June 2014
Available online 15 July 2014

Keywords:
Batch-size-dependent service
Blocking probability
Discrete-time
Finite-buffer
Steady-state
Queue

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, Mélange, Steyaert, Claeys, & Walraevens, 2012; Chang & Choi, 2005; Chaudhry & Chang, 2004; Claeys, Walraevens, Laevens, & Bruneel, 2010a, Claeys, Walraevens, Laevens, & Bruneel, 2010b, 2011; Clercq, Laevens, Steyaert, & Bruneel, 2013; Feyaerts, Vuyst, 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 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

☆ This manuscript was processed by Area Editor Manoj Tiwari.

* Corresponding author. Tel.: +91 3222283654.
E-mail addresses: anuradha.mat@iitbhu.ac.in, anuradha.iitkgp@yahoo.in (A. Banerjee), umesh@maths.iitkgp.ernet.in (U.C. Gupta), veena.goswami@yahoo.in (V. Goswami).

http://dx.doi.org/10.1016/j.cie.2014.06.009
0360-8352/

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 have 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, Claeys, Steyaert, Walraevens, Laevens, 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 $Geo/G_r^{(a,b)}/1/N$ 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 'a'.

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 $0, 1, \ldots, m, \ldots$, 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 set up and assume that the interarrival times (A) of customers (packets) are independent and geometrically distributed with probability mass function (pmf) $a_n = P(A = n) = \bar{\lambda}^{n-1}\lambda, \quad 0 < \lambda < 1, \; n \geqslant 1,$ and $\bar{\lambda} = 1 - \lambda$. The customers (packets) are served (transmitted) by a

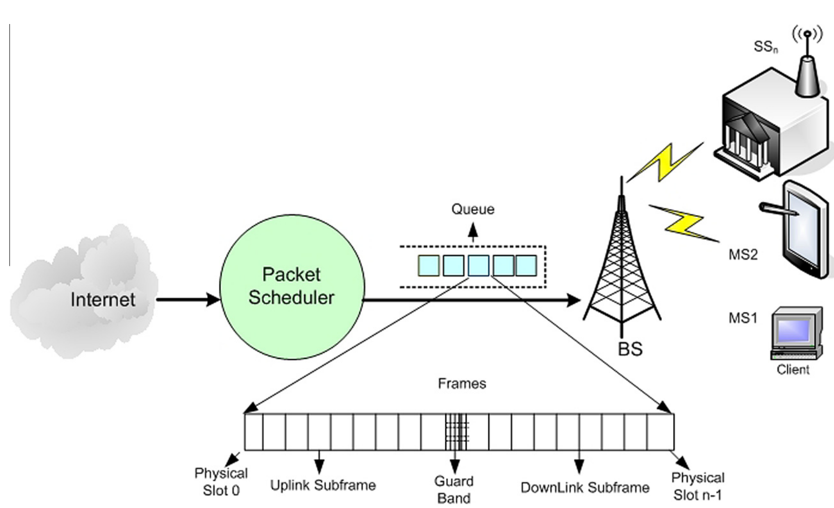

Fig. 1. Point to multipoint based IEEE 802.16 system architecture.