Computers & Operations Research 60 (2015) 138–149

Contents lists available at ScienceDirect

Computers & Operations Research

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

Analysis of a finite-buffer bulk-service queue under Markovian arrival process with batch-size-dependent service

A. Banerjee [a], U.C. Gupta [b,*], S.R. Chakravarthy [c]

[a] Department of Mathematical Sciences, Indian Institute of Technology (BHU), Varanasi 221005, India
[b] Department of Mathematics, Indian Institute of Technology, Kharagpur 721302, India
[c] Department of Industrial and Manufacturing Engineering, Kettering University, Flint MI 48504, USA

ARTICLE INFO

Available online 2 March 2015

Keywords:
Algorithmic probability
General bulk service rule
Markovian arrival process
Phase type distribution
Queue

ABSTRACT

We consider a finite capacity single server queue in which the customers arrive according to a Markovian arrival process. The customers are served in batches following a 'general bulk service rule'. The service times, which depend on the size of the batch, are generally distributed. We obtain, in steady-state, the joint distribution of the random variables of interest at various epochs. Efficient computational procedures in the case of phase type services are presented. An illustrative numerical example to bring out the qualitative nature of the model is presented.

1. Introduction

Bulk service queues have applications in many areas notably in production and manufacturing, transportation, package delivery, tourism, and amusement parks such as Disney World and Six Flags. For example, in production and manufacturing applications, it may not always be efficient to start the production as soon as the orders arrive. Instead, one has to wait until a certain number of orders are placed to start the process. Similarly, disposing hazardous petrochemicals and petroleum wastes (that arrive in containers or drums) may need thermal treatment using high temperatures, and hence processing them efficiently require grouping them with a minimum number and maximum number of containers. In the amusement park set up people queue up for going through various thrill rides and all of these are accommodated in groups of varying sizes with restrictions on the minimum and maximum numbers in each ride. In package delivery example, it is always better to fill the trucks (of varying sizes) to their capacity to balance the cost/efficiency to the delivery times of the packages. Thus, one can see the variety of scenarios that call for bulk service queues in real-life applications [1–6].

Queueing systems in which the services are offered in batches of varying sizes have been extensively studied in the literature [2–18]. In such queueing systems the customers are served in batches of sizes varying from a minimum size of '*a*' to a maximum size of '*b*' and this service rule is referred to as the 'general bulk service' (GBS) rule by Neuts [16]. The book by Chaudhry and Templeton [12] provides an in-depth study of bulk service queueing systems.

Thus, one can categorize the study of bulk service queueing systems as follows. The bulk service queueing systems in which the

1. buffer size is (a) finite or (b) infinite;
2. arrivals occur according to a (a) renewal or (b) correlated process;
3. arrivals occur (a) singly or (b) in batches;
4. services are (a) independent of the batch size or (b) dependent on the size of the batch being served;
5. service times are (a) exponential or (b) non-exponential.

While the literature is abundant with queueing systems dealing with cases falling in the intersection of (1), (2a), (3), (4a), and (5a), very few papers deal with case (2b) in combination with 4 (b) and 5(b).

Earlier work on bulk service queueing systems in which the services depend on the batch size includes Curry and Feldman [19] and Neuts [20]. Further, Bar-Lev et al. [3] and Chaudhry and Gai [21] both considered an $M/G_r^{(a,b)}/1$ queue and obtained the queue length distribution at departure epochs. One may note here that neither Bar-Lev et al. [3] nor Chaudhry and Gai [21] obtained the queue length distribution at an arbitrary epoch which is required to compute various performance measures. Also from their analysis one cannot obtain the joint distribution of the number of customers in the queue and the number in the departing batch at a departure epoch or the joint distribution of the number of customers in the queue and the number in the service at an arbitrary epoch. These distributions are needed in order to apply

* Corresponding author. Tel.: +91 3222 283654.
E-mail addresses: anuradha.iitkgp@yahoo.in, anuradha.mat@iitbhu.ac.in (A. Banerjee), umesh@maths.iitkgp.ernet.in (U.C. Gupta), schakrav@kettering.edu (S.R. Chakravarthy).

http://dx.doi.org/10.1016/j.cor.2015.02.012
0305-0548/

any batch-size-dependent service policy to a bulk service queueing system. In view of this Banerjee and Gupta [22] analyzed an $M/G_r^{(a,b)}/1/N$ queue and obtained all such distributions. Claeys et al. [4] analyzed an infinite buffer discrete-time bulk arrival and bulk service queue with GBS rule assuming the service time to be dependent on the size of the batch. They obtained the steady-state probability of the number in the system employing probability generating function approach. For recent development in this direction, see e.g., [23,24], and the references therein.

While the papers referenced above all dealt with renewal arrivals, there exists some literature on bulk service queueing systems [25–32] in which the customers arrive according to a versatile point process, namely, Markovian arrival process (*MAP*). The *MAP* is a rich class of point processes that includes many well-known processes such as Poisson process, phase type (PH) renewal processes, and Markov-modulated Poisson process. It has been observed that modern telecommunication networks/ATM networks support diverse traffics with different service characteristics such as voice, data and video. In such network, IP packets or cells of voice, video, data are sent over a common transmission channel on statistical multiplexing basis. These traffic streams are statistically multiplexed and transmitted in super high speed and also they are highly irregular (e.g., bursty and correlated). A good representation of such traffic is a Markovian arrival process. Further, *MAP* is very useful to model situations where the inter-arrival times are no longer independent. For example, in applications where the arrivals are generated from different sources the pooled input will not necessarily follow a renewal process (even if the individual sources are renewal). This is very common in production and manufacturing, computer communications, transportation, and other areas wherein the model under study is very much applicable. When the arrivals occur in batches in the context of *MAP*, we refer to that arrival process as *BMAP*. For more details on *BMAP* processes and their usefulness in stochastic modeling, we refer to [33,34], and for a review on *MAP* and *BMAP*, we refer the reader to [33,35–39].

Chakravarthy [27] and Gupta and Laxmi [28] both analyzed single server finite buffer bulk service queues with *MAP* arrivals and with GBS rule for the services. Chakravarthy [27] analyzed a finite-capacity $MAP/G^{(a,b)}/1$ queue where the upper threshold, 'b', is taken as the buffer size. Later, Gupta and Laxmi [28] considered a more general model than the one considered by Chakravarthy [27], and assumed that the buffer size, N, is greater than the upper threshold value 'b', and obtained the queue length distribution at various epochs. In the context of a multi-server queueing system with *MAP* arrivals, exponential group services with batches of sizes between two pre-determined thresholds, and a dynamic service rule that governs the size of the batch at the time of starting a service, Chakravarthy and Alfa [40] studied a bulk service queueing system using matrix-analytic methods. Recently, Banik [26] analyzed $BMAP/G^{(a,b)}/1/N$ and $BMAP/MSP^{(a,b)}/1/N$ queues and obtained the queue length distributions for both models. But in all of these papers the service times are assumed to be independent of the serving batch size.

In Chakravarthy and Alfa [41] a finite capacity queuing system with *MAP* arrivals attended by two exponential servers, who offer services in groups of varying sizes, is studied. Here the authors assume that the service times may depend on the number of customers in service. Efficient algorithmic procedures for computing the steady-state queue length densities and other system performance measures are discussed. A conjecture on the nature of the mean waiting time is proposed. Some illustrative numerical examples are presented. Alfa et al. [42] studied a discrete queueing system with *MAP* arrivals and PH services in batches of size ranging from 1 to a fixed threshold. Under a probabilistic service rule, which depends on the number of customers waiting in the queue, the authors show that the steady-state probability vector is (modified) matrix-geometric type [43]. Using a dynamic probabilistic rule associated with group (the size of which varies from a pre-determined threshold to the maximum buffer size) services, assuming *MAP* arrivals and exponential services whose parameter depends on the size of the batch, Chakravarthy and Bin [44] developed efficient algorithms for computing various performance measures such as throughput, mean number served, job overflow probability and server idle probability, useful in qualitative and quantitative interpretations of the model.

Most of the papers dealing with batch-size-dependent services assume the service times to be exponential except in Alfa et al. [42] where it is of PH type. But this one, as mentioned earlier, is studied in discrete time. In view of these, to fill the gap in the present literature, in this paper we study a continuous time finite capacity single server queue where arrivals occur according to a *MAP* with representation (C, D) of order m. The services are offered in batches of size varying from a minimum size a to a maximum value of b. The service time distribution of the batch is assumed to be arbitrarily distributed and dependent on the batch size. Henceforth, we denote this model by $MAP/G_r^{(a,b)}/1/N$. For analytic analysis of this model we use the embedded Markov chain technique and first obtain the joint distribution of the number of customers in the queue, the number in the departing batch, and the phase of the arrival process at a departure epoch. Then using the supplementary variable technique and considering the supplementary variable as remaining service time of the batch undergoing service, we develop relations between arbitrary and departure epoch probabilities in order to obtain the joint distribution of the number of customers in the queue, the number in service, and phase of the arrival process at an arbitrary epoch. Then we obtain arrival epoch probabilities by deriving relations between arbitrary and arrival epoch probabilities. Several performance measures of interest, such as mean number of customers in the queue (system), mean number of customers served in a batch, loss probability and mean waiting time in the queue (system), have been obtained.

For use in sequel, let $\boldsymbol{e}(r)$, $\boldsymbol{e}_j(r)$ and I_r denote, respectively, the column vector of dimension r consisting of 1's, column vector of dimension r with 1 in the jth position and 0 elsewhere, and an identity matrix of dimension r. When there is no need to emphasize the dimension of these vectors we will suppress the suffix. Thus, $\boldsymbol{e}$ will denote a column vector of 1's of appropriate dimension. The notation 'T' appearing in the superscript will stand for the transpose of a matrix; the notation '′' will stand for the derivative; and the symbols $\otimes$ and $\oplus$, respectively, will stand for the Kronecker product and Kronecker sum of two matrices. Thus, if A is a matrix of order $m \times n$ and if B is a matrix of order $p \times q$, then $A \otimes B$ will denote a matrix of order $mp \times nq$ whose (i,j)th block matrix is given by $a_{ij}B$; the Kronecker sum of two square matrices, say, G of order g and H of order h, is given by $G \otimes I + I \otimes H$, a square matrix of dimension gh. For more details on Kronecker products and sum, we refer the reader to Marcus and Minc [45].

The paper is organized as follows. In Section 2, the description of the model and its analysis at various epochs is given. Some key system performance measures to bring out the qualitative nature of the model are displayed in Section 3. The computational simplifications that arise when the services are of phase type are spelled out in Section 4, and an illustrative numerical example is discussed in Section 5. Some concluding remarks are presented in Section 6.

2. Model description and analysis

We consider a single server queueing system in which the customers arrive according to a *MAP* where the arrivals are governed by an underlying m-state Markov chain with transition rate c_{ij}, $1 \le i,j \le m$, $i \ne j$, there is a transition from state i to state j in the underlying Markov chain without an arrival, and with transition rate