Computers & Industrial Engineering 57 (2009) 748–761

Contents lists available at ScienceDirect

Computers & Industrial Engineering

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

Queueing analysis and optimal control of $BMAP/G^{(a,b)}/1/N$ and $BMAP/MSP^{(a,b)}/1/N$ systems

A.D. Banik *

School of Technology and Computer Science, Tata Institute of Fundamental Research, Homi Bhaba Road, Mumbai 400005, India

ARTICLE INFO

Article history:
Received 10 September 2008
Received in revised form 4 February 2009
Accepted 5 February 2009
Available online 13 February 2009

Keywords:
Finite-buffer
Queue
Batch Markovian arrival process
Markovian service process
General bulk service rule

ABSTRACT

We first consider a finite-buffer single server queue where arrivals occur according to batch Markovian arrival process (*BMAP*). The server serves customers in batches of maximum size '*b*' with a minimum threshold size '*a*'. The service time of each batch follows general distribution independent of each other as well as the arrival process. We obtain queue length distributions at various epochs such as, pre-arrival, arbitrary, departure, etc. Some important performance measures, like mean queue length, mean waiting time, probability of blocking, etc. have been obtained. Total expected cost function per unit time is also derived to determine the optimal value N^* of N at a minimum cost for given values of a and b. Secondly, we consider a finite-buffer single server queue where arrivals occur according to *BMAP* and service process in this case follows a non-renewal one, namely, Markovian service process (*MSP*). Server serves customers according to general bulk service rule as described above. We derive queue length distributions and important performance measures as above. Such queueing systems find applications in the performance analysis of communication, manufacturing and transportation systems.

1. Introduction

Bulk service queues have received considerable attention due to their wide applications in several areas including computer-communication, telecommunication, transportation and manufacturing systems. In many telecommunication systems, it is frequently observed that the server processes the packets in groups of random size. For example, in ATM networks with multiple input links where each link may serve messages that consist of several packets. Besides applications in telecommunication systems, bulk service queues have a wide range of applications in several areas including transportation systems, automatic manufacturing systems, etc. Chaudhry and Templeton (1983), Dshalalow (1997), etc. provide an extensive discussion of bulk-service systems. In such queues customers are served by a single server in batches of maximum size '*b*' with a minimum threshold size '*a*'. Such type of service rule is referred to as the general bulk service rule. In past several authors have analyzed similar kind of model described above, see e.g., Gold and Tran-Gia (1993), Chaudhry and Gupta (1999), Hébuterne and Rosenberg (1999), Chakravarthy (1992), etc.

Queueing models with non-renewal arrivals and service processes are often used to model such networks of complex computer and communication systems. In such systems, both the arrival and service processes may exhibit correlations which have significant impact on queueing behaviour. Lucantoni, Meier-Hellstern, and Neuts (1990) used Markovian arrival process (*MAP*) to capture correlation among the successive inter-arrival times. Similarly, to capture correlation among inter-batch arrival times Lucantoni (1991) introduces batch Markovian arrival process (*BMAP*), which is a convenient representation of Neuts (1979) versatile Markovian point processes. Like the *MAP*, Markovian service process (*MSP*) is a versatile service process which can capture correlation among successive service times. Several other service processes such as Poisson, Markov modulated Poisson, and PH-type renewal process can be considered as special cases of *MSP*. For details of *MSP* readers are referred to Bocharov (1996), Alberes-Velasco and Tajonar-Sanabria (2004), Gupta and Banik (2007), etc. More general non-renewal type of service process, e.g., semi-Markov process (*SM*) has been analyzed by several researcher. For example, $G/SM/1/\infty$ queueing system with vacations was considered by Machihara (1999). Recently, some studies have been done on departure process of $MAP/MSP/1$ type queueing systems, e.g., see Zhang, Heindl, and Smirni (2005) and references therein. In their paper, Zhang et al. (2005) have studied departure process of $BMAP/MAP/1$ queue using exact aggregate solution technique (called ETAQA truncation). Some studies of batch service queueing systems with *MAP* arrival have been done in past, see e.g., Chakravarthy (1993), Gupta and Vijaya Laxmi (2001), etc. Very few studies have been found on the batch service queue with a non-renewal service process, such as *MSP*. Recently, Banik, Chaudhry, and Gupta (2008) have analyzed $GI/BMSP/1/N$ queue in which service

* Tel.: +91 22 22782769.
E-mail addresses: banikad@gmail.com, abhijit@tcs.tifr.res.in

0360-8352/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.cie.2009.02.002

process *BMSP* refers to a variable batch-size service capacity rule. In the corresponding section we discuss the difference between *BMSP* and $MSP^{(a,b)}$ service rule.

This paper first carry out the analytic analysis of the $BMAP/G^{(a,b)}/1/N$ queue using the method of embedded Markov chain and the argument of total probability. As stated earlier, the service to the queueing system is provided in batches of minimum size a and maximum $b(1 \leqslant a \leqslant b \leqslant N)$. We obtain the steady-state queue length distribution at departure, arbitrary and pre-arrival epochs. Following that the construction of the expected cost function per unit time has been performed. An effective procedure is developed for searching a suitable thresholds N^* that minimizes the cost function for given values of a and b. Secondly, we consider $BMAP/MSP^{(a,b)}/1/N$ queue and obtain queue length distributions at departure, arbitrary and pre-arrival epochs. The queueing models discussed in this paper is more general in the sense that several other queueing models become special cases of the present one, e.g., $M^{[X]}/G^{(a,b)}/1/N$, $MAP/PH^{(a,b)}/1/N$, $PH^{[X]}/PH^{(a,b)}/1/N$, etc.

This paper is organized as follows. Section 2 and its subsections analyze queueing model $BMAP/G^{(a,b)}/1/N$ elaborately. Section 3 and its subsections present analytic analysis of the queueing model $BMAP/MSP^{(a,b)}/1/N$. Section 4 presents numerical results in the form of graphs.

2. Description and analysis of $BMAP/G^{(a,b)}/1/N$ queue

Let us consider a single server finite-buffer queue where input process is *BMAP* and service times are generally distributed and independent of the arrival process, whereas N is the capacity of the queue excluding those who are in service. The service takes place in batches of maximum size b with a minimum threshold equal to $a(1 \leqslant a \leqslant b \leqslant N)$. However, if fewer than $a(\geqslant 1)$ customers are present in the queue, the server waits till the number of customers in the queue reaches a and then initiates service for that group of customers. The arrival process *BMAP* is characterized by $m \times m$ matrices $\mathbf{D}_k, k \geqslant 0$ where (i,j)th $(1 \leqslant i,j \leqslant m, i \neq j)$ element of $\mathbf{D}_0$, is the state transition rate from state i to state j in the underlying Markov chain without an arrival and (i,j)th $(1 \leqslant i,j \leqslant m)$ element of $\mathbf{D}_k, k \geqslant 1$, is the state transition rate from state i to state j in the underlying Markov chain with an arrival of batch size k. The matrix $\mathbf{D}_0$ has nonnegative off-diagonal and negative diagonal elements, and the matrix $\mathbf{D}_k, k \geqslant 1$, has nonnegative elements. The negative of diagonal elements, i.e., (i,i)th $(1 \leqslant i \leqslant m)$ element of $\mathbf{D}_0$ represents the mean rate of exponential sojourn time in state i. Let $N(t)$ denote the number of arrivals in $[(0,t)]$ and $J(t)$ be the state of the underlying Markov chain at time t with state space $\{i: 1 \leqslant i \leqslant m\}$. Then $\{N(t), J(t)\}$ is a two-dimensional Markov process of *BMAP* with state space $\{(n,i): n \geqslant 0, 1 \leqslant i \leqslant m\}$. The infinitesimal generator of *BMAP* is given by

$$\mathbf{Q} = \begin{pmatrix} \mathbf{D}_0 & \mathbf{D}_1 & \mathbf{D}_2 & \mathbf{D}_3 & \cdots \\ \mathbf{0} & \mathbf{D}_0 & \mathbf{D}_1 & \mathbf{D}_2 & \cdots \\ \mathbf{0} & \mathbf{0} & \mathbf{D}_0 & \mathbf{D}_1 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}.$$

As $\mathbf{Q}$ is the infinitesimal generator of the *BMAP*, we have $\sum_{k=0}^{\infty} \mathbf{D}_k \mathbf{e}_m = \mathbf{0}$, where $\mathbf{e}_m$ is a $m \times 1$ vector with all its elements equal to 1. Throughout the paper we sometimes do not use the subscript m and write it as $\mathbf{e}$. However, when $\mathbf{e}$'s dimension is other than m we write its subscript as its dimension. Further, since $\mathbf{D} = \sum_{k=0}^{\infty} \mathbf{D}_k$ is the infinitesimal generator of the underlying Markov chain $\{J(t)\}$, there exists a stationary probability vector $\bar{\pi}$ such that $\bar{\pi}\mathbf{D} = \mathbf{0}$, $\bar{\pi}\mathbf{e} = 1$. Then the average arrival rate λ^* and average batch arrival rate λ_g of the stationary *BMAP* are given by $\lambda^* = \bar{\pi}\sum_{k=1}^{\infty} k\mathbf{D}_k\mathbf{e}$ and $\lambda_g = \bar{\pi}\sum_{k=1}^{\infty}\mathbf{D}_k\mathbf{e} = \bar{\pi}\mathbf{D}'_1\mathbf{e}$, respectively, where $\mathbf{D}'_n = \sum_{i=n}^{\infty}\mathbf{D}_i$. Let us define $\{\mathbf{P}(n,t), n \geqslant 0, t \geqslant 0\}$ as $m \times m$ matrix whose (i,j)th element is the conditional probability defined as

$$P_{ij}(n,t) = Pr\{N(t) = n, J(t) = j | N(0) = 0, J(0) = i\}.$$

These matrices satisfy the following system of difference-differential equations

$$\frac{d}{dt}\mathbf{P}(0,t) = \mathbf{P}(0,t)\mathbf{D}_0, \quad (1)$$

$$\frac{d}{dt}\mathbf{P}(n,t) = \sum_{i=0}^{n}\mathbf{P}(i,t)\mathbf{D}_{n-i}, \quad n \geqslant 1, \quad (2)$$

with $\mathbf{P}(0,0) = \mathbf{I}_m$, where $\mathbf{I}_m$ is the identity matrix of dimension m. Usually we write this matrix as $\mathbf{I}$ when its dimension is equal to m else we write $\mathbf{I}$'s dimension in its subscript. Let us define matrix generating function $\mathbf{P}^*(z,t)$ as

$$\mathbf{P}^*(z,t) = \sum_{n=0}^{\infty}\mathbf{P}(n,t)z^n, \quad |z| \leqslant 1, \quad (3)$$

From (1)–(3), we have

$$\frac{d}{dt}\mathbf{P}^*(z,t) = \mathbf{P}^*(z,t)\sum_{i=0}^{\infty}\mathbf{D}_i z^i, \quad (4)$$

$$\frac{d}{dt}\mathbf{P}^*(z,0) = \mathbf{I}_m. \quad (5)$$

Solving the above matrix-differential equations, we get

$$\mathbf{P}^*(z,t) = e^{\mathbf{D}(z)t}, \quad |z| \leqslant 1, t \geqslant 0, \quad (6)$$

where $\mathbf{D}(z) = \sum_{i=0}^{\infty}\mathbf{D}_i z^i$.

Since we deal with finite-buffer queue with batch arrival, one may consider different batch acceptance/rejection strategies when buffer is going to be full and a batch arrives for service. Batches upon arrival find not enough space in the buffer are, either fully rejected, or a part of the batch is rejected. Some queueing protocol are based on the former strategy and it is known as total batch rejection policy. Latter one is known as the partial batch rejection policy. Here we consider partial batch rejection policy, i.e., if an arriving batch finds not enough space in the buffer, some of the customers of that batch are accepted till the buffer is full, and rests are rejected. In case of partial batch rejection strategy and finite-buffer capacity let us define $\{\mathbf{P}(n,t)(0 \leqslant n \leqslant N, t \geqslant 0)\}$ as $m \times m$ matrix whose (i,j)th element is the probability to admit n customers in the system during the time interval $[(0,t)]$ and to have the state j of the underlying Markov chain of the $BMAPJ(t)$ at the epoch t conditional that the state of this process was i at the epoch 0. The matrices $\mathbf{P}(n,t)$ satisfy following system of difference-differential equations:

$$\mathbf{P}^{(1)}(n,t) = \sum_{k=0}^{n}\mathbf{P}(k,t)\mathbf{D}_{n-k}, \quad 0 \leqslant n \leqslant N-1, \quad (7)$$

$$\mathbf{P}^{(1)}(N,t) = \sum_{k=0}^{N}\mathbf{P}(k,t)\mathbf{D}'_{N-k}, \quad (8)$$

with $\mathbf{P}(0,0) = \mathbf{I}_m$ for the above three cases and $\mathbf{P}^{(1)}(n,t) = \frac{d}{dt}\mathbf{P}(n,t)$.

The service times S of batches are independent identically distributed (i.i.d.) random variables (r.v.'s) with probability distribution function $S(x)$, density function $s(x)(x \geqslant 0)$ and mean service time $E(S)$. Let ρ be the traffic intensity, then $\rho = \lambda^* E(S)/b$.

2.1. Queue length distribution at departure epoch

Consider the system at a service completion instant of a batch. Let $t_0, t_1, t_2, \ldots$ be the time epochs at which service completion of a batch occurs. Let t_i^+ denotes the time epoch just after a service completion epoch of a batch. The state of the system at