*Applied Mathematics*, 2012, 3, 2169-2173
http://dx.doi.org/10.4236/am.2012.312A297 Published Online December 2012 (http://www.SciRP.org/journal/am)

# A Complete and Simple Solution to a Discrete-Time Finite-Capacity BMAP/D/c Queue

**Nam K. Kim[1], Mohan L. Chaudhry[2], Bong K. Yoon[3], Kilhwan Kim[4*]**
[1]Department of Industrial Engineering, Chonnam National University, Gwangju, South Korea
[2]Department of Mathematics and Computer Science, Royal Military College of Canada, Kingston, Canada
[3]Defense Management College, Korea National Defense University, Seoul, South Korea
[4]Department of Management Engineering, Sangmyung University, Cheonan, South Korea
Email: freedom@chonnam.ac.kr, chaudhry-ml@rmc.ca, bkyoon@kndu.ac.kr, *khkim@smu.ac.kr

Received November 10, 2012; revised December 10, 2012; accepted December 17, 2012

## ABSTRACT

We consider a discrete-time multi-server finite-capacity queueing system with correlated batch arrivals and deterministic service times (of single slot), which has a variety of potential applications in slotted digital telecommunication systems and other related areas. For this queueing system, we present, based on Markov chain analysis, not only the steady-state distributions but also the transient distributions of the system length and of the system waiting time in a simple and unified manner. From these distributions, important performance measures of practical interest can be easily obtained. Numerical examples concerning the superposition of certain video traffics are presented at the end.

**Keywords:** Discrete-Time Queue; Batch Markovian Arrival Process; Deterministic Service Time; Multiple Server

## 1. Introduction

The discrete-time multi-server queue with deterministic service times has gained importance in view of a number of potential practical applications to slotted digital telecommunication systems and other related areas (Bruneel and Wuyts [1]). It has been observed that arrival streams to these systems, in particular, tend to be correlated (see, e.g., Wittevrongel and Bruneel [2,3]). To model this correlated nature, a versatile point process called the *discrete-time batch Markovian arrival process* (D-BMAP) is introduced by Blondia and Casals [4] and widely used for analytical studies. This rich class of arrival processes contains a number of well-known arrival processes such as the Bernoulli process with independent identically distributed batch arrivals, the Markov modulated Bernoulli process, and a superposition of D-BMAP themselves (for more details, see Blondia and Casals [4]).

In this paper, we consider the D-BMAP/D/c/N queue, in which customers arrive according to D-BMAP and are served by one of $c$ servers. The system capacity is $N \geq c$ so that no more than $N$ customers can be accommodated in the system at the same time. Customers who arrive to find the system full are assumed to depart the system immediately on arrivals. Specifically, we assume a *partial acceptance model* (Takagi [5], p. 367) such that customers of an arriving batch are accepted until they fill all the available capacity, and the remaining customers, if any, are lost. Every accepted customer requires one slot for service that is assumed to start and end at slot boundaries.

For this queueing model, we present, based on an elementary Markov chain analysis, both steady-state and transient solutions to the system length (*i.e.*, the number of customers in system) as well as to the system waiting time (*i.e.*, the number of slots a customer spends in system) in a simple and unified manner. For similar discrete-time multi-server queueing models with infinite capacity, there have been some contributions by other authors: For a similar queue with infinite capacity (and correlated batch arrivals that belong to a subclass of B-DMAP arrivals), Sohraby and Zhang [6] analyze in transform domain the transient behavior of the system length and present an efficient numerical inversion method to calculate a few performance measures of interest. (They briefly discuss the finite capacity case as well.) For the D-BMAP/D/c queue, Alfa [7] assumes constant service times of multiple slots and presents an efficient algorithm making clever use of the structural property of the system to obtain the steady-state distributions of the system length and the system waiting time. Gao *et al*. [8] (Gao *et al*. [9]) assume constant service times of multiple slots (geometric service times) with a two-state Markovian arrival process to present a steady-state analysis of

*Corresponding author.

 

the system length and the system waiting time.

In this paper, we assume the finite-capacity model for the following three practical reasons. First and foremost, as demonstrated in this paper, the finite-capacity model of this paper is much simpler to analyze than its corresponding infinite-capacity counterpart (see Remark 1 below; also, see Sohraby and Zhang [6] for sophisticated analysis of the infinite-capacity model). Second, queueing models with finite capacity can serve as excellent approximations (by taking the system capacity $N$ sufficiently large) for their corresponding infinite-capacity counterparts (see Remark 2 at the end of this paper). Third, all the queueing systems in reality have finite capacity.

We organize the paper as follows: In Section 2, we first present the steady-state system-length distribution based on the elementary Markov chain analysis. From this, important performance measures of practical interest are obtained, including the steady-state system waiting-time distribution. In addition, the corresponding transient solutions are presented in a simple and unified manner. In Section 3, we present a set of numerical results with various system capacities and a few different numbers of servers. For this, we use D-BMAP arrivals that characterize the superposition of certain video traffics. We end the paper with a remark on the finite-capacity model.

## 2. Analysis

In discrete-time queueing models, the time axis is divided into fixed-length intervals, called *slots*. It is assumed that customer arrivals and departures take place only at slot boundaries; thus, nothing is assumed to happen in the middle of a slot.

In the D-BMAP/D/c/N queue, customers arrive according to a D-BMAP with representation $\{\boldsymbol{D}_k,\ k \ge 0\}$, where $\boldsymbol{D}_k$ is an $m \times m$ matrix with elements $(\boldsymbol{D}_k)_{ij}, 1 \le i, j \le m$. This arrival process is governed by an $m$-state (or $m$-phase) *underlying Markov chain* (UMC). Specifically, let us suppose that the UMC is in some phase $i$ in a certain slot. Then, with probabilities $(\boldsymbol{D}_k)_{ij}$, there are $k \ge 0$ arrivals during the slot with the phase of the UMC being $j$ in the next slot. (See Blondia and Casals [4] for more details.) Note that the number of arrivals per slot (including those who are lost) is given by

$$\lambda = \boldsymbol{\pi} \sum_{k=1}^{\infty} k \boldsymbol{D}_k \boldsymbol{e} \tag{1}$$

where $\boldsymbol{\pi}$ is the stationary probability vector of the UMC with the transition probability matrix (TPM) $\boldsymbol{D} = \sum_{k=0}^{\infty} \boldsymbol{D}_k$ and $\boldsymbol{e}$ is a column vector of 1's (note that $\boldsymbol{\pi}$ is obtained by solving simultaneously the equations $\boldsymbol{\pi} = \boldsymbol{\pi} \boldsymbol{D}$ and $\boldsymbol{\pi} \boldsymbol{e} = 1$ for $\boldsymbol{\pi}$).

Now, we consider the discrete-time bivariate process $\{(N_k, S_k); 0 \le N_k \le N, 1 \le S_k \le m, k \ge 1\}$, where $N_k$ and $S_k$ denote, respectively, the system length and the phase of the UMC just after the beginning of the $k$th slot. Then we consider the number of customers that arrive during the $k$th slot (denoted by $A_k$) and the number of customers that depart at the end of the same slot (which is given by $\min(N_k, c)$); as a result, we have $N_{k+1}$ in terms of $N_k$ as follows:

$$N_{k+1} = \min(N_k + A_k, N) - \min(N_k, c), k \ge 1. \tag{2}$$

Note that $A_k$ is dependent only on $S_k$; thus, we obtain the discrete-time Markov chain having the following TPM: (please see the formula below)
where $\boldsymbol{D}_{\ge l} = \boldsymbol{D}_l + \boldsymbol{D}_{l+1} + \cdots, l \ge 0$.

Let $\boldsymbol{p}^{(k)} = \left(\boldsymbol{p}_0^{(k)}, \boldsymbol{p}_1^{(k)}, \cdots, \boldsymbol{p}_N^{(k)}\right)$ denote the state probability vector of the bivariate process just after the beginning of the $k$th slot, where $\boldsymbol{p}_n^{(k)} = \left(p_{n,1}^{(k)}, \cdots, p_{n,m}^{(k)}\right)$ and $p_{n,i}^{(k)} = \Pr\{(N_k, S_k) = (n,\ i)\}$. Then it is immediate to have $\boldsymbol{p}^{(k+1)} = \boldsymbol{p}^{(k)} \cdot \boldsymbol{T}, k \ge 1$.

### 2.1. Steady-State Analysis

**1) The System-Length Distribution:** Let $\boldsymbol{p} = (\boldsymbol{p}_0, \boldsymbol{p}_1, \cdots, \boldsymbol{p}_N) = \lim_{k \to \infty} \boldsymbol{p}^{(k)}$. Then the steady-state system-length distribution is obtained by solving simultaneously the equations $\boldsymbol{p} = \boldsymbol{p}\boldsymbol{T}$ and $\boldsymbol{p}\boldsymbol{e} = 1$ for $\boldsymbol{p}$. This can be readily carried out by using mathematical software packages such as MATLAB, Mathematica, or even otherwise (see numerical examples in Section 3). Now, important performance measures of practical interest

$$\boldsymbol{T} = \begin{array}{c} \\ 0 \\ 1 \\ \vdots \\ c \\ c+1 \\ \vdots \\ N-1 \\ N \end{array} \begin{pmatrix} & & & & & & & \\ \boldsymbol{D}_0 & \boldsymbol{D}_1 & \cdots & \boldsymbol{D}_{N-c-1} & \boldsymbol{D}_{N-c} & \cdots & \boldsymbol{D}_{N-1} & \boldsymbol{D}_{\ge N} \\ \boldsymbol{D}_0 & \boldsymbol{D}_1 & \cdots & \boldsymbol{D}_{N-c-1} & \boldsymbol{D}_{N-c} & \cdots & \boldsymbol{D}_{N-1} & \boldsymbol{0} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \boldsymbol{D}_0 & \boldsymbol{D}_1 & \cdots & \boldsymbol{D}_{N-c-1} & \boldsymbol{D}_{\ge N-c} & \boldsymbol{0} & \cdots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{D}_0 & \cdots & \boldsymbol{D}_{N-c-2} & \boldsymbol{D}_{\ge N-c-1} & \boldsymbol{0} & \cdots & \boldsymbol{0} \\ \vdots & \vdots & \ddots & \boldsymbol{D}_{N-c-2} & \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{D}_0 & \boldsymbol{D}_{\ge 1} & \boldsymbol{0} & \cdots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{0} & \boldsymbol{D}_{\ge 0} & \boldsymbol{0} & \cdots & \boldsymbol{0} \end{pmatrix} \tag{3}$$

 

can be obtained from $\boldsymbol{p}_n$ as given below.

**2) The Effective Arrival Rate and the Loss Probability:** The number of arrivals (including those who are lost) per slot is given by

$$\lambda = \sum_{n=0}^{N} \boldsymbol{p}_n \sum_{k=1}^{\infty} k\boldsymbol{D}_k \boldsymbol{e} . \tag{4}$$

(Equation (4) reduces to (1) due to $\sum_{n=0}^{N} \boldsymbol{p}_n = \boldsymbol{\pi}$). Under the assumption of partial acceptance, the number of accepted arrivals per slot (*i.e.*, the *effective arrival rate*), on the other hand, is given by

$$\begin{aligned} \lambda_e &= \sum_{n=0}^{N-1} \boldsymbol{p}_n \left( \sum_{k=1}^{N-n} k\boldsymbol{D}_k + \sum_{k=N-n+1}^{\infty} (N-n)\boldsymbol{D}_k \right) \boldsymbol{e} \\ &= \sum_{n=0}^{N-1} \boldsymbol{p}_n \left( \sum_{k=1}^{N-n} \boldsymbol{D}_{\geq k} \right) \boldsymbol{e}. \end{aligned} \tag{5}$$

From (4) and (5), it is immediate to have the loss probability (*i.e.*, the probability that a customer is lost) as follows:

$$P_{\text{loss}} = 1 - \frac{\lambda_e}{\lambda} . \tag{6}$$

**3) Moments of the System Length:** Among others, the first moments of the numbers of customers in system and in service just after a slot boundary, denoted by $L$ and $L_S$ respectively, are given by

$$L = \sum_{n=1}^{N} n\boldsymbol{p}_n \boldsymbol{e} \tag{7}$$

and

$$L_S = \sum_{n=1}^{c} n\boldsymbol{p}_n \boldsymbol{e} + \sum_{n=c+1}^{N} c\boldsymbol{p}_n \boldsymbol{e} . \tag{8}$$

Along the same lines, higher moments such as the variances of the numbers of customers in system and in service can be easily obtained.

Note that $L_S$ also represents the number of departures (by service completions) per slot, *i.e.*, the *departure rate*, which, of course, should match with the effective arrival rate; that is, $L_S = \lambda_e$. Thus, the loss probability can be alternatively obtained from $P_{\text{loss}} = 1 - L_S / \lambda$.

**4) The System Waiting-Time Distribution of an Accepted Customer:** Let $W$ denote the system waiting time of an accepted customer, *i.e.*, the number of slots an accepted customer spends in system (we do not count, as a part of the system waiting time, the slot in which she arrives). The probability that an accepted customer spends at most $w$ slots in system can be interpreted as the long-run fraction of such customers out of all accepted customers. That is,

$$P(W \leq w) = E\big(A(w)\big) / \lambda_e \tag{9}$$

where $A(w)$ denotes the number of accepted customers per slot who are to spend at most $w$ slots in system. Note that because the service times are single slot, one can foresee whether the system waiting time of an accepted customer will exceed $w$ or not. Let $N_\infty$ and $S_\infty$ denote, respectively, the system length and the phase of the UMC just after the beginning of a slot at steady state. Then, conditioning on $N_\infty$ and $S_\infty$, we have $E\big(A(w)\big)$ as follows;

$$E\big(A(w)\big) = \sum_{n=0}^{N} \sum_{i=1}^{m} E\big(A(w) \big| N_\infty = n, S_\infty = i\big) \cdot p_{n,i} , \tag{10}$$

where

$$E\big(A(w) \big| N_\infty = n, S_\infty = i\big) = \begin{cases} \displaystyle\sum_{j=1}^{m} \left\{ \sum_{k=1}^{\min(cw, N-n)} k\left(\boldsymbol{D}_k\right)_{i,j} \right. \\ \left. \quad + \displaystyle\sum_{k=\min(cw, N-n)+1}^{\infty} \min\big(cw, N-n\big)\left(\boldsymbol{D}_k\right)_{i,j} \right\} \\ 0 \leq n \leq c, \\ \displaystyle\sum_{j=1}^{m} \left\{ \sum_{k=1}^{\min(c(w+1)-n, N-n)} k\left(\boldsymbol{D}_k\right)_{i,j} \right. \\ \left. \quad + \displaystyle\sum_{k=\min(c(w+1)-n, N-n)+1}^{\infty} \min\big(c(w+1)-n, N-n\big)\left(\boldsymbol{D}_k\right)_{i,j} \right\} \\ c+1 \leq n \leq N-1. \end{cases}$$

After simplifications (*i.e.*, following the same procedure as used in getting the last term of (5)), we have

$$\begin{aligned} E\big(A(w)\big) = &\sum_{n=0}^{c} \boldsymbol{p}_n \sum_{k=1}^{\min(cw, N-n)} \boldsymbol{D}_{\geq k} \boldsymbol{e} \\ &+ \sum_{n=c+1}^{N-1} \boldsymbol{p}_n \sum_{k=1}^{\min(c(w+1)-n, N-n)} \boldsymbol{D}_{\geq k} \boldsymbol{e}. \end{aligned} \tag{11}$$

Substituting (11) into (9), we have the steady-state system waiting-time distribution of an accepted customer. From this, one can get performance measures of interest, such as the mean $\big(E(W)\big)$ of the system waiting time and its tail probabilities. Also, one can get $L$ alternatively by Little's formula, $L = \lambda_e \cdot E(W)$.

## 2.2. Transient Analysis

Note that $\boldsymbol{p}^{(k)}$ is obtained from $\boldsymbol{p}^{(k)} = \boldsymbol{p}^{(1)} \cdot \boldsymbol{T}^{k-1}, k \geq 1$, where $\boldsymbol{p}^{(1)}$, the initial probability vector, is assumed to be given. Putting $\boldsymbol{p}_n^{(k)}$ in place of $\boldsymbol{p}_n$ of (4) through (11) derived for the steady state, one can immediately obtain the corresponding transient results: the expected numbers of total, accepted, and lost arrivals during the $k$ th slot, the moments of the numbers of customers in system as well as in service just after the beginning of the same slot, and the system waiting-time distribution of customers that are accepted during that slot.

 

**Remark 1:** Transient analyses of queueing models are, in general, much more demanding than their stationary counterparts, because the formers need to take an additional variable (time) into consideration. See, e.g., Sohraby and Zhang [6] for a transient analysis of the queue of a similar kind with infinite capacity. This is not the case for the finite-capacity case, which can be analyzed in a remarkably simple and unified manner as presented in this paper.

## 3. Numerical Examples

For numerical work, we use the same D-BMAP arrival as the one given in Example 2 of Blondia and Casals [4]. In this example, they approximate the superposition of 3 video sources by the superposition of 30 independent identical on/off sources. The latter is then characterized by the D-BMAP, where the phase of the UMC corresponds to the number of active sources. (See Blondia and Casals [4] for the representation of this D-BMAP.)

For various system capacities $N = \{4,6,8,10,12\}$ with different numbers of servers $c = \{1,2,3\}$, **Table 1** gives the steady-state loss probabilities. From this, one can see how fast the loss probability decreases as the system capacity *N* increases and as the number of servers *c* increases.

**Table 2**, with fixed capacity $N = 10$ and the numbers of servers $c = \{1,2,3\}$, gives the steady-state probabilities that the system waiting time of an accepted customer exceeds given threshold values $\{1,2,3,4,5\}$. Such tail probabilities are one of the important performance measures of practical interest (particularly, in telecommunication area) that represent the quality of service.

For $\boldsymbol{p}_n^{(1)} = \mathbf{0}, 0 \le n \le N-1$, and $\boldsymbol{p}_N^{(1)} = \boldsymbol{\pi}$ (*i.e.*, the system is initially full of customers) with fixed capacity $N = 10$ and the numbers of servers $c = \{1,2,3\}$, **Figure 1** displays how fast the expected system length of each system just after the beginning of the *k*th slot $(k = 1,2,\cdots,20)$ converges to its corresponding steady-state quantity. Other measures of practical interest can be obtained along the same lines.

**Remark 2:** For each system with the number of servers $c = \{1,2,3\}$, we observe that the loss probability tends to converge to zero pretty quickly with a moderate increase in the system capacity (see **Table 1**); in addition, we observe that the tail probability of each waiting-time distribution decays pretty quickly as well (see **Table 2**). This seems to be mainly due to the extreme regularity of the (constant) service times and the multiple numbers of servers, both of which, individually as well as jointly, absorb burstiness of the arrival process considerably. Consequently, in such cases, one can effectively reduce the loss probability, the tail probability of waiting-time distribution, or both with a slight increase in the system capacity or the number of severs. Besides, in such cases as the loss probabilities are practically zero, a finite-capacity model can serve as an excellent approximation for the corresponding infinite-capacity counterpart. Then one can avoid sophisticated analyses for the infinite-

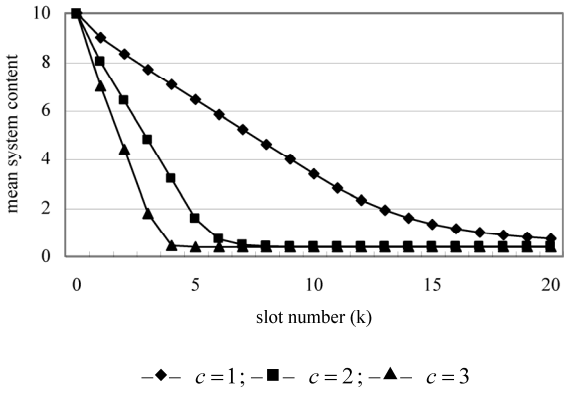


**Figure 1. Transient mean system lengths.**

**Table 1. Steady-state loss probabilities.**

| Number of servers | Sys. capacity | | | | |
|---|---|---|---|---|---|
| | $N = 4$ | $N = 6$ | $N = 8$ | $N = 10$ | $N = 12$ |
| $c = 1$ | 7.388627E−03 | 6.05570E−04 | 6.98359E−05 | 1.09435E−05 | 2.21558E−06 |
| $c = 2$ | 3.682965E−03 | 6.69900E−05 | 8.55853E−07 | 1.01555E−08 | 1.35632E−10 |
| $c = 3$ | 3.611292E−03 | 5.86295E−05 | 5.58459E−07 | 3.60230E−09 | 3.01454E−11 |

**Table 2. Steady-state tail probabilities of the system waiting time.**

| Number of servers | $P(W > 1)$ | $P(W > 2)$ | $P(W > 3)$ | $P(W > 4)$ | $P(W > 5)$ |
|---|---|---|---|---|---|
| $c = 1$ | 0.262522 | 0.062382 | 0.015792 | 0.004463 | 0.001390 |
| $c = 2$ | 0.021443 | 0.000175 | 1.23E−06 | 2.64E−10 | 0 |
| $c = 3$ | 0.001665 | 2.98E−07 | 4.31E−12 | 0 | 0 |

 

capacity model and get both steady-state and transient solutions in a remarkably simple and unified manner as presented in this paper.

We hope that the elementary Markov-chain based analysis we present in this paper for the finite-capacity D-BMAP/D/c/N queue would turn out to be beneficial to both theoreticians and practitioners who would like simple and straightforward practical solutions to their complex queueing systems.

## 4. Acknowledgements

The first author acknowledges with thanks the support provided by Prof. Sungjune Park and BISOM (Business Information Systems and Operations Management) of University of North Carolina at Charlotte, where he held an Adjunct-Visiting Professorship during his sabbatical year 2012 and where part of this work was done. The second author's research was supported in part by grant 4010 NSERC RG.

## REFERENCES

[1] A H. Bruneel and I. Wuyts, "Analysis of Discrete-Time Multiserver Queueing Models with Constant Service Times," *Operations Research Letters*, Vol. 15, No. 5, 1994, pp. 231-236. doi:10.1016/0167-6377(94)90082-5

[2] S. Wittevrongel and H. Bruneel, "Exact Calculation of Buffer Contents Variance and Delay Jitter in a Discrete-Time Queue with Correlated Input Traffic," *Electronics Letters*, Vol. 32, No. 14, 1996, pp. 1258-1259. doi:10.1049/el:19960848

[3] S. Wittevrongel and H. Bruneel, "Discrete-Time Queues with Correlated Arrivals and Constant Service Times," *Computers & Operations Research*, Vol. 26, No. 2, 1999, pp. 93-108. doi:10.1016/S0305-0548(98)00053-7

[4] C. Blondia and O. Casals, "Statistical Multiplexing of VBR Sources: A Matrix-Analytic Approach," *Performance Evaluation*, Vol. 16, No. 1-3, 1992, pp. 5-20. doi:10.1016/0166-5316(92)90064-N

[5] H. Takagi, "Queueing Analysis," Vol. 3, Discrete-Time Systems, North-Holland, Amsterdam, 1993.

[6] K. Sohraby and J. Zhang, "Spectral Decomposition Approach for Transient Analysis of Multi-Server Discrete-Time Queues," *Performance Evaluation*, Vol. 21, No. 1-2, 1994, pp. 131-150. doi:10.1016/0166-5316(94)90031-0

[7] A. S. Alfa, "Algorithmic Analysis of the BMAP/D/k System in Discrete Time," *Advances in Applied Probability*, Vol. 35, No. 4, 2003, pp. 1131-1152. doi:10.1239/aap/1067436338

[8] P. Gao, S. Wittevrongel and H. Bruneel, "On the Behavior of Multiserver Buffers with Geometric Service Times and Bursty Input Traffic," *IEICE TRANSACTIONS on Communications*, Vol. E87-B, No. 12, 2004, pp. 3576-3583.

[9] P. Gao, S. Wittevrongel, J. Walraevens and H. Bruneel, "Analytic Study of Multiserver Buffers with Two-State Markovian Arrivals and Constant Service Times of Multiple Slots," *Mathematical Methods of Operations Research*, Vol. 67, No. 2, 2008, pp. 269-284. doi:10.1007/s00186-007-0163-z

 