Performance Analysis of a Discrete-Time Two-Phase Queueing System

Tae-Sung Kim, Seok Ho Chang, and Kyung Chul Chae

This paper introduces the modeling and analysis of a discrete-time, two-phase queueing system for both exhaustive batch service and gated batch service. Packets arrive at the system according to a Bernoulli process and receive batch service in the first phase and individual services in the second phase. We derive the probability generating function (PGF) of the system size and show that it is decomposed into two PGFs, one of which is the PGF of the system size in the standard discrete-time Geo/G/1 queue without vacations. We also present the PGF of the sojourn time. Based on these PGFs, we present useful performance measures, such as the mean number of packets in the system and the mean sojourn time of a packet.

Manuscript received July 5, 2002; revised Feb. 4, 2003.

Tae-Sung Kim (phone: +82 43 261 3343, email: kimts@cbnu.ac.kr) is with the Department of Management Information Systems, Chungbuk National University, Chungbuk, Korea.

Seok Ho Chang (email: soccer9@mit.edu) is with the Lab for Manufacturing and Productivity, Massachusetts Institute of Technology, Massachusetts, USA.

Kyung Chul Chae (email: kcchae@kaist.ac.kr) is with the Department of Industrial Engineering, KAIST, Daejeon, Korea.

I. Introduction

Recently, discrete-time queueing systems have been extensively used in the design and performance analysis of telecommunication networks. In particular, they can be used to model various mechanisms in Broadband Integrated Services Digital Network (B-ISDN), which is expected to integrate the transmission of voice, video, and data in a single network. Asynchronous Transfer Mode (ATM) is the standard switching technique adopted by ITU-T [1] for the implementation of B-ISDN. It provides high flexibility of network access, dynamic bandwidth allocation on demand, and flexible capacity allocation. In such an environment, all information, such as continuous data stream, voice, and video, is digitalized and segmented into small packets called cells [2]. Since ATM is based on a packet switching principle, all events, such as arrivals and departures of packets, are allowed only at regularly spaced points in time. Thus, the underlying mechanism of this system is represented adequately by discrete-time queues [3]-[7].

Two-phase queueing systems have been discussed in the past for their applications in various areas, such as computer, communication, manufacturing, and other stochastic systems. In many computer and communication service systems, the situation in which arriving packets receive batch mode service in the first phase followed by individual services in the second phase is common. Recent applications of this queueing system have been discussed by Krishna and Lee [8], Doshi [9], and Kim and Chae [10], to name a few. Most of the papers on two-phase queueing systems have mainly concentrated on continuous-time models, but to the best of our knowledge, no studies have dealt with performance analysis of the discrete-time two-phase queueing system.

In this paper, we consider a discrete-time, two-phase Geo/G/1 queueing system. Packets arrive at the system according to a Bernoulli process and receive batch service in the first phase followed by individual services in the second phase. This type of queueing problem can be easily found in various practical situations.

Consider a central processor connected to a number of peripherals or distributed subprocessors. The central processor collects the jobs arriving at the peripherals or the distributed subprocessors in batches and processes them sequentially. When there is no online job for collection, the processor can be switched to process offline jobs, to update storage devices, or to attend to maintenance or repair work. As soon as a job arrives, the server is turned on and starts to serve jobs in batch mode.

Another application is the inventory problem in which the arriving orders are collected, and when the order arrives, their service requirements, such as due date, quantity, and quality, are analyzed initially in batch mode. This is followed by individual services of the batch. As the system empties, which means there are no orders, stock can be replenished to prepare for the next orders.

One special case of our model is a discrete-time, two-phase queueing model in which an arriving packet requires a deterministic service time that could be greater than one slot. This queueing model has gained importance in recent years because of a number of potential practical applications in slotted digital communication systems, such as ATM switching elements, circuit-switched TDMA systems, and traffic concentrators [11].

One motivation of this paper is the fact that discrete-time queues are more useful in modeling and performance analysis of communication networks and computer systems in a digitalized environment. In addition, continuous-time results can be obtained from their discrete-time counterparts by imposing appropriate limits.

The rest of the paper is organized as follows. Section II gives assumptions of our system. Section III presents the probability generating functions (PGFs) of the system size and the system sojourn time for the exhaustive batch service model. Section IV presents the results for the gated batch service model. Finally, section V concludes this paper.

II. The System and Assumptions

We assume that the time axis is divided into fixed-length time intervals called slots and that service times can be started and completed only at slot boundaries and that their durations are integral multiples of slot durations. We adopt a late arrival system with delayed access (LAS-DA) where packets arrive late during a slot and get delayed access to the server if they arrive to find the system empty. The slot in which a packet arrives is not counted in calculating its sojourn time. [3], [4], [12].

This paper considers a system that satisfies the following assumptions.

Assumptions

Packets arrive at the system according to a Bernoulli process with a mean interarrival time $1/\lambda$. All arrivals go into Q1, the batch service queue. The batch service times $\{B_i, i=1,2, \dots\}$ are independent and identically distributed (i.i.d.) random variables with distribution function $B(t)$, finite mean $E(B)$, and PGF $B(z)$. On the completion of the batch service, the entire batch is transferred to Q2, the individual service queue. The transferring time from Q1 to Q2 is assumed to be zero. The individual service times $\{S_i, i=1,2, \dots\}$ are i.i.d. random variables with distribution function $S(t)$, finite mean $E(S)$, and PGF $S(z)$. We assume that the service in Q2 is FIFO based on the original order of arrival. The server is assumed to work in one of the following modes:

i) Case A: Exhaustive batch service

When Q2 is empty, the server returns to Q1. The transferring time from Q2 to Q1 is also assumed to be zero. If one or more packets are in Q1, the server will start the batch service immediately. Packets arriving during the on-going batch service are included in the ongoing batch service. If no packet is in Q1 when the server transfers from Q2, the server waits until a packet arrives. On the arrival of a packet, the batch service begins and proceeds as before. On completion of this batch service, all packets that received the batch service are moved to Q2 and individual services start immediately. When all individual services for this batch are completed, the server moves to Q1, and so on.

ii) Case B: Gated batch service

The basic mechanism is similar to that of Case A. The only difference is that the batch service includes only those packets that were present when the batch service started. Packets arriving during the batch service have to wait for the next batch service to start.

Finally, we assume that the system has an infinite queueing capacity and it is stationary. Thus, all packets arriving at the system are eventually served, so that $\rho = \lambda E(S) < 1$, where $E(S)$ denotes the mean individual service time. Remark 1 makes some comments on this statement.

Remark 1. The system under study belongs to a class of Geo/G/1 queues with generalized vacations in the sense that both idle periods and first-phase service periods act as if they were a vacation period. Therefore, the stability condition ($\lambda E(S)$<1) of the Geo/G/1 queue with generalized vacations is applicable to our system [4].

The notations in this paper are listed as follows:

Q: the system size right after the beginning of a batch service

Q_1: the system size right after the completion of a batch service

Q_2: the system size right after the end of the individual services of the batch

B: the batch service time

S: the individual service time

M: the system size at an arbitrary slot boundary

W_q: the sojourn time of the test packet until its individual service

$Q(z), Q_1(z), Q_2(z)$: PGF of Q, Q_1, Q_2

$r(z)$: PGF of M

$B(z)$, $S(z)$: PGF of B, S

$W_q(z)$: PGF of W_q

$q_0 \equiv \Pr(Q_2 = 0)$: the probability that the system is empty at the end of a second phase service

$\gamma = \lambda E(B)$, $\rho = \lambda E(S)$

III. Exhaustive Batch Service: System Size and Sojourn Time Distributions

1. Regeneration Cycle Analysis

From the definitions of the above notation, the following relations can easily be seen:

$Q_1 = Q +$ the number of arrivals during the batch service

$Q_2 =$ the number of arrivals during the individual services of the Q_1 packets

$$Q = \begin{cases} Q_2, & \text{if } Q_2 > 0 \\ 1, & \text{if } Q_2 = 0. \end{cases}$$

Because the PGFs of the number of arrivals during B and S are given by $B(\lambda z + 1 - \lambda)$ and $S(\lambda z + 1 - \lambda)$, respectively, [4], the above relations are translated into

$$Q_1(z) = Q(z) \cdot B(\lambda z + 1 - \lambda). \tag{1}$$

$$Q_2(z) = Q_1(S(\lambda z + 1 - \lambda)). \tag{2}$$

$$Q(z) = [Q_2(z) - q_0] + q_0 z. \tag{3}$$

Combining (1), (2) and (3), we get the following functional relationship for $Q(z)$:

$$Q(z) = Q(S(\lambda z + 1 - \lambda)) \cdot B(\lambda S(\lambda z + 1 - \lambda) + 1 - \lambda) - q_0(1 - z). \tag{4}$$

We will now present the closed form expression for $Q(z)$ in (4). To this end, we first present the closed form expression for q_0 in (4) using the procedures in Sumita [13] and Doshi [9].

Let us define the following notation.

$$\begin{aligned} &L(z) = 1 - z, \\ &b^{(0)}(z) = z,\ b(z) = B(\lambda z + 1 - \lambda), \\ &s^{(0)}(z) = z,\ s(z) = s^{(1)}(z) = S(\lambda z + 1 - \lambda), \\ &s^{(n)}(z) = s(s^{(n-1)}(z)) = s^{(n-1)}(s(z)), n \ge 1. \end{aligned}$$

We can use successive substitutions in (4). Then we get

$$Q(z) = \prod_{j=1}^{\infty} b(s^{(j)}(z)) - q_0 \sum_{k=0}^{\infty} L(s^{(k)}(z)) \prod_{j=1}^{k} b(s^{(j)}(z)). \tag{5}$$

See appendix for the detailed derivation of (5).

If we set $z = 0$ in both sides of (4), we get

$$Q\left(s^{(1)}(0)\right) b\left(s^{(1)}(0)\right) - q_0 = 0,$$

from which we get

$$q_0 = b\left(s^{(1)}(0)\right) Q\left(s^{(1)}(0)\right). \tag{6}$$

Combining (5) and (6), we get

$$\begin{aligned} q_0 &= b\left(s^{(1)}(0)\right) Q\left(s^{(1)}(0)\right) \\ &= b\left(s^{(1)}(0)\right) \left[\prod_{j=1}^{\infty} b\left(s^{(j+1)}(0)\right) - q_0 \sum_{k=0}^{\infty} L\left(s^{(k+1)}(0)\right) \prod_{j=1}^{k} b\left(s^{(j+1)}(0)\right) \right] \\ &= \prod_{j=1}^{\infty} b\left(s^{(j)}(0)\right) - q_0 \sum_{k=0}^{\infty} L\left(s^{(k+1)}(0)\right) \prod_{j=1}^{k+1} b\left(s^{(j)}(0)\right), \end{aligned}$$

from which we get

$$q_0 = \frac{\prod_{j=1}^{\infty} b\left(s^{(j)}(0)\right)}{1 + \sum_{k=0}^{\infty} L\left(s^{(k+1)}(0)\right) \prod_{j=1}^{k+1} b\left(s^{(j)}(0)\right)}. \tag{7}$$

If we substitute (7) into (5), we get the final closed form expression for $Q(z)$ as follows:

$$Q(z) = \prod_{j=1}^{\infty} b\left(s^{(j)}(z)\right) - \frac{\prod_{j=1}^{\infty} b\left(s^{(j)}(0)\right)}{1 + \sum_{k=0}^{\infty} L\left(s^{(k+1)}(0)\right) \prod_{j=1}^{k+1} b\left(s^{(j)}(0)\right)} \times \sum_{k=0}^{\infty} L\left(s^{(k)}(z)\right) \prod_{j=1}^{k} b\left(s^{(j)}(z)\right). \tag{8}$$

Differentiating (1), (2), and (3) with respect to z and setting $z = 1$, we get

$$E(Q_1) = E(Q) + \gamma, \tag{9}$$

$$E(Q_2) = \rho E(Q_1), \tag{10}$$

$$E(Q) = E(Q_2) + q_0. \tag{11}$$

From (9), (10), and (11), we get

$$E(Q_1) = \frac{\gamma + q_0}{1-\rho}. \tag{12}$$

Based on the quantities (9), (10), (11), and (12), we present a regeneration cycle analysis to derive the probabilities (16a), (16b), and (16c), which are crucial to determining the system size distribution.

Note that the regeneration points of our system are those instants at which the system becomes empty right after the completions of individual services. We call the interval between two such successive regeneration points a regeneration cycle. Dividing the regeneration cycle into the initial idle period and K subservice cycles, each of which consists of a batch service and the individual services of the batch, we see that $E(K)=1/q_0$. Let D be the number of arrivals during the initial delay, which consists of the idle period and K first-phase batch service periods in the cycle and Γ be the number of arrivals during the delay cycle (or the regeneration cycle). Then, based on the delay cycle arguments [4], we can find the following expected values.

$$E(D) = 1 + E(K) \cdot \lambda E(B) = 1 + \frac{1}{q_0}\gamma, \tag{13}$$

$$E(\Gamma) = \frac{E(D)}{1-\rho} = \frac{q_0 + \gamma}{1-\rho} \cdot \frac{1}{q_0} = E(Q_1)E(K). \tag{14}$$

Let T_c be the length of the regeneration cycle. Using Wald's equation, the expected length of the regeneration cycle, $E(T_c)$, is given by the following:

$$E(T_c) = E(\Gamma)E(A), \tag{15}$$

where $E(A)$ denotes the expected interarrival time, which is given by λ^{-1}.

Let TP stand for the test packet. Based on renewal reward arguments and on the property of Bernoulli arrivals see time averages (BASTA) [14], we can derive the following probabilities from (12), (13), (14), and (15).

Pr(TP arrives during an individual service period)

$$= \frac{E(\Gamma)E(S)}{E(T_c)} = \lambda E(S) = \rho. \tag{16a}$$

Pr(TP arrives during a batch service period)

$$= \frac{E(K)E(B)}{E(T_c)} = \frac{E(B)}{E(Q_1)E(A)} = \frac{\lambda E(B)}{E(Q_1)} = \frac{(1-\rho)\gamma}{\gamma + q_0}. \tag{16b}$$

Pr(TP arrives during an idle period)

$$= \frac{E(A)}{E(T_c)} = \frac{1}{E(Q_1)E(K)} = \frac{q_0}{E(Q_1)} = \frac{(1-\rho)q_0}{\gamma + q_0}. \tag{16c}$$

2. The System Size Distribution

In this subsection, we present the system size distribution based on the probabilities (16a), (16b), and (16c).

As stated earlier, the key observation is that the system under study belongs to a class of Geo/G/1 queues with generalized vacations [4, pp. 90-93] such that both idle periods and first-phase periods act as if they were vacation periods. In Geo/G/1 queues with generalized vacations, the PGF of the system size at an arbitrary slot boundary is given by the product of two PGFs: one is the PGF of the system size of an ordinary Geo/G/1 queue (without vacations) at an arbitrary slot boundary, and the other is the conditional PGF of the system size at an arbitrary slot boundary given that the server is in a vacation period. This property is called stochastic decompositions [4]. Note that this decomposition property holds for a broad class of continuous- and discrete-time queues with generalized vacations [15], [16].

The system size PGF at an arbitrary slot boundary for the standard discrete-time Geo/G/1 queue without vacations is given by $\dfrac{(1-\rho)(1-z)s(z)}{s(z)-z}$, where $s(z) = S(\lambda z + 1 - \lambda)$ [4]. To obtain the conditional system size PGF at an arbitrary slot boundary given that the server is in a vacation period, we use the approach of Chae et al. [17], which is based on the conditioning of the system state. This approach holds for a broad class of discrete-time queues with generalized vacations [16].

Note that the conditional system size PGF at an arbitrary slot boundary given that the server is in a vacation period is equal to the conditional system size PGF at TP's arrival-epoch given that TP arrives during a vacation period by BASTA [14]. The conditional system size PGF at TP's arrival-epoch given that TP arrives during a vacation period consists of two parts depending on whether TP arrives during an idle period or during a first-phase period. If TP arrives during an idle period with a conditional probability $\frac{q_0}{\gamma+q_0}$, it is clear that the PGF is given by (1). If TP arrives during a first-phase period with a conditional probability $\frac{\gamma}{\gamma+q_0}$, the PGF is given by $Q(z)\frac{1-b(z)}{\gamma(1-z)}$, where $b(z)=B(\lambda z+1-\lambda)$.

Putting all these together, we finally get the PGF of the system size at an arbitrary slot boundary (or at TP's arrival epoch),

$$r(z)=\frac{(1-\rho)(1-z)s(z)}{s(z)-z}\left[\frac{q_0}{\gamma+q_0}+\frac{\gamma}{\gamma+q_0}\cdot Q(z)\cdot\frac{1-b(z)}{\gamma(1-z)}\right]. \quad (17)$$

Note that the terms q_0 and $Q(z)$ contained in (17) are, respectively, given by (7) and (8). Differentiating (17) with respect to z and setting it at $z=1$, we get the expected system size,

$$E(M)=\rho+\frac{\lambda^2E(S^2)-\lambda\rho}{2(1-\rho)}+\frac{\lambda^2E(B^2)}{2(\gamma+q_0)}+\frac{\gamma}{1-\rho}-\frac{\gamma^2}{\gamma+q_0}. \quad (18)$$

Note that the term q_0 contained in (17) and (18) is given by (7).

Remark 2. As stated earlier, using appropriate limits, we can get results corresponding to (17) and (18) for the continuous-time, two-phase M/G/1 queue. If we assume the length of a slot is equal to a constant Δ, we can consider the transition of the above results (17) and (18) for the continuous-time system by using the limit $\Delta\to 0$ [4].

3. The System Sojourn Time Distribution

To find W_q, the sojourn time of the test packet until its individual service, we use the arrival time approach of Chae et al. [17]. We need to consider three cases:

- Case 1: The arriving TP that finds the server is idle has to wait during the batch service time for itself. Therefore, the PGF of the sojourn time of a packet that arrives during the idle period is given by

$$W_q(z|idle)=B(z). \quad (19)$$

- Case 2: The test packet that arrives during the first phase batch service has to wait

i) the remaining batch service time, plus
ii) the individual service times of the packets that arrive during the elapsed batch service time, plus
iii) the individual service times of Q packets.

i) + ii) is easily obtained by using the joint PGF of the elapsed time and the remaining time [4]. Thus, the PGF of the sojourn time of a packet that arrives during the batch service time is given by the following:

$$W_q(z|batch\ service)=Q(S(z)).\frac{z\{B(z)-B(u)\}}{(z-u)E(B)}\bigg|_{u=\lambda S(z)+1-\lambda}. \quad (20)$$

- Case 3: The test packet that arrives during the second phase individual service has to wait if we let T be the total individual service times of the packets that arrive during the first phase batch service plus Q packets,

i) the remaining time of T period, plus
ii) the individual service times of the packets that arrive during the elapsed time of T period, plus
iii) the next batch service time.

The PGF of the sojourn time of a packet that arrives during the individual services is given by the following.

$$W_q(z|individual\ service)=B(z).\frac{z\{T(z)-T(u)\}}{(z-u)E(T)}\bigg|_{u=\lambda S(z)+1-\lambda}, \quad (21)$$

where $T(z)=B(\lambda S(z)+1-\lambda)\cdot Q(S(z))$.

The probabilities of each case are determined by (16a), (16b), and (16c), respectively. Unconditioning (19), (20), and (21) with the probabilities of each case, the PGF of W_q, the sojourn time of an arbitrary packet until its individual service, becomes

$$\begin{aligned}W_q(z)&=(1-\rho)\cdot\frac{q_0}{\gamma+q_0}\cdot B(z)\\&+(1-\rho)\cdot\frac{\gamma}{\gamma+q_0}\cdot\frac{z[B(z)-B(\lambda S(z)+1-\lambda)]}{[z-\lambda S(z)-1+\lambda]E(B)}\cdot Q(S(z))\\&+\rho\cdot\frac{z[T(z)-T(\lambda S(z)+1-\lambda)]}{[z-\lambda S(z)-1+\lambda]E(T)}\cdot B(z).\end{aligned} \quad (22)$$

Note that the terms q_0 and $Q(z)$ contained in (22) are, respectively, given by (7) and (8). Differentiating (22) with respect to z and setting $z=1$, we obtain the mean sojourn time $E(W_q)$ as follows:

$$E(W_q)=\frac{\lambda E(S^2)-\rho}{2(1-\rho)}+\frac{\lambda E(B^2)}{2(\gamma+q_0)}+\frac{\gamma}{\lambda(1-\rho)}-\frac{\gamma^2}{\lambda(\gamma+q_0)}. \quad (23)$$

Note that the term q_0 contained in (23) is given by (7). Equation (23) can also be obtained by Little's law using (18).

We can also get results corresponding to (22) and (23) for the continuous-time, two-phase M/G/1 queue using appropriate limits. Thus, if we assume the length of a slot equal to a constant Δ, we can consider the transition of the above results of (22) and (23) for the continuous-time system by taking the limit $\Delta \to 0$ [4], [14].

IV. Gated Batch Service: System Size and Sojourn Time Distributions

In this section, we consider the gated batch service model, which we described in section II.

From the definitions of the notation in section II, the following relations are clear:

$Q_1 = Q$ + (the number of arrivals during the batch service)

Q_2 =(the number of arrivals during the individual services of the Q_1 packets) + (the number of arrivals during the batch service)

$$Q = \begin{cases} Q_2, & \text{if } Q_2 > 0 \\ 1, & \text{if } Q_2 = 0. \end{cases}$$

The above relations are translated into the following PGF forms:

$$Q_1(z) = Q(z) \cdot B(\lambda z + 1 - \lambda), \tag{24}$$

$$Q_2(z) = Q(S(\lambda z + 1 - \lambda)) \cdot B(\lambda z + 1 - \lambda), \tag{25}$$

$$Q(z) = [Q_2(z) - q_0] + q_0 z. \tag{26}$$

Combining (25) and (26), we get the following functional relationship for $Q(z)$.

$$Q(z) = Q(S(\lambda z + 1 - \lambda)) \cdot B(\lambda z + 1 - \lambda) - q_0(1 - z). \tag{27}$$

Following a procedure close to that in section III, the closed form expression for $Q(z)$ and q_0 are given, respectively, by the following:

$$Q(z) = \prod_{j=0}^{\infty} b\left(s^{(j)}(z)\right) - q_0 \sum_{k=0}^{\infty} L\left(s^{(k)}(z)\right) \prod_{j=0}^{k-1} b\left(s^{(j)}(z)\right) \tag{28}$$

and
$$q_0 = \frac{b(0) \prod_{j=1}^{\infty} b\left(s^{(j)}(0)\right)}{1 + b(0) \sum_{k=0}^{\infty} L\left(s^{(k+1)}(0)\right) \prod_{j=1}^{k} b\left(s^{(j)}(0)\right)}. \tag{29}$$

The appendix gives the detailed derivations of (28) and (29).

If we substitute (29) into (28), we get the final closed form expression for $Q(z)$ for gated batch service as follows:

$$Q(z) = \prod_{j=0}^{\infty} b\left(s^{(j)}(z)\right) - \frac{b(0) \prod_{j=1}^{\infty} b\left(s^{(j)}(0)\right)}{1 + b(0) \sum_{k=0}^{\infty} L\left(s^{(k+1)}(0)\right) \prod_{j=1}^{k} b\left(s^{(j)}(0)\right)} \sum_{k=0}^{\infty} L\left(s^{(k)}(z)\right) \prod_{j=0}^{k-1} b\left(s^{(j)}(z)\right). \tag{30}$$

Differentiating (25) and (26) with respect to z and setting $z = 1$, we obtain

$$E(Q_2) = \rho E(Q) + \gamma, \tag{31}$$

$$E(Q) = E(Q_2) + q_0. \tag{32}$$

Combining (31) and (32), we get

$$E(Q) = \frac{q_0 + \gamma}{1 - \rho}. \tag{33}$$

Taking the same steps as given in section III.2, we can derive the PGF of the system size $r(z)$ and its expected value $E(M)$ as follows:

$$r(z) = \frac{(1-\rho)(1-z)s(z)}{s(z) - z} \left[\frac{q_0}{\gamma + q_0} + \frac{\gamma}{\gamma + q_0} \cdot Q(z) \cdot \frac{1 - b(z)}{\gamma(1-z)} \right], \tag{34}$$

$$E(M) = \rho + \frac{\lambda^2 E(S^2) - \lambda\rho}{2(1-\rho)} + \frac{\lambda^2 E(B^2)}{2(\gamma + q_0 N)} + \frac{\gamma}{1-\rho}. \tag{35}$$

Note that the terms q_0 and $Q(z)$ contained in (34) and (35) are, respectively, given by (29) and (30). $Q(z)$ and $E(Q)$ determine the system size distribution. Note further that the expressions for $r(z)$ in (17) and (34) are the same, but the contents of $Q(z)$ are different.

The analysis for the system sojourn time distribution is similar to that for the exhaustive batch service. Specifically, (19) is still valid and only (20) and (21) need minor modifications. We now need to multiply $B(z)$ and $B(\lambda S(z) + 1 - \lambda)$ to the right hand sides of (20) and (21), respectively. Consequently, we have the PGF of the sojourn time $W_q(z)$ and its expected value $E(W_q)$ as follows:

$$\begin{aligned} W_q(z) &= (1-\rho) \cdot \frac{q_0}{\gamma + q_0} \cdot B(z) \\ &+ (1-\rho) \cdot \frac{\gamma}{\gamma + q_0} \cdot \frac{zB(z) - zB(\lambda S(z) + 1 - \lambda)}{[z - \lambda S(z) - 1 + \lambda]E(B)} \cdot Q(S(z)) \cdot B(z) \\ &+ \rho \cdot \frac{zT(z) - zT(\lambda S(z) + 1 - \lambda)}{[z - \lambda S(z) - 1 + \lambda]E(T)} \cdot B(\lambda S(z) + 1 - \lambda) \cdot B(z). \end{aligned} \tag{36}$$

$$E(W_q)=\frac{\lambda E(S^2)-\rho}{2(1-\rho)}+\frac{\lambda E(B^2)}{2(\gamma+q_0)}+\frac{\gamma}{\lambda(1-\rho)}. \quad (37)$$

Note that the terms q_0 and $Q(z)$ contained in (36) and (37) are, respectively, given by (29) and (30). Equation (37) can also be obtained by Little's law using (35).

Remark 3. As stated earlier, using appropriate limits, we can get results corresponding to (33), (34), (35) and (36) for the continuous-time, two-phase M/G/1 queue. Thus, if we assume the length of a slot is equal to a constant Δ, we can consider the transition of the above results of (33), (34), (35), and (36) for the continuous-time system by using the limit $\Delta\to 0$ [4], [14].

Remark 4. From (18) and (35) (or (23) and (37)), we conclude that implementing exhaustive batch service reduces the mean system size and the mean sojourn time compared with gated batch service. A simple comparison between the gated batch service model and the exhaustive batch service model shows that the mean system size of the gated batch service model is larger than that of the exhaustive batch service model by $\frac{\gamma^2}{\gamma+q_0}$, and that the mean sojourn time of the gated batch service model is larger than that of the exhaustive batch service model by $\frac{\gamma^2}{\lambda(\gamma+q_0)}$. These differences are consistent with the discrete-time version of Little's formula [4].

V. Conclusions

In this paper, we analyzed a discrete-time, two-phase Geo/G/1 queueing system for both exhaustive batch service and gated batch service. We identified our model as belonging to a class of Geo/G/1 queue with generalized vacations. From this observation, we presented the PGFs of the system size and the system sojourn time based on a decomposition property. Based on these PGFs, we presented useful performance measures, such as the mean number of packets in the system and mean sojourn time of a packet.

The results in this paper may be useful for system designers and practitioners involved in investigating the performance of slotted digital communication systems and related areas. The results and methodology presented in this paper can be used to study discrete-time, two-phase queueing systems with various threshold policies, such as multiple and single vacations and N-policy.

Appendix 1. Detailed Derivation of (5)

If we express (4) using the notation defined in section III.1, we get

$$Q(z)=Q\left(s^{(1)}(z)\right)\cdot b\left(s^{(1)}(z)\right)-q_0L\left(s^{(0)}(z)\right). \quad \text{(A1)}$$

If we set $z=s^{(1)}(z)$ in both sides of (A1), we get

$$Q\left(s^{(1)}(z)\right)=Q\left(s^{(2)}(z)\right)\cdot b\left(s^{(2)}(z)\right)-q_0L\left(s^{(1)}(z)\right). \quad \text{(A2)}$$

Substituting (A2) into (A1), we get

$$\begin{aligned}Q(z)=&\,Q\left(s^{(2)}(z)\right)\cdot b\left(s^{(2)}(z)\right)b\left(s^{(1)}(z)\right)\\&-q_0L\left(s^{(1)}(z)\right)b\left(s^{(1)}(z)\right)-q_0L\left(s^{(0)}(z)\right).\end{aligned} \quad \text{(A3)}$$

If we set $z=s^{(1)}(z)$ in both sides of (A3), we get

$$\begin{aligned}Q\left(s^{(1)}(z)\right)=&\,Q\left(s^{(3)}(z)\right)\cdot b\left(s^{(3)}(z)\right)b\left(s^{(2)}(z)\right)\\&-q_0L\left(s^{(2)}(z)\right)b\left(s^{(2)}(z)\right)-q_0L\left(s^{(1)}(z)\right).\end{aligned} \quad \text{(A4)}$$

Substituting (A4) into (A1), we get

$$\begin{aligned}Q(z)=&\,Q\left(s^{(3)}(z)\right)\cdot b\left(s^{(3)}(z)\right)b\left(s^{(2)}(z)\right)b\left(s^{(1)}(z)\right)\\&-q_0L\left(s^{(2)}(z)\right)b\left(s^{(2)}(z)\right)b\left(s^{(1)}(z)\right)\\&-q_0L\left(s^{(1)}(z)\right)b\left(s^{(1)}(z)\right)-q_0L\left(s^{(0)}(z)\right).\end{aligned} \quad \text{(A5)}$$

Note that, if $\rho<1$, $z=1$ is the unique solution, inside the unit circle $|z|\le 1$, of

$$z=s^{(1)}(z).$$

This implies that for any z with $|z|\le 1$,

$$s^{(n)}(z)\to 1, \quad \text{as } n\to\infty, \quad \text{(A6)}$$

and
$$L\left(s^{(n)}(z)\right)\to 0, \quad \text{as } n\to\infty. \quad \text{(A7)}$$

Thus, we can use successive substitutions in (A1) using (A6) and (A7). Then we get

$$Q(z)=\prod_{j=1}^{\infty}b\left(s^{(j)}(z)\right)-q_0\sum_{k=0}^{\infty}L\left(s^{(k)}(z)\right)\prod_{j=1}^{k}b\left(s^{(j)}(z)\right),$$

which is reduced to (5).

Appendix 2. Detailed Derivations of (28) and (29)

If we express (27) using the notation defined in section III.1, we get

$$Q(z)=Q\left(s^{(1)}(z)\right)\cdot b(z)-q_0L\left(s^{(0)}(z)\right). \quad \text{(A8)}$$

If we set $z=s^{(1)}(z)$ in both sides of (A9), we get

$$Q\left(s^{(1)}(z)\right)=Q\left(s^{(2)}(z)\right)\cdot b\left(s^{(1)}(z)\right)-q_0L\left(s^{(1)}(z)\right). \quad \text{(A9)}$$

Substituting (A9) into (A8), we get

$$Q(z)=Q\left(s^{(2)}(z)\right)\cdot b\left(s^{(1)}(z)\right)b(z)-q_0L\left(s^{(1)}(z)\right)b(z)-q_0L\left(s^{(0)}(z)\right). \quad \text{(A10)}$$

If we set $z=s^{(1)}(z)$ in both sides of (A10), we get

$$\begin{aligned}Q\left(s^{(1)}(z)\right)=&\ Q\left(s^{(3)}(z)\right)\cdot b\left(s^{(2)}(z)\right)b\left(s^{(1)}(z)\right)\\&-q_0L\left(s^{(2)}(z)\right)b\left(s^{(1)}(z)\right)-q_0L\left(s^{(1)}(z)\right).\end{aligned} \quad \text{(A11)}$$

Substituting (A11) into (A8), we get

$$\begin{aligned}Q(z)=&\ Q\left(s^{(3)}(z)\right)\cdot b\left(s^{(3)}(z)\right)b\left(s^{(2)}(z)\right)b\left(s^{(1)}(z)\right)b(z)\\&-q_0L\left(s^{(2)}(z)\right)b\left(s^{(1)}(z)\right)b(z)-q_0L\left(s^{(1)}(z)\right)b(z)\\&-q_0L\left(s^{(0)}(z)\right).\end{aligned} \quad \text{(A12)}$$

Thus, we can use successive substitutions in (A8) using (A6) and (A7). Then we get

$$Q(z)=\prod_{j=0}^{\infty}b\left(s^{(j)}(z)\right)-q_0\sum_{k=0}^{\infty}L\left(s^{(k)}(z)\right)\prod_{j=0}^{k-1}b\left(s^{(j)}(z)\right), \quad \text{(A13)}$$

which is reduced to (28).

If we set $z=0$ in both sides of (A8), we get

$$Q(s(0))b(0)-q_0=0,$$

from which we get

$$q_0=b(0)Q(s(0)). \quad \text{(A14)}$$

Combining (A13) and (A14), we get

$$\begin{aligned}q_0&=b(0)Q(s(0))\\&=b(0)\left[\prod_{j=1}^{\infty}b\left(s^{(j)}(0)\right)-q_0\sum_{k=0}^{\infty}L\left(s^{(k+1)}(0)\right)\prod_{j=1}^{k}b\left(s^{(j)}(0)\right)\right],\end{aligned}$$

from which we get

$$q_0=\frac{b(0)\prod_{j=1}^{\infty}b\left(s^{(j)}(0)\right)}{1+b(0)\sum_{k=0}^{\infty}L\left(s^{(k+1)}(0)\right)\prod_{j=1}^{k}b\left(s^{(j)}(0)\right)}.$$

This is (29).

Acknowledgements

The authors would like to thank the anonymous referees for their valuable comments. Their careful readings and valuable comments greatly enhanced the richness and readability of this paper. Reference [13] was brought to the authors' attention by one of the referees.

The research of Seok Ho Chang was supported by the Post-doctoral Fellowship Program of Korea Science & Engineering Foundation (KOSEF).

References

[1] IUT-T, *I-Series Recommendation on B-ISDN*, Geneva, Switzerland, 1994.

[2] Daniel W.-K. Hong, C.S. Hong, D.S. Yun, J.H. Yoo, and W.-S. Kim, "An ATM Network Management System for Point-to-Multipoint Reservation Service," *ETRI J.*, vol. 24, no. 4, Aug. 2002, pp. 299-310.

[3] H. Bruneel and B. Kim, *Discrete-Time Models for Communication Systems Including ATM*, Kluwer Academic Publishers, Boston, MA, 1993.

[4] H. Takagi, *Queueing Analysis, Vol. III: Discrete-Time Systems*, North-Holland, 1993.

[5] F. Ishizaki, T. Takine, and T. Hasegawa, "Analysis of Discrete-Time Queue with Gated Priority," *Performance Evaluation*, vol. 13, no. 2, 1995, pp. 121-143.

[6] M. Miyazawa and H. Takagi, "Advances in Discrete Time Queue," *Queueing Systems*, vol. 18, 1994.

[7] J.K. Choi, "Discrete-Time Queueing Analysis of Dual-Plane ATM Switch with Synchronous Connection Control," *ETRI J.*, vol. 19, 1997, pp. 326-343.

[8] C.M. Krishna and Y.H. Lee, "A Study of Two-Phase Service," *Operations Research Lett.*, vol. 9, 1990, pp. 91-97.

[9] B.T. Doshi, "Analysis of a Two Phase Queueing System with General Service Times," *Operations Research Lett.*, vol. 10, 1991, pp. 265-272.

[10] T.S. Kim and K.C. Chae, "Two-Phase Queueing System with Generalized Vacation (in Korean)," *Journal of Korean Institute of Industrial Engineers*, vol. 22, 1996, pp. 95-104.

[11] H. Bruneel and I. Wuyts, "Analysis of Discrete-Time Multiserver Queueing Models with Constant Service Times," *Operations Research Lett.*, vol. 15, 1994, pp. 231-236.

[12] J.J. Hunter, *Mathematical Technique of Applied Probability, Vol. II: Discrete-Time Models*, Academic Press, New York, 1983.

[13] S. Sumita, "Performance Analysis of Interprocessor Communications in an Electronic Switching System with Distributed Control," *Performance Evaluation*, vol. 9, 1988/89, pp. 83-91.

[14] O.J. Boxma and W.P. Groenendijk, "Sojourn Times in Discrete-Time Cyclic-Service Systems," *IEEE Trans. on Communications*, vol. 36, 1988, pp. 164-170.

[15] S.H. Chang, T. Takine, K.C. Chae, and H.W. Lee, "A Unified Queue Length Formula for BMAP/G/1 Queue with Generalized Vacations," *Stochastic Models*, vol. 18, 2002, pp. 369-386.

[16] S.H. Chang, N.K. Kim, K.C. Chae, H.W. Lee, and T. Takine, "Factorization Properties in the Discrete-Time BMAP/G/1 Queue with Generalized Vacations," *Technical Report 04*, Dept. of

Industrial Engineering, Korea Advanced Institute of Science and Technology, KAIST, 2002.

[17] K.C. Chae, H.W. Lee, and C.W. Ahn, “An Arrival Time Approach to M/G/1-Type Queues with Generalized Vacations,” *Queueing Systems*, vol. 38, 2001, pp. 91-100.

Tae-Sung Kim has been working for Department of Management Information Systems at Chungbuk National University since September 2000. He received his bachelor, master, and doctoral degrees in engineering from the Department of Management Science at KAIST in 1991, 1993 and 1997. He worked for Internet Economics Team at ETRI (Electronics and Telecommunications Research Institute) as a Senior Researcher from February 1997 to August 2000. His research areas include performance evaluation, telecommunications management, telecommunications policy and information security management. He is currently a member of Korean Institute of Communication Sciences (KICS), Korea Institute of Information Security & Cryptology (KIISC), The Korea Society of Management Information System (KMIS), and Korean Operations Research and Management Science Society (KORMS).

Seok Ho Chang received the BS degree from the Department of Industrial Engineering of Hanyang University in Korea in 1998 and the MS and PhD degrees from the Department of Industrial Engineering at the Korea Advanced Institute of Science and Technology (KAIST), Korea in 2000 and 2003. He is currently an invited visiting scholar in the Laboratory for Manufacturing and Productivity, Massachusetts Institute of Technology (MIT), in the USA, where he is researching on the performance analysis of large-scale manufacturing systems and the real-time scheduling in large-scale manufacturing systems. His research interests include operations research, with emphasis on queueing theory and its applications to the performance analysis of high-speed telecommunication networks and manufacturing systems. His recent research papers have appeared in international journals, such as Stochastic Models, Operations Research Letters, and Stochastic Analysis and Applications. He is currently a member of Korean Institute of Industrial Engineers (KIIE) and Korean Institute of Communication Sciences (KICS).

Kyung Chul Chae is a Professor in the Department of Industrial Engineering, KAIST. He received the MS and PhD in industrial engineering from Ohio State University. His research interests include analysis of queueing systems and stochastic modeling. Detailed activities can be found in http://osl7.kaist.ac.kr.