4

Delay and overflow of discrete-time priority queue with burst arrivals and partial buffer sharing

Hideaki Yamashita
Faculty of Business Administration, Komazawa University
Setagaya, Tokyo 154, JAPAN, Tel:+81 3 3418 9437
Fax:+81 3 3418 9127, E-mail: i38666@m-unix.cc.u-tokyo.ac.jp

abstract

We study a discrete-time, single-server queue with partial buffer sharing. There are two priority classes of jobs. Though class 1 jobs in the queue have higher priority for the next service than any of class 2 jobs, class 2 jobs are allowed to occupy their own part of buffer when the shared part of buffer is full. We characterize a bursty arrival process using bursts which consist of the same class of jobs. Once the first job of a burst arrives at the queue, the successive jobs will arrive on every time slot until the last job of the burst arrives. The numbers of jobs of a burst and the inter-arrival times of bursts are assumed to be i.i.d., respectively, and the service time is assumed to be equal to one slot. This model targets the buffer management to meet the quality of service requirments of different traffic types as video, voice and data in ATM multiplexer. In particular, class 1 jobs may correspond to cells with the strict delay requirments. On the other hand, class 2 jobs may correspond to cells with the strict cell loss requirments. We propose an efficient numerical method to exactly obtain the job loss probability, the waiting time distribution and the mean queue length. Some numerical examples are also given.

keywords

Discrete-time, Burst arrival, Head of the line priority, Partial buffer sharing, ATM multiplexer, Markov chain.

1. INTRODUCTION

We study a discrete-time, single-server queue with partial buffer sharing. There are two priority classes of jobs, and we characterize a bursty arrival process using bursts which consist of the same class of jobs. Once the first job of a burst arrives at the queue, the successive jobs will arrive on every time slot

Performance and Management of Complex Communication Networks
T. Hasegawa, H. Takagi & Y. Takahashi (Eds.) © 1998 IFIP. Published by Chapman & Hall

until the last job of the burst arrives. The numbers of jobs of a burst and the inter-arrival times of bursts are assumed to be i.i.d., respectively, and the service time is assumed to be equal to one slot. The buffer consists of the shared part and the part for class 2 jobs only, whose capacitis are finite, so class 2 jobs are allowed to occupy their own part of buffer when the shared part of buffer is full. Class 1 jobs can occupy only the shared part, but they have a higher priority for the next service than any of class 2 jobs, i.e., non-preemptive head of the line priority.

This model was motivated by ATM(Asynchronous Transfer Mode) multiplexer (see, for example, Händel and Huber (1991)). In ATM networks, all information including voice, video, and data, is conveyed using a fixed-size block call a cell, and each type of information has its own quality of service (QOS) requirements, such as a cell loss probability and an end-to-end delay. For instance, the voice traffic has more strict delay requirment than the data traffic, but is tolerant for the cell loss requirment than others. The model targets the buffer management to meet the various QOS requirments of each traffic type. In particular, class 1 jobs may correspond to cells with the strict delay requirments. On the other hand, class 2 jobs may correspond to cells with the strict cell loss requirments. Such buffer management strategy that combines the head of the line priority and the partial buffer sharing (or the pushout priority) was studied for the Poisson arrival cases by Gravey and Hebuterne (1991).

The switch architecture is synchronized. Between two synchronization points any incoming cells that are in process of arriving at the input ports are written to the memory, and each output port transmits cells (if there are any for the output port). Because of the synchronization, the discrete-time queueing system is more suitable for the model of ATM multiplexer than the continuous-time one. The service time of the job is assumed to be equal to one slot, since the length of cells is fixed in ATM switch.

ATM uses short fixed length cells to transmit the variable length packets generated at higher layers. The arrival process of cells cannot be renewal in general, because of a correlation between inter-arrival times of cells. This is one of the important characteristics of ATM traffic and makes the performance analysis difficult. In this model, the packet and the cell correspond to the burst and the job, respectively. The burst represents the sequential incoming cells from an input port. In order to model a superposition of each arrival stream from an input port, we consider an arrival stream in which the inter-arrival time of bursts is generally distributed. The continuous-time version of this input process called 'Gradual Input' has been analyzed by Kino and Miyazawa (1993).

A number of models have been proposed to capture the effect of correlated input processes. However, most of them have considered the continuous-time queue, and there are several results for the descrete-time queue. A model with a geometrically distributed burst size and a Poisson burst arrival has been analyzed by Miyazawa and Yamazaki (1992). Morris (1981) has modeled a correlated input process by considering the source to be a function of a Markov chain and has obtained queue length distributions numerically. Neuts (1990) has obtained an explicit formula for the mean waiting time in a queue whose input is generated by N heterogeneous Markovian on-off sources. The generating functions of both the queue length and waiting time distributions have derived for SBBP(Switched Batch Bernoulli Process)/G/1 queue by Hashida, Takahashi, and Shimogawa (1991) and for SBBP/G/1 priority queue by Hashida and Takahashi (1991). Brandt, Brandt and Sulanke (1990) have studied the batch arrival of messages with the geometrically distributed number of packets using the generating function. Johnson and Narayama (1996) analyzed discrete-time Markovian arrival processes as descriptors of discrete-time bursty arrival processes. The discrete-time priority queues with correlated arrivals have been also studied in Takine, Sengupta, and Hasegawa (1994). The input processes proposed above are simpler than the one considered in this paper. Yamashita (1994) has modeled the single-class burst arrival system, and numerically obtained the performance measures. Yamashita (1994) has extended this analysis to the two class model with shared buffer. Finally, we mention that a performance model with partial buffer sharing has been analyzed by Kröner (1990).

In this paper, we propose an efficient numerical method to exactly obtain the job loss probability, the waiting time distribution, and the mean queue length. For this purpose, we derive an embedded Markov chain at the arrival instants of bursts, which enables us to save a lot of space and computational efforts. The remainder of the paper is organized as follows. In the following section, we show the queueing model under consideration. In section 3, we derive the embedded Markov chain at the arrival instants of bursts, and we exactly obtain some stationary performance measures in section 4. In section 5, some numerical examples are illustrated. Finally, the concluding remarks are given in section 6.

2. MODEL DESCRIPTION

The queueing model under consideration is a discrete-time, single-server queue with partial buffer sharing. There are two priority classes of jobs. After a service completion, class 1 jobs in the queue have higher priority for the next service than any of class 2 jobs (non-preemptive head of the line priority). This means that class 2 jobs may start their service only if there is no class 1 job in the queue. On the other hand, class 2 jobs are allowed to occupy their own part of buffer when the shared part of buffer is full. The capacity of the shared part and total capacity of buffer are denoted by M_1 and M_2 $(0 \leq M_1 \leq M_2)$, respectively. Then, the capacity $M_2 - M_1$ is used for class 2 jobs only. Note that the case of $M_1 = 0$ corresponds to the loss system for class 1 jobs.

We characterize a bursty arrival process using bursts which consist of the same kind of jobs, i.e., class 1 or class 2. Once the first job of a burst arrives at the queue, the successive jobs will arrive on every time slot until the last job of the burst arrives. The probability that a burst consists of priority class i $(i = 1, 2)$ jobs is denoted by r_i. The number of jobs of the nth burst is denoted by S_i^n if the burst consists of class i jobs, which is assumed to be independent and identically distributed (*i.i.d.*) with a general distribution. We assume that there exists a positive number S_{max} such that $Pr[S_1^n > S_{max}] = Pr[S_2^n > S_{max}] = 0$. The inter-arrival time between nth and $(n+1)$st bursts is denoted by T^{n+1}, which is assumed to be *i.i.d.* with a general distribution. We allow that T^n may take 0, i.e., more than one burst may arrive on the same slot.

Servers are synchronized so that they start and end services at the same time. The service time of the job is assumed to be equal to one slot. The jobs arrive at the queue at the beginning of a slot and leave the queue at the end of a slot. When the first job of a burst arrives at the queue, the burst tries to keep a server, or buffer space if the server has been kept already, for all jobs belonging to the burst. That is, the jobs of the nth burst have the higher priority to enter the queue than any job of the (n+1)st burst whenever they arrive. We call the rule FIFO discipline on a burst basis. An arriving job is lost if both the buffer for its class (only the shared part for class 1) and the server are occupied (or reserved) by other jobs belonging to prior bursts, even if they have not been in the system. Indeed, if $T^n + t < S_i^{n-1}$, the jobs after $(T^n + t)$th of the $(n-1)$st burst have not arrived yet when the tth job of the nth burst arrives. Note that even if a class 2 job of the nth burst succeeds in keeping the server, the job may be pushed out to the shared part of buffer if class 1 jobs of the bursts after nth arrive before its service and they can enter the shared part of buffer. In this case, the class 2 job is never lost, but is only made to wait in the buffer.

In the following two sections, we propose an efficient numerical method to analyze the queueing model described above.

3. EMBEDDED MARKOV CHAIN

In this section, we construct a finite state embedded Markov chain, which will be used for obtaining some stationary performance measures of the queue described in the previous section. First of all, let us consider an embedded Markov chain by giving attention to all active bursts, i.e., bursts with remaining jobs (which have not arrived yet). If we keep track of the number of remaining jobs of each active burst, the priority class which each active burst belongs to, and the number of jobs in the buffer for each priority class at the arrival instant of bursts, the process has a Markov property. It might be possible to obtain some stationary performance measures, e.g., the job loss probability, the queue length distribution and the waiting time distribution from the steady state probability distribution of the process. However, the process becomes intractable as the number of active bursts increases. Therefore, it is important to reduce the state space of the Markov chain in order to efficiently obtain some performance measures such as the job loss probability.

For this purpose, the methodology proposed by Yamashita (1994) is available. He analyzed the single class queue with the same arrival stream. The basic idea of his method is as follows: Let us consider the embedded point of the nth burst arrival instant. In order to know whether the jobs of the nth burst are lost or not, we need to know the number of jobs in the buffer because the jobs which find that the buffer is full are lost. In the case of single server queue, the number of jobs never decreases while at least one burst is active whenever the $(n+1)$st burst arrives. Therefore, we may only keep track of the largest number of remaining jobs of active bursts and the number of jobs in the buffer on the last slot when at least one burst is active. It is much more effective than keeping track of the number of remaining jobs of each active burst and the number of jobs in the buffer. We extend the basic idea for the priority queue model.

Let v_1^n denote the largest number of remaining jobs among active bursts of class 1 at the arrival instant of the nth burst. In other words, v_1^n means the time until the last slot when at least one burst of class 1 is active counting from the arrival instant of the nth burst, excluding the $(n+1)$st burst and all the bursts after $(n+1)$st. Similarly, let v_2^n denote the largest number of remaining jobs among all active bursts (of class 1 or class 2) at the arrival instant of the nth burst. Note that if $0 < v_1^n < v_2^n$, then the largest number of remaining jobs among active bursts of class 2 is v_2^n at the arrival instant of the nth burst, but if $0 < v_1^n = v_2^n$, then whether there are active bursts of class 2 is not clear, which we do not mind.

Now, let us obtain the relationship between v_j^n and v_j^{n+1} $(j = 1, 2)$ given T^{n+1} and S_i^{n+1}. As the first case, we assume that $(n+1)$st burst belongs to class 2. The last job of the $(n+1)$st burst arrives on the $(T^{n+1} + S_2^{n+1})$th slot counting from the arrival instant of the nth burst. If $v_2^n \le T^{n+1} + S_2^{n+1}$, then the burst which has the largest number of remaining jobs at the arrival instant of the $(n+1)$st burst becomes the $(n+1)$st burst. Otherwise, it is not the $(n+1)$st burst but still the same burst at the arrival instant of the nth burst. Accordingly, we have following relations:

$$v_2^{n+1} = \begin{cases} S_2^{n+1}, & \text{if } \; v_2^n \le T^{n+1} + S_2^{n+1}, \\ v_2^n - T^{n+1}, & \text{if } \; v_2^n > T^{n+1} + S_1^{n+1}. \end{cases} \tag{3.1}$$

Since no new burst of class 1 arrives during T^{n+1} slots, we have

$$v_1^{n+1} = (v_1^n - T^{n+1})^+, \tag{3.2}$$

where

$$(N)^+ = \max(0, N).$$

As the same way, for the case of $(n+1)$st burst belonging to class 1, we have

$$v_j^{n+1} = \begin{cases} S_1^{n+1}, & \text{if } \; v_j^n \le T^{n+1} + S_1^{n+1}, \\ v_j^n - T^{n+1}, & \text{if } \; v_j^n > T^{n+1} + S_1^{n+1}, \end{cases} \tag{3.3}$$

where $j = 1, 2$.

Here, we introduce another kind of variables. Let w_1^n be the number of jobs of class 1 in the buffer on the v_1^nth slot counting from the arrival instant of the nth burst, excluding the $(n+1)$st burst and all the bursts after $(n+1)$st even if they have arrived already on the v_1^nth slot. w_1^n takes into account the arrival jobs which do not keep the server. Similarly, let w_2^n be the number of jobs (of class 1 or class 2) in the buffer on the v_2^nth slot counting from the arrival instant of the nth burst, excluding the $(n+1)$st burst and all the bursts after $(n+1)$st. Again, as the first case, we assume that $(n+1)$st burst belongs to class 2. If $v_1^n < T^{n+1}$, the server will serve the class 1 jobs in the buffer, if any, on every slot by one from $(v_1^n + 1)$st to T^{n+1}th slots counting from the arrival instant of the nth burst. However, if $v_1^n \geq T^{n+1}$, the class 1 jobs in the buffer are not served until T^{n+1}th slot, so $w_1^{n+1} = w_1^n$. Therefore, w_1^{n+1} can be written as follows:

$$w_1^{n+1} = \begin{cases} [w_1^n - (T^{n+1} - v_1^n)]^+, & \text{if } v_1^n < T^{n+1}, \\ w_1, & \text{if } v_1^n \geq T^{n+1}. \end{cases} \tag{3.4}$$

Now, let us obtain w_2^{n+1} given T^{n+1}, S_2^{n+1}, v_2^n, and w_2^n. If $v_2^n \leq T^{n+1}$, then w_2^{n+1} is equivalent to the number of jobs in the buffer on the T^{n+1}th slot counting from the arrival instant of the nth burst and is less than w_2^n since the jobs in the buffer will be served after the v_2^nth slot. If $v_2^n > T^{n+1}$, on the other hand, w_2^{n+1} is equivalent to the number of jobs in the buffer on the v_2^nth slot counting from the arrival instant of the $(n+1)$st burst, and is greater than w_2^n since the number of jobs increases on every slot by one from the T^{n+1}th to $\min(v_2^n, T^{n+1} + S_2^{n+1})$th slots counting from the arrival instant of the nth burst, as long as there is enough space in the buffer. From the above discussion, we have

$$w_2^{n+1} = \begin{cases} [w_2^n - (T^{n+1} - v_2^n)]^+, & \text{if } v_2^n - T^{n+1} \leq 0, \\ \min[w_2^n + v_2^n - T^{n+1}, M_2], & \text{if } 0 < v_2^n - T^{n+1} \leq S_2^{n+1}, \\ \min[w_2^n + S_2^{n+1}, M_2], & \text{if } S_2^{n+1} < v_2^n - T^{n+1}. \end{cases} \tag{3.5}$$

Next, we suppose that $(n+1)$st burst belongs to class 1. In this case, it is a little complicated to obtain w_i^{n+1}'s, since we have to take care not only of the total number of jobs in the buffer not to exceed M_2 but also of the number of class 1 jobs not to exceed M_1. After some straightforward considerations we can get the following relations for w_2^{n+1}:

$$w_2^{n+1} = \begin{cases} [w_2^n - (T^{n+1} - v_2^n)]^+, & \text{if } v_2^n \leq T^{n+1}, \\ \min[w_2^n + v_2^n - T^{n+1}, M_2], & \text{if } v_1^n \leq T^{n+1} < v_2^n \leq T^{n+1} + S_1^{n+1}, \\ \min[w_2^n + S_1^{n+1}, M_2], & \text{if } v_1^n \leq T^{n+1}, T^{n+1} + S_1^{n+1} < v_2^n, \\ \min[w_2^n + \min(v_1^n - T^{n+1},\ M_1 - w_1^n) + v_2^n - v_1^n, M_2], & \\ & \text{if } T^{n+1} < v_1^n, v_2^n \leq T^{n+1} + S_1^{n+1}, \\ \min[w_2^n + \min(v_1^n - T^{n+1},\ M_1 - w_1^n) + S_1^{n+1} + T_{n+1} - v_1^n, M_2], & \\ & \text{if } T^{n+1} < v_1^n \leq T^{n+1} + S_1^{n+1} < v_2^n, \\ w_2^n + \min(S_1^{n+1}, M_2 - w_2^n,\ M_1 - w_1^n) & \\ & \text{if } T^{n+1} + S_1^{n+1} < v_1^n. \end{cases} \tag{3.6}$$

In the first three cases of (3.6), the number of class 1 jobs never increase, so the formulae are the same as (3.5). In the fourth and fifth cases, the jobs from first to $(v_1^n - T^{n+1})$th in the $(n+1)$st burst arrive during the class 1 bursts are active, so the number of class 1 jobs possiblly increases by $v_1^n - T^{n+1}$ as long as the remaining buffer capacity $M_1 - w_1^n$ is enough and the total number of jobs in the buffer does not exceed M_2 . The jobs from $(v_1^n - T^{n+1} + 1)$st to $(v_2^n - T^{n+1})$th arrive during only the class 2 bursts are active. Therefore, they do not increase the number of class 1 jobs, but possiblly increase the toal number of jobs, since they push out the class 2 jobs from the server to the buffer and receive the services immediately. In the last case, all S_1^{n+1} jobs arrive during the class 1 bursts are active.

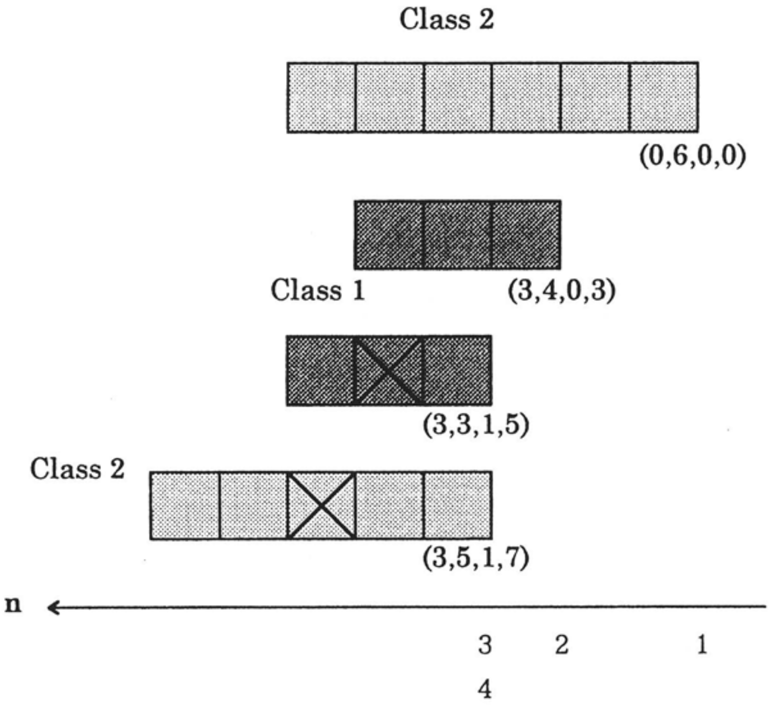

Figure 1 Sample path of $(v_1^n, v_2^n, w_1^n, w_2^n)$ $(M_1 = 1$ and $M_2 = 7)$.

Using (3.6), w_2^{n+1} can be expressed by

$$w_1^{n+1} = \begin{cases} [w_1^n - (T^{n+1} - v_1^n)]^+, & \text{if } v_2^n \le T^{n+1}, \\ [w_1^n + w_2^{n+1} - w_2^n - \ (v_2^n - v_1^n)]^+, & \\ \qquad \text{if } T^{n+1} < v_2^n \le T^{n+1} + S_1^{n+1}, & \\ [w_1^n + w_2^{n+1} - w_2^n - \ (S_1^{n+1} + T^{n+1} - v_1^n)]^+, & \\ \qquad \text{if } v_1^n \le T^{n+1} + S_1^{n+1} < v_2^n, & \\ w_2^n + \min(S_1^{n+1}, M_2 - w_2^n, M_1 - w_1^n), & \\ \qquad \text{if } T^{n+1} + S_1^{n+1} < v_1^n. & \end{cases} \tag{3.7}$$

Note that in the second case of (3.7), $(w_2^{n+1} - w_2^n)$ is the number of class 1 jobs which join the system until v_2^nth slot counting from the arrival instant of the nth burst, and $(v_2^n - v_1^n)$ is the time slots during which the class 1 jobs in the shard buffer receive the service if any. The sample path of $(v_1^n, v_2^n, w_1^n, w_2^n)$ is illusrated in Fig.1 for the case of $M_1 = 1$ and $M_2 = 7$, where the lost jobs are marked by ×.

$(v_1^n, v_2^n, w_1^n, w_2^n)$ has the Markov property, because $(v_1^{n+1}, v_2^{n+1}, w_1^{n+1}, w_2^{n+1})$ depends only on $(v_1^n, v_2^n, w_1^n, w_2^n)$ given T^{n+1} and S_i^{n+1}. Let us denote the set of all posible states of (v_1, v_2, w_1, w_2) by $\mathcal{U}$, and denote the relationship by:

$$(v_1^{n+1}, v_2^{n+1}, w_1^{n+1}, w_2^{n+1}) = f(v_1^n, v_2^n, w_1^n, w_2^n, S_1^{n+1}, T^{n+1}).$$

for the case that $(n+1)$st burst belongs to class 1, and

$$(v_1^{n+1}, v_2^{n+1}, w_1^{n+1}, w_2^{n+1}) = g(v_1^n, v_2^n, w_1^n, w_2^n, S_2^{n+1}, T^{n+1}).$$

for the case that $(n+1)$st burst belongs to class 2.

Since v_i^n's are bounded by S_{max} and $w_i^n \leq M_i$, $(v_1^n, v_2^n, w_1^n, w_2^n)$ is a finite state embedded Markov chain at the arrival instant of bursts with less than $(S_{max}+1)^2(M_1+1)(M_2+1)$ states, i.e., $O(S_{max}^2 M_1 M_2)$. $T^{n+1} \geq S_{max} + M_2$ is a sufficient condition for $v_2^{n+1} = S_i^{n+1}$ and $w_i^{n+1} = 0$ $(i = 1, 2)$. Therefore, it is sufficient to consider the case $T^{n+1} = 0, 1, \cdots, S_{max} + M_2$, $S_i^{n+1} = 1, 2, \cdots, S_{max}$ for every state $(v_1^n, v_2^n, w_1^n, w_2^n)$ to calculate the coefficients of the equilibrium equations using (3.1) ~ (3.7), which requires $O(S_{max}^3 M_1 M_2 (S_{max} + M_2))$ time. Once we calculate the coefficients of the equilibrium equations, we can get the steady state probability distribution of $(v_1^n, v_2^n, w_1^n, w_2^n)$, denoted by $P(v_1, v_2, w_1, w_2)$, by solving the system of stationary equilibrium equations:

$$\begin{aligned}
&P(v_1, v_2, w_1, w_2) \\
&= r_1 \sum_{S_1=1}^{S_{max}} \sum_{T=0}^{\infty} P(S_1)P(T) \sum_{(v_1', v_2', w_1', w_2') \in \Lambda(v_1, v_2, w_1, w_2, S_1, T)} P(v_1', v_2', w_1', w_2') \\
&+ r_2 \sum_{S_2=1}^{S_{max}} \sum_{T=0}^{\infty} P(S_2)P(T) \sum_{(v_1, v_2', w_1', w_2') \in \Gamma(v_1, v_2, w_1, w_2, S_2, T)} P(v_1', v_2', w_1', w_2')
\end{aligned}$$

for $(v_1, v_2, w_1, w_2) \in \mathcal{U}$, where

$$\begin{aligned}
&\Lambda(v_1^{n+1}, v_2^{n+1}, w_1^{n+1}, w_2^{n+1}, S_1^{n+1}, T^{n+1}) \\
&= \{(v_1^n, v_2^n, w_1^n, w_2^n) \mid (v_1^{n+1}, v_2^{n+1}, w_1^{n+1}, w_2^{n+1}) \\
&= f(v_1^n, v_2^n, w_1^n, w_2^n, S_1^{n+1}, T^{n+1})\}, \\
&\Gamma(v_1^{n+1}, v_2^{n+1}, w_1^{n+1}, w_2^{n+1}, S_2^{n+1}, T^{n+1}) \\
&= \{(v_1^n, v_2^n, w_1^n, w_2^n) \mid (v_1^{n+1}, v_2^{n+1}, w_1^{n+1}, w_2^{n+1}) \\
&= g(v_1^n, v_2^n, w_1^n, w_2^n, S_2^{n+1}, T^{n+1})\},
\end{aligned}$$

and $P(S_i)$ and $P(T)$ denote the probability that the number of jobs of a burst is S_i and the probability that the inter-arrival time of bursts is T, respectively.

We note that this method is still much more efficient than the straightforward way mentioned at the beginning of this section, though the number of states of the embedded markov chain $(v_1^n, v_2^n, w_1^n, w_2^n)$ increases in polynomial order as the maximum burst size and/or the capacity of the buffer increases.

4. PERFORMANCE MEASURES

In this section, we get the performance measures: the job loss probability, the waiting time distribution, and the mean number of jobs in the buffer, using steady state probability distribution $P(v_1, v_2, w_1, w_2)$ obtained in the previous section.

4.1 Job Loss Probabilities

We first calculate the job loss probability defined to be the ratio between the average number of jobs out of a burst that are lost and the average number of jobs arriving in a burst. Let C_i^{n+1} be the number of lost jobs in the $(n+1)$st burst belonging to class i given $(v_1^n, v_2^n, w_1^n, w_2^n)$, S_i^{n+1}, and T^{n+1}. Here, we again assume that the $(n+1)$st burst belongs to class 2. Because of FIFO discipline on a burst basis, the jobs of the $(n+1)$st burst are lost when they find, at their arrival slot, that the other job excluding all the bursts after $(n+1)$st keeps the server and that the buffer capacity for the class is full of jobs of the $(n+1)$st burst which already arrived and/or jobs of all the bursts before $(n+1)$st. If $v_2^n \le T^{n+1}$, all the arriving jobs can keep the server. If $v_2^n > T^{n+1}$, however, $\min(v_2^n - T^{n+1}, S_2^{n+1})$ jobs can not keep the server and try to enter the buffer. Then, since the available buffer capacity is $M_2 - w_2^n$, we have

$$C_2^{n+1} = \begin{cases} 0, & \text{if } v_2^n - T^{n+1} \le 0, \\ (v_2^n - T^{n+1} - M_2 + w_2^n)^+, & \text{if } 0 < v_2^n - T^{n+1} \le S_2^{n+1}, \\ (S_2^{n+1} - M_2 + w_2^n)^+, & \text{if } S_2^{n+1} < v_2^n - T^{n+1}. \end{cases} \quad (4.1)$$

Next, we consider the case that $(n+1)$st burst belongs to calss 1. In this case, the arriving jobs can not enter the buffer if the total number of jobs in the buffer is M_2 or the number of class 1 jobs in the buffer is M_1 excluding jobs of all the bursts after $(n+1)$st. The arriving jobs in the $(n+1)$st burst may find the three possible situations, that is, among the bursts before $(n+1)$st the class 1 bursts are active (until v_1^nth slots), or only the class 2 bursts are active (from (v_1^n+1)st to v_2^nth slots), or no burst is active (from (v_2^n+1)st slots). We define these three periods by period 1, period 2, and period 3, respectively. Let $C_{1,k}^{n+1}$ be the number of lost jobs in the $(n+1)$st burst belonging to class 1 which arrive at the queue during period k, given $(v_1^n, v_2^n, w_1^n, w_2^n)$, S_i^{n+1}, and T^{n+1}. Clearly $C_{1,3}^{n+1} = 0$ since the arriving jobs during period 3 can keep the server, and then we have

$$C_1^{n+1} = C_{1,1}^{n+1} + C_{1,2}^{n+1}. \quad (4.2)$$

During period 1, the available buffer capacity is $\min(M_2 - w_2^n, M_1 - w_1^n)$. Then we get

$$C_{1,1}^{n+1} = \begin{cases} 0, & \text{if } \ v_1^n - T^{n+1} \leq 0, \\ [v_1^n - T^{n+1} - \min(M_2 - w_2^n, M_1 - w_1^n)]^+, \\ \qquad \text{if } \ T^{n+1} < v_1^n \leq T^{n+1} + S_1^{n+1}, \\ [S_1^{n+1} - \min(M_2 - w_2^n, M_1 - w_1^n)]^+, \\ \qquad \text{if } \ T^{n+1} + S_1^{n+1} < v_1^n. \end{cases} \tag{4.3}$$

The number of lost jobs during periods 2 can be obtained as we did in (4.1):

$$C_{1,2}^{n+1} = \begin{cases} (v_2^n - T^{n+1} - M_2 + w_2^n - C_{1,1}^{n+1})^+, & \text{if } \ 0 < v_2^n - T^{n+1} \leq S_1^{n+1}, \\ (S_1^{n+1} - M_2 + w_2^n - C_{1,1}^{n+1})^+, & \text{if } \ v_1^n \leq T^{n+1} + S_1^{n+1} < v_2^n, \\ 0, & \text{otherwise.} \end{cases} \tag{4.4}$$

Using (4.1) $\sim$ (4.4), the loss probability for the class i jobs denoted by $P_{loss,i}$ is obtained by

$$P_{loss,i} = \sum_{(v_1,v_2,w_1,w_2)\in\mathcal{U}} \sum_{T=0}^{v_2-1} \sum_{S_i=1}^{S_{max}} P(v_1,v_2,w_1,w_2)P(T)P(S_i)C_i / \sum_{S_i=1}^{S_{max}} P(S_i)S_i$$

where C_i means C_i^{n+1} given $(v_1^n, v_2^n, w_1^n, w_2^n) = (v_1, v_2, w_1, w_2)$, $S_i^{n+1} = S_i$, and $T^{n+1} = T$.

4.2 Waiting Time Distributions for Class 1 Jobs

Now, we get the waiting time distribution for the class 1 jobs denoted by W_1, assuming FIFO discipline on a burst basis, that is, the jobs of the nth burst have the higher priority than any jobs of the $(n+1)$st burst whenever they arrive. We define the waiting time distribution so that it satisfies the following equation:

$$\sum_{k=0}^{\infty} Pr[W_1 = k] + P_{loss,1} = 1.$$

Because of FIFO discipline on a burst basis, the waiting times of all the jobs in a burst are same as long as no jobs in the burst is lost, which are equal to $(v_1^n - T^{n+1} + w_1^n)^+$ for the jobs of the $(n+1)$st burst. If some jobs are lost, the waiting time of the jobs arriving after the loss are less than the jobs arriving before the loss by the number of lost jobs. Here we classify three periods as we did in calculating the job loss probabilities, and denote the waiting time of the jobs of the nth burst belonging to class 1 which arrive at the queue during period k by $W_{1,k}^n (k = 1, 2, 3)$. Note that the waiting time of jobs in the same period is same since the lost jobs, if any, are the jobs which arrive on the last slots of each period. Then we get

$$W_{1,k}^{n+1} = (v_1^n - T^{n+1} + w_1^n - \sum_{j=1}^{k-1} C_{1,j}^{n+1})^+.$$

Another necessary information for the waiting time distribution is the number of jobs which can enter the queue, that is, the number of jobs whose waiting time is $W_{1\ k}^{n+1}$. Let $D_{1\ k}^n$ be the number of the jobs of the the nth burst belonging to class 1 which enter the queue during period k. $D_{1,k}^n$ can be easily written using the number of lost jobs during each period, (4.3) and (4.4), as follows:

$$D_{1,1}^{n+1} = \begin{cases} 0, & \text{if } v_1^n - T^{n+1} \le 0, \\ v_1^n - T^{n+1} - C_{1,1}^{n+1}, & \text{if } T^{n+1} < v_1^n \le T^{n+1} + S_1^{n+1}, \\ S_1^{n+1} - C_{1,1}^{n+1}, & \text{if } T^{n+1} + S_1^{n+1} < v_1^n, \end{cases} \tag{4.5}$$

$$D_{1,2}^{n+1} = \begin{cases} v_2^n - T^{n+1} - C_{1,2}^{n+1}, & \text{if } v_1^n \le T^{n+1} < v_2^n \le T^{n+1} + S_1^{n+1}, \\ S_1^{n+1} - C_{1,2}^{n+1}, & \text{if } v_1^n \le T^{n+1}, T^{n+1} + S_1^{n+1} < v_2^n, \\ v_2^n - v_1^n - C_{1,2}^{n+1}, & \text{if } T^{n+1} < v_1^n, v_2^n \le T^{n+1} + S_1^{n+1}, \\ S_1^{n+1} + T^{n+1} - v_1^n - C_{1,2}^{n+1}, & \text{if } T^{n+1} < v_1^n \le T^{n+1} + S_1^{n+1} < v_2^n, \\ 0, & \text{otherwise}, \end{cases} \tag{4.6}$$

and

$$D_{1,3}^{n+1} = \begin{cases} S_1^{n+1}, & \text{if } v_2^n \le T^{n+1}, \\ S_1^{n+1} + T^{n+1} - v_2^n, & \text{if } T^{n+1} < v_2^n \le T^{n+1} + S_1^{n+1}, \\ 0, & \text{otherwise.} \end{cases} \tag{4.7}$$

Using (4.5) $\sim$ (4.7), the waiting time distribution for the class 1 jobs is obtained by

$$\begin{aligned} Pr[W_1 = j] \quad = \quad & \sum_{(v_1,v_2,w_1,w_2)\in\mathcal{U}} \sum_{T=0}^{\infty} \sum_{S_1=1}^{S_{max}} P(v_1,v_2,w_1,w_2)P(T)P(S_1) \\ \times \quad & \sum_{k=1}^{3} \mathbf{1}_{[W_{1,k}=j]} D_{1,k} / \sum_{S_1=1}^{S_{max}} P(S_1)S_1 \\ & (j = 0, 1, \cdots, S_{max} + M_1), \end{aligned}$$

where $D_{1,k}$ and $W_{1,k}$ means $D_{1,k}^{n+1}$ and $W_{1,k}^{n+1}$ given $(v_1^n, v_2^n, w_1^n, w_2^n) = (v_1, v_2, w_1, w_2)$, $S_1^{n+1} = S_1$, and $T^{n+1} = T$,respectively, and

$$\mathbf{1}_{[W_{1,k}=j]} = \begin{cases} 1, & \text{if } W_{1,k} = j, \\ 0, & \text{otherwise.} \end{cases}$$

Note that if $T^{n+1} \ge S_{max} + M_1$, then $W_{1,k}^{n+1} = 0$ because the jobs find no class 1 jobs in bursts before $(n+1)$st. For the case of FIFO disciplineon on a job basis the maximum waiting time is bounded by M_1, but it is difficult to obtain the waiting time distribution since we use the embedded Markov chain at the arrival instant of bursts.

4.3 Mean Waiting Times for Class 2 Jobs

Next, let consider the waiting time for class 2 jobs. If we neglect the arrivals after the tagged bursts, we can obtain the pseudo waiting time distribution for class 2 jobs in the same way as the previous subsection. This is not the real waiting time distribution, because the class 2 jobs may be passed by class 1 jobs of the posterior burst. The additional delay time by an arrival of the class 1 jobs in the (n+1)st burst during period k is the same as $D^{n+1}_{1,k}$, the number of the jobs of the the $(n+1)$st burst belonging to class 1 which enter the queue during period k. The number of class 2 jobs passed by the class 1 jobs in the $(n+1)$st burst during period k denoted dy $N^{n+1}_k (k = 1, 2, 3)$ can be also obtained by

$$N^{n+1}_1 = \begin{cases} \alpha^n, & \text{if } T^{n+1} < v^n_1, \\ 0, & \text{otherwise,} \end{cases} \tag{4.8}$$

$$N^{n+1}_2 = \begin{cases} \alpha^n - (\beta^n)^+, & \text{if } v^n_1 \leq T^{n+1} < v^n_2, \\ \alpha^n, & \text{if } T^{n+1} < v^n_1 \leq T^{n+1} + S^{n+1}_1, \\ 0, & \text{otherwise,} \end{cases} \tag{4.9}$$

and

$$N^{n+1}_3 = \begin{cases} [\alpha^n - (\beta^n)^+]^+, & \text{if } v^n_2 \leq T^{n+1}, \\ \alpha^n - (\beta^n + C^{n+1}_2)^+, & \text{if } v^n_1 \leq T^{n+1} < v^n_2 \leq T^{n+1} + S^{n+1}_1, \\ \alpha^n - (\beta^n + C^{n+1}_1 + C^{n+1}_2)^+, & \\ & \text{if } T^{n+1} < v^n_1, v^n_2 \leq T^{n+1} + S^{n+1}_1, \\ 0, & \text{otherwise,} \end{cases} \tag{4.10}$$

where $\alpha^n = v^n_2 - v^n_1 + w^n_2 - w^n_1$ and $\beta^n = T^{n+1} - v^n_1 - w^n_1$. However, it is difficult to get the joint distribution of the pseudo waiting time and the additional delay time. Therefore, we can not get the waiting time distribution for class 2 jobs, but the mean waiting time can be calculated as follows: Let us assume that $(v^n_1, v^n_2, w^n_1, w^n_2)$, S^{n+1}_1, and T^{n+1} are given. If the $(n+1)$st burst belongs to class 2, the total amount of pseudo waiting time of the jobs in $(n+1)$st burst becomes $(v^n_2 - T^{n+1} + w^n_2)^+(S^{n+1}_2 - C^{n+1}_2) - C^{n+1}_2(S^{n+1}_2 + T^{n+1} - v^n_2)^+$. If the $(n+1)$st burst belongs to class 1, the total amount of additional delay which the burst bring to class 2 jobs is expressed by $\sum_{k=1}^{3} D^{n+1}_{i,k} N^{n+1}_k$ using (4.5) ~ (4.7) and (4.8) ~ (4.10). Since the ratio between arrival rates of class 1 bursts and class 2 bursts is r_1/r_2, the mean waitin time for class 2 jobs is obtained by

$$\begin{aligned} \bar{W_2} &= \sum_{(v_1,v_2,w_1,w_2)\in\mathcal{U}} \sum_{T=0}^{\infty} \sum_{S_2=1}^{S_{max}} P(v_1, v_2, w_1, w_2) P(T) P(S_2) \\ &\times \quad [\frac{r_1}{r_2} \sum_{k=1}^{3} D_{1,k} N_k + (v_2 - T + w_2)^+(S_2 - C_2) - C_2(S_2 + T - v_2)^+] \\ &/ \quad \sum_{S_2=1}^{S_{max}} P(S_2) S_2, \end{aligned}$$

where $D_{1,k}$, N_k, and C_2 means $D^{n+1}_{1,k}$ N^{n+1}_k and C^{n+1}_2 given $(v^n_1, v^n_2, w^n_1, w^n_2) = (v_1, v_2, w_1, w_2)$, $S^{n+1}_1 = S_1$, and $T^{n+1} = T$,respectively.

Using the mean waiting time, we can get the mean queue length for class i jobs, $\bar{L_i}$ using the Little's law, i.e.,

$$\bar{L}_i = \bar{W}_i \bar{S}_i r_i / \bar{T},$$

where $\bar{S}_i$, $\bar{T}$, and $\bar{W}_i$ denote the first moments of S_i, T, and W_i, respectively.

5. NUMERICAL EXAMPLES

In this section, we present some numerical examples. We consider two examples as shown in Table 1. We assume that the inter-arrival time of bursts, T, is uniformly distributed from T_{min} to T_{max} and that the number of jobs in bursts of each class, S_1 and S_2, are uniformly distributed from S_{min} to S_{max}. C_T^2 and C_S^2 respectively denote the squared coefficient of variation of T and S_i. Note that C_S^2 in Ex.2 is larger than the one in Ex.1.

For both examples, we fix the total buffer capacity as $M_2 = 5$ and the probability of the classes that the bursts belong to as $r_1 = r_2 = 0.5$. Then, we observe the behaviors of the job loss probability and the mean waiting time as the capacity of the shared part increases. Table 2 shows the number of states of the embedded Markov chain (v_1, v_2, w_1, w_2) for each example. We illustrate the job loss probability and the mean waiting time for each class in Fig.2 and Fig.3 respectively.

Table 1 Parameters of examples

Ex.	Inter-arrival time				The number of jobs				Traffic intensity
	T_{min}	T_{max}	$\bar{T}$	C_T^2	S_{min}	S_{max}	$\bar{S}$	C_S^2	
1	1	9	5	0.2666	2	5	3.5	0.1020	0.7
2	1	9	5	0.2666	1	6	3.5	0.2381	0.7

Table 2 The number of states (v_1, v_2, w_1, w_2)

Capacity of the shared part	0	1	2	3	4	5
Ex.1	88	158	211	249	275	293
Ex.2	127	228	305	361	400	427

We can conclude the numerical results as follows:

1. The job loss probability as well as the mean waiting time increases as the coefficient of variation of S_i increases when the traffic intensity is fixed. Though it is not shown here, we can say the same thing for C_T^2.

2. As the capacity of the shared part increases, the job loss probability for class 1 drastically decreases, but the one for class 2 slightly increases. They become exactly same when $M_1 = M_2 = 5$, which is most reasonable.

3. As the capacity of the shared part increases, the mean waiting time for each class slightly increases, because the total job loss probability de-

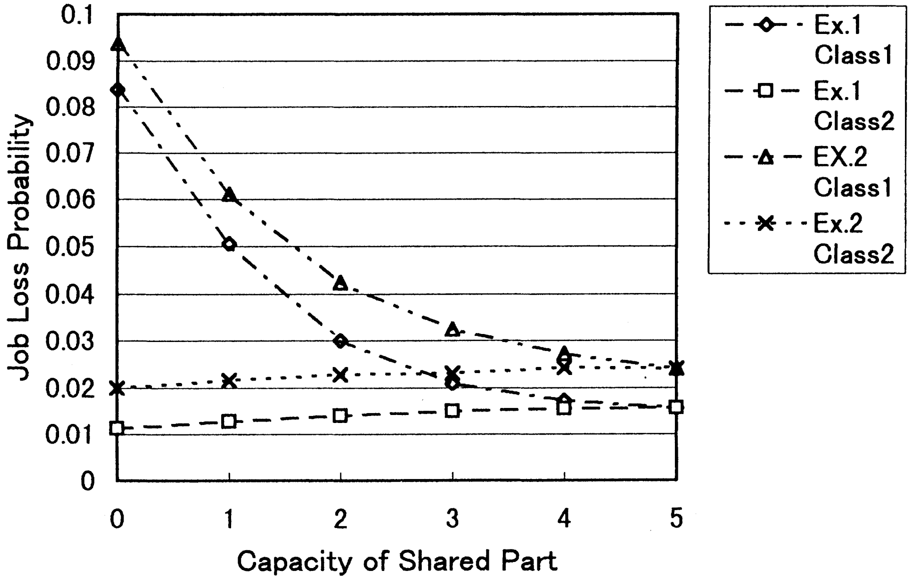

Figure 2 Job Loss Probability

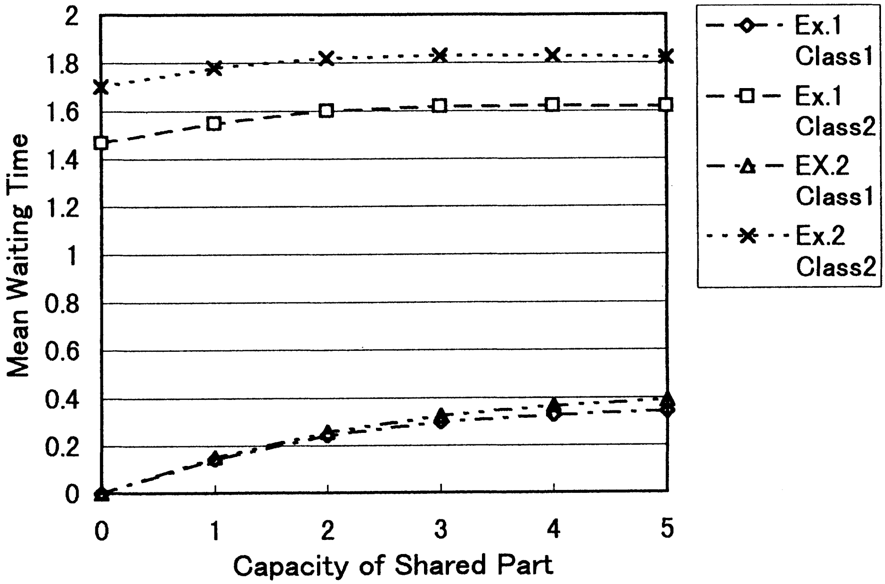

Figure 3 Mean Waiting Time

creases, and the substanial traffic intensity increases, consequently. When $M_1 = 0$, the waiting time for any class 1 jobs is zero since the class 1 jobs are not allowed to wait in the buffer in this case.

4. It is shown that the proposed buffer management is efficient to meet the quality of service requirments of different traffic types by setting the capcity of each part of buffer appropriately.

6. CONCLUDING REMARKS

We studied a discrete-time, single-server priority queue with partial buffer sharing, and proposed an efficient numerical method to exactly obtain some performance measures. We keep track of the largest numbers of remaining jobs of active bursts for each priority class rather than the numbers of remaining jobs of every active burst. Hence, we can save a lot of space and computational effort, compared with the straightforward way mentioned in Section 3. Though the number of states of the embedded Markov chain $(v_1^n, v_2^n, w_1^n, w_2^n)$ increases as the maximum burt size and/or the capacity of the buffer increases, the computational complexity is still polynomial order.

We can extend this work to multi-server systems. In the two-server systems, we have to keep track of the first and second largest numbers of remaining jobs of active bursts for each priority class, and the resulting embedded Markov chain is $(v_{1,1}^n, v_{1,2}^n, v_{2,1}^n, v_{2,2}^n, w_1^n, w_2^n)$, where $v_{i,j}^n$ is the jth largest number of remaining jobs among active bursts of class 1 to class i. The extension to three classes job systems is also possible. In these systems, however, the number of states rapidly increases as the maximum burt size and/or the capacity of the buffer increases, and they become intractable even for small systems.

7. REFERENCES

Brandt,A., Brandt,M. and Sulanke,H. (1990) A Single Server Model for Packetwise Transmission of Messages. *Queueing Systems.* **6**, 287-310.

Devault,M., Cochennec,J.-Y. and Servel,M. (1988) The 'Prelude' ATD Experiment: Assessments and Future Prospects. *IEEE J. SAC* **6**, 1528-1537.

Gravey,A. and Hebuterne,G. (1991) Mixing Time and Priorities in a Single Server Queue, in *Proc. of ITC-13,* Copenhagen, Denmark, 47-52.

Händel,H. and Huber,M.N. (1991) *Integrated Broadband Networks,* Addison-Wesley.

Hashida,O. and Takahashi,Y. (1991) A Discrete-time Priority Queue with Switched Batch Bernoulli Process Inputs and Constant Service Time, in *Proc. of ITC-13,* Copenhagen, Denmark, 521-526.

Hashida,O., Takahashi,Y. and Shimogawa,S. (1991) Switched Batch Bernoulli Process(SBBP) and the Discrete-Time SBBP/G/1 Queue with Application to Statistical Multiplexer Performance. *IEEE J. Select. Areas Commun.,* **SAC-9** 394-401.

Johnson,M.A. and Narayana,S. (1996) Descriptors of Arrival-Process Burstiness with Application to the Discrete Markovian Arrival Process. *Queueing Systems,* **23** 107-130.

Kino,I. and Miyazawa,M. (1993) The Stationary Work in System of a G/G/1 Gradual Input Queue. *J. Appl. Prob.,* **30** 207-222.

Kröner,H. (1990) Comparative Performance Study of Space Priority Mechanisms for ATM Networks, in *Proc. of INFOCOM'90,* San Francisco, 1136-1143.

Miyazawa,M and Yamazaki,G. (1992) Loss Probability of a Burst Arrival Finite Queue with Synchronized Service. *Probability in the Engineering and Informational Sciences,* **6** 201-216.

Morris,R. (1981) An Algorithmic Technique for a Class of Queueing Models with Packet Switching Applications, in *Proc. IEEE ICC'81,* 41.2.1-41.2.8.

Neuts,M.F. (1990) On Viterbi's Formula for the Mean Delay in a Queue of Data Packets. *Commun. Statist.-Stochastic Models,* **6** 87-98.

Takine,T., Sengupta,B. and Hasegawa,T. (1994) An Analysis of a Discrete-time Queue for Broadband ISDN with Priorities among Traffic Classes. *IEEE Trans. Comm.,* **42** 1837-1845.

Yamashita,H. (1994) Numerical Analysis of a Discrete-Time Finite Capacity Queue with a Burst Arrival. *Annals of Operations Research* **49** 101-110.

Yamashita,H. (1994) Discrete-Time Analysis of a Classified Multi-server Queue with Burst Arrivals and a Shared Buffer. *Queueing Systems,* **18** 167-182.

8. BIOGRAPHY

Hideaki Yamashita is currently Associate Professor in the Department of Business Administration at Komazawa University, Tokyo, Japan. He received his B.S., M.S., and ph.D. in Mechanical Engineering from Sophia University, Tokyo. His research interests include queueing analyses of production systems, high speed networks, and road traffics. He is a member of the Operations Research Society of Japan, the Institute of Electronics, Information and Communication Engineers, and the Japan Society of Mechanical Engineers.