Performance Analysis of the DAR(1)/D/c priority queue under Partial Buffer Sharing Policy

Gang Uk Hwang[a1] and Bong Dae Choi[b]

[a]Division of Applied Mathematics
Korea Advanced Institute of Science and Technology
373-1 Guseong-dong, Yuseong-gu, Taejeon, 305-701, South Korea

[b]Department of Mathematics and
Telecommunication Mathematics Research Center
Korea University
1 Anam-dong, Sungbuk-ku, Seoul, 136-701, South Korea

Abstract

We analyze a multi-server priority queueing system with partial buffer sharing, where the input is a discrete autoregressive process of order 1 (DAR(1)) which is known as a good mathematical model for a VBR-coded teleconference video traffic. We assume that arriving packets are classified into two priority classes, say, high priority class and low priority class based on their importance. A threshold T is set to limit the accessibility of low priority packets to the buffer. When the partial buffer sharing is applied to the real time traffic such as teleconference video traffic, it is known that it can decrease the queueing delay at the expense of the loss of low priority packets which is less important. Since the queueing delay is more important than the loss probability for real time traffic, it is important to analyze the queueing delay of DAR(1) arrivals under the partial buffer sharing policy. Based on the Wiener-Hopf factorization of the GI/GI/1 queue, we obtain the waiting time distribution of a packet which is not discarded at its arrival in the steady state. Numerical examples are provided to show the feasibility of our analysis. We also show that the partial buffer sharing policy significantly decreases the waiting time as the value of threshold increases.

Keywords: Discrete Autoregressive Arrivals, Partial Buffer Sharing, Performance, Queueing Delay, Priority Queue

[1]Corresponding author

1 Introduction

The partial buffer sharing policy is an overload control mechanism to meet the required QoS (Quality of Service) of different traffic in B-ISDNs (Broadband - Integrated Services Digital Networks). In the partial buffer sharing policy, the input packets are classified into priority classes according to their importance, and threshold value(s) is(are) assigned to each priority class to limit their access to the network resources. So, when a packet, belonging to a certain priority class with threshold value T, arrives at the system, the packet is accepted to be stored in the buffer only if the queue length, defined by the total number of packets in the system, at the packet arrival time is less than or equal to T. Otherwise, the packet is discarded. By doing this, higher priority packets have less loss probability and shorter queueing delay at the expense of loss of lower priority packets. It has been shown that the partial buffer sharing policy not only performs well to meet the various QoS requirements for multiple priority classes, but also can be implemented easily, so it has been proposed as a good candidate for an overload control mechanism and analyzed intensively [5, 7, 8, 16, 17]. Lucantoni *et al.* [8] analyzed the partial buffer sharing policy when the arrival process is a Poisson process. Kröner *et al.* [7] examined the performance of the partial buffer sharing policy based on the M/G/1/N queueing system. They also considered an ON and OFF source as input traffic and provided an approximation for the loss probability of each class. Hou *et al.* [5] analyzed the partial buffer sharing policy with a multiplexing of ATM CBR (Continuous Bit Rate) traffic and bursty traffic. Yin *et al.* [17] considered a packet voice multiplexer where packets are selectively discarded during periods of congestion. They assumed that the input traffic is a Markov modulated fluid flow (MMFF) and analyzed the system to show that their control mechanism achieves significant improvement in system performance. Yegani [16] considered an ATM multiplexer whose inputs consist of voice and data traffic. He modelled the input traffic as a Markov modulated Poisson process (MMPP) and analyzed the MMPP/G/1/N queue to examine the system performance. Some variants of the partial buffer sharing policy and other overload control mechanisms can be found in the literature. Interesting readers refer to [1, 3, 4, 10, 6, 14, 15, 1] and the references therein.

Although the earlier studies have been focused on the MMPP or its variants as the input traffic model, another traffic models should be also considered to more deeply understand the performance behavior of the partial buffer sharing policy. One of important traffic models other than the MMPP or its variants is the discrete autoregressive process of order 1 (shortly, DAR(1)). The DAR(1) process was shown to be a good candidate to model VBR-coded teleconference video traffic at frame level by Elwalid *et al.* [9]. It is shown that the DAR(1) process is a Markov process whose autocorrelation function is geometrically decaying and it exhibits general marginal distribution [12, 13]. In spite of such merits and its importance as a stochastic model, relatively little attention has been paid to the DAR(1) process in performance analysis because the analysis of a queueing system with the DAR(1) process is not easy to develop. Recently, some analytic techniques for the queueing system with the DAR(1) process and no control mechanism, were developed by Hwang *et al.* [12, 13]. However, to the best of the authors' knowledge there is no study on the performance analysis of the partial buffer sharing policy when the input is according to the DAR(1) process.

In this paper, we analyze an infinite buffer multi-server queueing system with the partial buffer sharing policy where the input traffic is modelled by the DAR(1) process. We assume that, when the input generates packets, it classifies the packets into two classes, say a high priority class and a low priority class, according to their importance. There is a single infinite size buffer in the system which accommodates arriving packets and serves packets in the buffer in a FIFO (First In First Out) manner. In the analysis, a threshold T is set to limit the accessibility of low priority packets to the buffer and there is no limit for high priority packets to access to the buffer because we consider an infinite size buffer. Since for real time traffic the queueing delay is more important QoS (Quality of Service) parameter than the loss probability, in our model the partial buffer sharing policy is used to decrease the waiting time of packets in the buffer (i.e., the queueing delay) at the expense of the loss of low priority packets. Note that a previous study revealed that the partial buffer sharing policy can reduce the waiting time in the buffer while maintaining the least degradation in information loss when it is applied to the buffer which accommodates real time traffic such as packetized voice traffic [17]. Furthermore, we can easily implement the partial buffer sharing in real situation. Therefore, it is worth while to analyze the waiting time distribution under the partial buffer sharing policy when we have real time traffic.

The rest of the paper is organized as follows: In section 2, we describe the DAR(1) arrival process and the system modelling. In section 3, we analyze the system by using the Wiener-Hopf factorization of the GI/GI/1 queue and in section 4 we compute the waiting time distribution of an arbitrary packet which is not discarded at its arrival instant. In section 5, we provide numerical results to show the feasibility of our analysis and to examine the performance behaviors of the partial buffer sharing policy when the input traffic is according to the DAR(1) arrivals. Conclusions are given in section 6.

2 The Model

We consider an infinite buffer multi-server queueing system with the partial buffer sharing policy where the time axis is divided into slots of equal size. The packet arrival process is according to a discrete autoregressive process of order 1 (shortly, DAR(1) process).

In order to define a DAR(1) process $\{A_m\}_{m\geq 0}$, we should introduce two independent random sequences $\{\alpha_m\}$ and $\{B_m\}$. Here, $\{\alpha_m\}$ is an i.i.d. (independent and identically distributed) sequence of Bernoulli random variables with state space $\{0,1\}$ and the distribution of α_m is given by:

$$P\{\alpha_m = 1\} = p, \qquad 0 < p \leq 1.$$

$\{B_m\}$ is an i.i.d. sequence with probability mass function denoted by

$$P\{B_m = k\} = b_k, \qquad k \geq 0.$$

Then, the discrete autoregressive process $\{A_m\}$ of order 1 is defined as follows:

$$\begin{aligned} A_0 &= B_0 \\ A_{m+1} &= (1-\alpha_m)A_m + \alpha_m B_m, \quad m \geq 0. \end{aligned}$$

That is, A_m is equal to its previous value A_{m-1} with probability $1-p$, and it is equal to a new random variable B_m with probability p. Accordingly, the higher p the lower correlation between A_m and A_{m+1}. For other interesting properties of the DAR(1) process, refer to [12, 13]. Since we view the DAR(1) process as our packet arrival process in this paper, A_m is interpreted as the number of packets (or the size of a batch) arriving in the m-th slot.

We assume that, when a packet is generated, it is classified into two priority groups according to its importance. So, to model such situation we further assume that the sequence $\{B_m\}$ consists of two sub-processes $\{B_m^{(H)}\}$ and $\{B_m^{(L)}\}$ satisfying $B_m = B_m^{(H)} + B_m^{(L)}$. Here, $B_m^{(H)}$ denotes the number of high priority packets and $B_m^{(L)}$ denotes the number of low priority packets, both of which are (possibly) generated in the m-th slot. Note that $B_m^{(H)}$

and $B_m^{(L)}$ may be dependent with each other, in general. Hence, the discrete autoregressive arrivals $\{A_m\}$ considered in the paper is actually given by:

$$A_{m+1} = (1 - \alpha_m)A_m + \alpha_m(B_m^{(H)} + B_m^{(L)}).$$

For the analysis, we assume the joint probability mass function of $B_m^{(H)}$ and $B_m^{(L)}$ is independent of m and denoted by

$$P\{B_m^{(H)} = k_h, B_m^{(L)} = k_l\} = b(k_h, k_l).$$

Further, the marginal distributions for both sub-processes are denoted by

$$P\{B_m^{(H)} = k_h\} = b_k^{(H)}, P\{B_m^{(L)} = k_l\} = b_k^{(L)}, \qquad k_h \geq 0, k_l \geq 0.$$

We assume that there are $c(\geq 1)$ server(s) in the system and accordingly, at most c packets in the queue can be served in a slot simultaneously. The service discipline is FIFO (First In First Out). We also assume that a newly arriving packet enters the system immediately after the beginning of a slot, i.e., early arrival model, so it can be served immediately after its arrival if the queue length is less than c at its arrival time. We assume that the system can always accept at least c newly arriving packets in each slot. We will explain the details of accepting packets shortly.

We assume a threshold T is given in the system. So, when a batch consisting of high and low priority packets arrives at the system and the queue length at its arrival epoch does not exceed the threshold T, all the packets in the batch are stored in the queue. However, when the queue length at the batch arrival epoch exceeds the threshold T, the number of packets allowed to be stored depends on the number of high priority packets in the batch as follows: If the queue length at the batch arrival epoch exceeds the threshold T and the number c_h of high priority packets in the batch is greater than or equal to c, then only high priority packets are stored. On the other hand, if the number c_h of high priority packets in the batch is less than c, then $\min(c - c_h, c_l)$ low priority packets in the batch are stored as well as c_h high priority packets where c_l is the number of low priority packets in the batch. Therefore, the system tries to accept as many low priority packets as possible as long as the system performance does not become worse (or the queue length does not increase due to the acceptance of low priority packets) even when the queue length exceeds the threshold T. By doing this we can achieve a good overload control mechanism which decrease the queueing delay while degrading the loss probability of low priority packets.

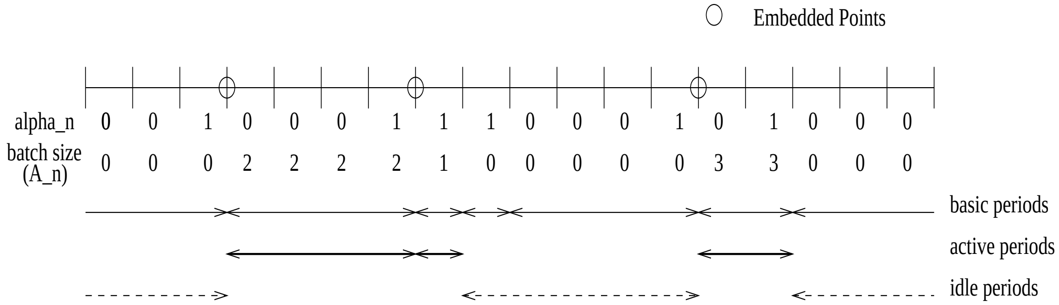

Figure 1: Basic Periods and Embedded Points

3 Analysis

3.1 Preliminaries

For the analysis we first explain basic, active and idle periods of the arrival process introduced in [13]. First, let $1 = m_0 < m_1 < m_2 < \cdots$ be the slot numbers satisfying $\alpha_{m_k-1} = 1$ for $k \geq 1$. A basic period is then defined by the time interval from slot m_k to slot $m_{k+1} - 1$ for $k \geq 0$. Refer to figure 1. In figure 1, we have $m_0 = 1, m_1 = 4, m_2 = 8, m_3 = 9, m_4 = 10, \cdots$. Hence the first basic period consists of slot 1 to slot 3, the second basic period consists of slot 4 to slot 7, and so on.

Based on the concept of the basic period, an active period is defined by a basic period during which there arrives a batch of size greater than 0. An idle period is then defined by the time interval between two consecutive active periods. Note that we can have an active period immediately followed by the next active period. In this case, we have an idle period of length 0. Without loss of generality we may assume that the first active period starts from slot 1.

For later use, we introduce two random variables X_n and Y_n which denote the lengths of the n-th active period and the n-th idle period, respectively, for $n \geq 1$. Then it is easy to show

$$\begin{aligned} P\{X_n = l\} &= (1-p)^{l-1}p, \qquad l \geq 1, \\ P\{Y_n = 0\} &= 1 - b(0,0), \\ P\{Y_n = m\} &= b(0,0)[1 - p(1 - b(0,0))]^{m-1}p(1 - b(0,0)), m \geq 1. \end{aligned}$$

Let $\bar{B}_n^{(H)}$ denote the number of high priority packets, and $\bar{B}_n^{(L)}$ denote the number of low priority packets arriving during a slot of the n-th active period. We use $\bar{B}_n = \bar{B}_n^{(H)} + \bar{B}_n^{(L)}$. Note that $\bar{B}_n^{(H)}$ and $\bar{B}_n^{(L)}$ remain the same during the n-th active period and that $\bar{B}_n = \bar{B}_n^{(H)} + \bar{B}_n^{(L)}$ should be greater than 0. We then have

$$P\{\bar{B}_n^{(H)} = k_h, \bar{B}_n^{(L)} = k_l\} = \frac{P\{B_n^{(H)} = k_h, B_n^{(L)} = k_l\}}{1 - P\{B_n^{(H)} + B_n^{(L)} = 0\}}, k_h, k_l \geq 0, k_h + k_l > 0.$$

Note that $\bar{B}_n^{(H)}$ and $\bar{B}_n^{(L)}$ are essentially the same as $B_n^{(H)}$ and $B_n^{(L)}$, respectively, except for conditioning on $B_n^{(H)} + B_n^{(L)} > 0$.

3.2 The Queue Length Distribution

Now we are ready to analyze our system. For doing this, we first construct an embedded Markov chain for the queue length where the beginnings of active periods are considered as our embedded points. For the time being we assume that the service discipline is LIFO (Last In First Out) for convenience and simplicity (The reason why we do this will be clear in the analysis). Note that this assumption does not make any difference in the stationary distribution of the queue length at the embedded points.

Let q_n denote the queue length at the n-th embedded point for $n \geq 1$. Then, if $q_n \leq T$, then both priority packets in a batch newly arriving at the system are stored in the queue until the queue length exceeds the threshold value T during the n-th active period. Once the queue length exceeds T during the n-th active period, some of low priority packets in a batch newly arriving at the system, depending on the number of high priority packets in the same batch, are discarded during the remaining slots of the n-th active period.

For the analysis, when $q_n = m(\leq T)$, we introduce a random variable $\eta_{n,m}$ defined as follows:

$$\eta_{n,m} = \begin{cases} (\bar{B}_n - c)X_n, & \text{if } \bar{B}_n \leq c, \text{ or } \bar{B}_n > c \text{ and} \\ & \quad X_n \leq \lceil (T+1-m)/(\bar{B}_n - c) \rceil \\ & \\ (\bar{B}_n - c)\lceil (T+1-m)/(\bar{B}_n - c) \rceil & \text{if } \bar{B}_n > c \text{ and} \\ +(X_n - \lceil (T+1-m)/(\bar{B}_n - c) \rceil) & \quad X_n > \lceil (T+1-m)/(\bar{B}_n - c) \rceil, \\ \qquad \times (\bar{B}_n^{(H)} - c)^+, & \end{cases}$$

where $\lceil x \rceil$ means the smallest integer greater than or equal to x. The meaning of $\eta_{n,m}$ is as follows: When $\bar{B}_n \leq c$, all arriving packets are immediately served after their arrival and $\eta_{n,m}$ denotes the number of packets in the queue which are *served* during the n-th active period. When $\bar{B}_n > c$, $\eta_{n,m}$ denotes the number of packets *stored* in the queue ($\eta_{n,m}$ does not include the packets served immediately after its arrival) during the n-th active period. So, if the length X_n of the n-th active period is less than or equal to $\lceil (T+1-m)/(\bar{B}_n-c)\rceil$, then all arriving packets are accepted. Otherwise, all arriving packets are accepted during the slots $(1, \lceil (T+1-m)/(\bar{B}_n - c)\rceil)$ of the n-th active period, and only $(\bar{B}_n^{(H)} - c)^+$ packets in each slot are stored during the rest of slots of the n-th active period.

When $q_n > T$, some of low priority packets in a batch newly arriving at the system, depending on the number of high priority packets in the same batch, are discarded during the n-th active period. Similarly above, when $q_n > T$, we introduce a random variable ζ_n defined as follows:

$$\zeta_n = \begin{cases} (\bar{B}_n^{(H)} - c)X_n, & \text{if } \bar{B}_n^{(H)} \geq c \\ 0, & \text{if } \bar{B}_n^{(H)} < c, \bar{B}_n \geq c \\ (\bar{B}_n - c)X_n, & \text{if } \bar{B}_n^{(H)} < c, \bar{B}_n < c \end{cases} .$$

Based on the above observation $\{q_n\}$ have the following evolution equation:

$$\begin{aligned} q_1 &= 0, \\ q_{n+1} &= \begin{cases} (q_n + \zeta_n - cY_n)^+ & \text{if } q_n > T \\ (q_n + \eta_{n,m} - cY_n)^+ & \text{if } q_n = m, 0 \leq m \leq T. \end{cases} \end{aligned} \tag{1}$$

The general solution of the Markov chain defined as in (1) can be found in [2]. In [2] they presented an analytic method of getting the P.G.F. (Probability Generating Function) of the stationary distribution of the Markov chain, based on the Wiener-Hopf factorization. In the paper, we follow the method in [2] to obtain the P.G.F. of $\{q_n\}$. Based on the results, we also present a numerical algorithm to compute performance measures to investigate the effect of the threshold on the system performance. In the analysis, we assume $E[\zeta_n - cY_n] < 0$ for the stability of our system.

Let $g_m(z)$ be the P.G.F. of the random variable $\eta_{n,m} - cY_n$, and $f(z)$ be the P.G.F. of the random variable $\zeta_n - cY_n$. For computing $g_m(z)$ let's consider the case of $(\bar{B}_n^{(H)}, \bar{B}_n^{(L)}) = (k_h, k_l), k_h \geq 0, k_l \geq 0, k_h + k_l > 0$ in the

n-th active period. When $0 < k_h + k_l \leq c$, for $1 - p < |z|^{c-k_h-k_l}$

$$\begin{aligned}
&E[z^{\eta_{n,m}-cY_n}|(\bar{B}_n^{(H)}, \bar{B}_n^{(L)}) = (k_h, k_l)] \\
&\quad = E[z^{-cY_n}]E[z^{\eta_{n,m}}|(\bar{B}_n^{(H)}, \bar{B}_n^{(L)}) = (k_h, k_l)] \\
&\quad = E[z^{-cY_n}]\sum_{i=1}^{\infty}(1-p)^{i-1}pz^{(k_h+k_l-c)i} \\
&\quad = E[z^{-cY_n}]\frac{pz^{k_h+k_l-c}}{1-(1-p)z^{k_h+k_l-c}}, \qquad (2)
\end{aligned}$$

where $E[z^{-cY_n}]$ is given by, for $1 - p(1 - b(0,0)) < |z|^c$,

$$\begin{aligned}
E[z^{-cY_n}] &= 1 - b(0,0) + \sum_{m=1}^{\infty} b(0,0)p(1-b(0,0))[1-p(1-b(0,0))]^{m-1}z^{-mc} \\
&= 1 - b(0,0) + \frac{b(0,0)p(1-b(0,0)}{z^c - [1-p(1-b(0,0))]}.
\end{aligned}$$

When $k_h + k_l > c$, we have

$$\begin{aligned}
&E[z^{\eta_{n,m}-cY_n}|(\bar{B}_n^{(H)}, \bar{B}_n^{(L)}) = (k_h, k_l)] \\
&\quad = E[z^{-cY_n}]\sum_{i=1}^{\infty}(1-p)^{i-1}pz^{(k_h+k_l-c)\times\min\{i,\lceil(T+1-m)/(k_h+k_l-c)\rceil\}} \\
&\qquad\qquad \times z^{(k_h-c)^+\times(i-\lceil(T+1-m)/(k_h+k_l-c)\rceil)^+}. \qquad (3)
\end{aligned}$$

Therefore, from (2) and (3) we obtain $g_m(z)$.

Next we derive the P.G.F. $f(z)$ of the random variable $\zeta_n - cY_n$. From the definition of ζ_n we have

$$\begin{aligned}
f(z) &= \sum_{k_h=0}^{c-1}\sum_{k_l=0}^{\infty} P\{\bar{B}_n^{(H)} = k_h, \bar{B}_n^{(L)} = k_l\}\sum_{i=1}^{\infty}(1-p)^{i-1}pz^{(k_h+k_l-c)^-\times i}E[z^{-cY_n}] \\
&\quad + \sum_{k_h=c}^{\infty} P\{\bar{B}_n^{(H)} = k_h\}\sum_{i=1}^{\infty}(1-p)^{i-1}pz^{(k_h-c)\times i}E[z^{-cY_n}], \qquad (4)
\end{aligned}$$

where$(x)^- = \min(0, x)$ and $P\{\bar{B}_n^{(H)} = 0, \bar{B}_n^{(L)} = 0\} = 0$.

Introduce the P.G.F. of the distribution of q_n as follows:

$$P_n(z) = \sum_{k=0}^{\infty} z^k P\{q_n = k\}, \qquad n \geq 1, |z| \leq 1.$$

Then, it follows from (1) that

$$\begin{aligned} P_{n+1}(z) &= \sum_{m=0}^{T} P\{q_n = m\}P\{\eta_{n,m} - cY_n + m < 0\} \\ &\quad + \sum_{m=T+1}^{\infty} P\{q_n = m\}P\{\zeta_n - cY_n + m < 0\} \\ &\quad + \sum_{k=0}^{\infty}\sum_{m=0}^{T} z^k P\{q_n = m\}P\{\eta_{n,m} - cY_n = k - m\} \\ &\quad + \sum_{k=0}^{\infty}\sum_{m=T+1}^{\infty} z^k P\{q_n = m\}P\{\zeta_n - cY_n = k - m\} \end{aligned}$$

$$\begin{aligned} &= (P_n(z)f(z))^{[0,\infty)} \\ &\quad + \sum_{m=0}^{T} P\{q_n = m\} \\ &\qquad \times \sum_{k=0}^{\infty} z^k \left(P\{\eta_{n,m} - cY_n = k - m\} - P\{\zeta_n - cY_n = k - m\}\right) \\ &\quad + \sum_{m=0}^{T} P\{q_n = m\}\left(P\{\eta_{n,m} - cY_n + m < 0\} - P\{\zeta_n - cY_n + m < 0\}\right) \\ &\quad + \sum_{m=0}^{\infty} P\{q_n = m\}P\{\zeta_n - cY_n + m < 0\} \qquad (5) \end{aligned}$$

where

$$\left(\sum_{k=-\infty}^{\infty} a_k z^k\right)^D = \sum_{k\in D} a_k z^k.$$

Let $P(z)$ be the P.G.F. of the stationary distribution of $\{q_n\}$, i.e.,

$$P(z) = \lim_{n\to\infty} P_n(z).$$

Introducing

$$\begin{aligned} p_m &= \lim_{n\to\infty} P\{q_n = m\}, \\ \phi_m(z) &= P\{\eta_{n,m} - cY_n + m < 0\} - P\{\zeta_n - cY_n + m < 0\} \\ &\quad + z^m \left(g_m(z) - f(z)\right)^{[-m,\infty)}, \\ l(z) &= (P(z)f(z))^{(-\infty,-1]}, \end{aligned}$$

and letting $n \to \infty$ in (5) we get, for $|z| = 1$

$$P(z) = (P(z)f(z))^{[0,\infty)} + \sum_{m=0}^{T} p_m \phi_m(z) + l(1), \tag{6}$$

or, equivalently,

$$P(z)(1 - f(z)) = \sum_{m=0}^{T} p_m \phi_m(z) + l(1) - l(z), \tag{7}$$

Consider the Wiener-Hopf factorization of $1 - f(z)$ as follows:

$$1 - f(z) = R_+(z)R_-(z), \tag{8}$$

where

$$\begin{aligned} R_+(z) &= 1 - E[z^{\chi^+} I\{\nu^+ < \infty\}], \quad |z| \le 1, \\ R_-(z) &= 1 - E[z^{\chi^-}], \quad |z| \ge 1. \end{aligned}$$

Here, ν^+ is defined to be the first (strong) ascending ladder index and ν^- is defined to be the first (weak) descending ladder index, and $\chi^+ = \zeta_1 + \cdots + \zeta_{\nu^+}, \chi^- = \zeta_1 + \cdots + \zeta_{\nu^-}$. Note that both sides of (7) vanished when $z = 1$ and that, from our stability condition of the system we have $R_-(1) = 0$. As in [2], introducing, for $|z| \ge 1$

$$\tilde{R}_-(z) = \frac{R_-(z)}{1 - z},$$

and substituting (8) into (7), we get

$$P(z)R_+(z) = \sum_{m=0}^{T} p_m h_m(z) + \frac{l(1) - l(z)}{(1 - z)\tilde{R}_-(z)}, \quad |z| = 1, \tag{9}$$

where

$$h_m(z) = \frac{\phi_m(z)}{(1 - z)\tilde{R}_-(z)}. \tag{10}$$

Rewriting (9) as

$$P(z)R_+(z) - \sum_{m=0}^{T} p_m (h_m(z))^{[0,\infty)} = \sum_{m=0}^{T} p_m (h_m(z))^{(-\infty,0)} + \frac{l(1) - l(z)}{(1 - z)\tilde{R}_-(z)}, \quad |z| = 1$$

and applying Liouville's theorem, we have

$$P(z)R_+(z) - \sum_{m=0}^{T} p_m (h_m(z))^{[0,\infty)} = K,$$

where K is a constant, and finally we get

$$P(z) = \sum_{m=0}^{T} p_m R_+^{-1}(z) \left(h_m(z)\right)^{[0,\infty)} + KR_+^{-1}(z), \qquad |z| \le 1. \tag{11}$$

The values of $p_m, m = 0, \cdots, T$ in (11) can be computed from (11) itself as follows:

$$\begin{aligned} R_+^{-1}(z) \left(h_m(z)\right)^{[0,\infty)} &\triangleq \sum_{k=0}^{\infty} z^k \gamma_k^{(m)}, \\ R_+^{-1}(z) &\triangleq \sum_{k=0}^{\infty} z^k r_k, \\ p_k &= \sum_{m=0}^{T} p_m \gamma_k^{(m)} + Kr_k, k = 0, \cdots, \qquad (12) \\ P(1) &= 1. \qquad (13) \end{aligned}$$

For the details of the analysis, interested readers may refer to [2].

3.3 Numerical Algorithm

Even though we obtain the P.G.F. $P(z)$ of the stationary distribution of $\{q_n\}$, to compute the distribution numerically we should compute $\{\gamma_k^{(m)}\}$ and $\{r_k\}$. In this subsection we focus our attention on how to compute those values numerically. For the numerical computations, we first need to truncate the distribution of $f(z)$ by $[-B, A]$, i.e.,

$$f(z) = \sum_{j=-B}^{A} f_j z^j.$$

Here, we take A and B sufficiently large, so that $\sum_{j=A+1}^{\infty} f_j + \sum_{j=-\infty}^{-B-1} f_j < \epsilon$ for a small value of ϵ. Then it immediately follows that [11]

$$\begin{aligned} R_+(z) &= 1 - \sum_{i=1}^{A} a_i z^i, \qquad (14) \\ R_-(z) &= 1 - \sum_{i=-B}^{0} d_i z^i. \qquad (15) \end{aligned}$$

When $R_-(z) \neq 0$ for $|z| \geq 1, z \neq 1$ (which holds in most practical situations), we have for $|z| \geq 1, z \neq 1$

$$R_-^{-1}(z) \triangleq \frac{1}{R_-(z)} = \sum_{i=-\infty}^{\infty} s_i z^i.$$

Observing the fact that $R_-(z) < \infty$ as $|z| \to \infty$ we have $s_i = 0$ for $i > 0$. Therefore, for $|z| \geq 1, z \neq 1$

$$R_-^{-1}(z) = \sum_{i=-\infty}^{0} s_i z^i. \tag{16}$$

From the fact that $R_-^{-1}(z)R_-(z) = 1$ for $|z| = 1, z \neq 1$ and (16), we have the following equations for $\{s_n\}_{n=-\infty}^{0}$:

$$s_0 = \frac{1}{1-d_0}, \tag{17}$$

$$s_{-i} = \frac{\sum_{j=0}^{i-1} s_{-j} d_{-i+j}}{1-d_0}, \qquad i > 0, \tag{18}$$

where the sequence $\{d_i\}$ is given in (15). Hence, we can compute $\{s_n\}$ by iteration from (17) and (18).

Now letting

$$\phi_m(z) \triangleq \sum_{n=0}^{A+m} \phi_n^{(m)} z^n,$$

$$h_m(z) \triangleq \sum_{k=-\infty}^{\infty} h_k^{(m)} z^k,$$

from (10) we can obtain $\{h_k^{(m)}\}$ from the following equations:

$$h_k^{(m)} = \sum_{n=k}^{A+m} \phi_n^{(m)} s_{-n+k}, \qquad k \geq 0. \tag{19}$$

Our next step is to compute $\{\gamma_k^{(m)}\}$ and $\{r_k\}$. Observe that [11]

$$W(z) = \frac{1-\psi}{R_+(z)}, \qquad \psi = \sum_{k=1}^{A} a_k, \tag{20}$$

where $W(z) \overset{\Delta}{=} \sum_{k=0}^{\infty} w_k z^k$ is the P.G.F. of the waiting time of the GI/GI/1 queue generated by the random walk $\{\zeta_n - cY_n\}$ and the sequence $\{a_i\}$ is given in (14). We then have

$$\frac{W(z)}{1-\psi}\left(h_m(z)\right)^{[0,\infty)} = \sum_{k=0}^{\infty} z^k \gamma_k^{(m)},$$

from which $\gamma_k^{(m)}$ can be computed as follows:

$$\gamma_k^{(m)} = \frac{1}{1-\psi}\sum_{k=0}^{k} w_l h_{k-l}^{(m)}, k \geq 0 \tag{21}$$

Further, from (20) we have

$$r_k = \frac{w_k}{1-\psi}. \tag{22}$$

Finally, summing over k in (12) yields

$$\sum_{k=0}^{\infty} p_k = \sum_{m=0}^{T} p_m \sum_{k=0}^{\infty} \gamma_k^{(m)} + K \sum_{k=0}^{\infty} r_k,$$

from which and (13) we have

$$K = (1-\psi) - (1-\psi)\sum_{m=0}^{T} p_m \sum_{k=0}^{\infty} \gamma_k^{(m)}. \tag{23}$$

Again from the first $T+1$ equations of (12) and (23) we obtain $\{p_m\}_{k=1}^{T}$ and consequently the constant K. The probabilities $p_m, m > T$ can be computed from the remaining equations of (12).

4 Performance Measures

Since we are interested in the queueing delay of a packet stored in the buffer, we first have to compute the loss probability of low priority packets. For doing this, we tag an arbitrary slot during active periods in the steady state. Let $k(n)$ be the probability that the elapsed life time of the random variable X is n at our tagged slot where X is the generic random variable for $\{X_n\}$. Then we have

$$k(n) = \frac{P\{X > n\}}{E[X]} = p(1-p)^n, n \geq 0.$$

Next, let $E[\bar{B}]$ denote the mean of $\bar{B} = \bar{B}^{(H)} + \bar{B}^{(L)}$ where $\bar{B}^{(H)}$ and $\bar{B}^{(L)}$ are the generic random variables of $\{\bar{B}_n^{(H)}\}$ and $\{\bar{B}_n^{(L)}\}$, respectively, and $E[B]$ be the mean of the random variable B where B is the generic random variable of $\{B_m\}$.

Given that there are m packets in the beginning of the active period to which our tagged slot belongs, if $m \leq T$, then the loss probability $D(m)$ is given by

$$\begin{aligned} D(m) &= \sum_{k=c+1}^{\infty} \sum_{l=0}^{k} \frac{k}{E[\bar{B}]} P\{\bar{B}^{(H)} = l, \bar{B}^{(L)} = k-l\} \\ &\quad \times \sum_{n=\lceil (T+1-m)/(k-c) \rceil}^{\infty} k(n) \frac{k-l-(c-l)^+}{k} \\ &= \sum_{k=c+1}^{\infty} \sum_{l=0}^{k} \frac{1}{E[B]} P\{B^{(H)} = l, B^{(L)} = k-l\} \\ &\quad \times \sum_{n=\lceil (T+1-m)/(k-c) \rceil}^{\infty} p(1-p)^n \{k-l-(c-l)^+\} \end{aligned} \tag{24}$$

Here, the first equation is obtained from the following:

- $\frac{k}{E[\bar{B}]} P\{\bar{B}^{(H)} = l, \bar{B}^{(L)} = k-l\}$ is the probability that an arbitrary batch consists of l high priority packets and $k-l$ low priority packets for $k \geq 1$.

- $\sum_{n=\lceil (T+1-m)/(k-c) \rceil}^{\infty} k(n)$ is the probability that the queue length exceeds the threshold value T before our tagged slot.

- $\frac{k-l-(c-l)^+}{k}$ is the probability that an arbitrary packet in the batch arriving at our tagged slot is lost.

Similarly, if $m \geq T+1$, then the loss probability $D(m)$ is given by

$$\begin{aligned} D(m) &= \sum_{k=c+1}^{\infty} \sum_{l=0}^{k} \frac{k}{E[\bar{B}]} P\{\bar{B}^{(H)} = l, \bar{B}^{(L)} = k-l\} \frac{k-l-(c-l)^+}{k} \\ &= \sum_{k=c+1}^{\infty} \sum_{l=0}^{k} \frac{1}{E[B]} P\{B^{(H)} = l, B^{(L)} = k-l\} \{k-l-(c-l)^+\} \end{aligned} \tag{25}$$

Hence, from (24) and (25) the loss probability D is given by

$$D = \sum_{m=0}^{\infty} p_m D(m). \tag{26}$$

Next we derive the P.G.F. $W_a(z)$ of the waiting time W_a of an arbitrary packet (which is not discarded at its arrival) under the FIFO (First In First Out) service discipline. We tag an arbitrary packet. Let $W_a(z, k|m) = E[z^{W_a} I\{\tilde{B} = k\}|q = m]$ where q denotes the steady state version of the random variable sequence $\{q_n\}$ and $\tilde{B}$ denotes the size of the batch to which our tagged packet belongs. Then, for $k \leq c$

$$\begin{aligned} W_a(z, k|m) &= \frac{k}{E[\bar{B}]} P\{\bar{B} = k\} \sum_{n=0}^{\infty} k(n) z^{(m-(c-k)n)^+} \sum_{s=0}^{k-1} \frac{1}{k} z^s \\ &= \frac{1}{E[B]} P\{B = k\} \sum_{n=0}^{\infty} p(1-p)^n \sum_{s=0}^{k-1} z^{(m-(c-k)n)^+ + s}. \quad (27) \end{aligned}$$

Here, the value of s denotes the number of packets in front of our tagged packet in the batch to which they belong.

For $m \leq T$ and $k > c$, by the same argument given in the derivation of (24) we have

$$\begin{aligned} &W_a(z, k|m) \\ &= \sum_{l=0}^{k} \frac{k}{E[\bar{B}]} P\{\bar{B}^{(H)} = l, \bar{B}^{(L)} = k - l\} \\ &\quad \times \left\{ \sum_{n=0}^{\lceil (T+1-m)/(k-c) \rceil - 1} k(n) \sum_{s=0}^{k-1} \frac{1}{k} z^{m+(k-c)n+s} \right. \\ &\qquad + \sum_{n=\lceil (T+1-m)/(k-c) \rceil}^{\infty} k(n) \frac{l + (c-l)^+}{k} \sum_{s=0}^{l+(c-l)^+ - 1} \frac{1}{l + (c-l)^+} \\ &\qquad\qquad \times z^{m+(k-c)(\lceil (T+1-m)/(k-c) \rceil)} \\ &\qquad\qquad \left. \times z^{(l+(c-l)^+ - c)(n - \lceil (T+1-m)/(k-c) \rceil) + s} \right\}. \quad (28) \end{aligned}$$

For $m \geq T + 1$ and $k > c$ we have

$$\begin{aligned} W_a(z, k|m) &= \frac{k}{E[\bar{B}]} \sum_{l=0}^{k} P\{\bar{B}^{(H)} = l, \bar{B}^{(L)} = k - l\} \\ &\times \sum_{n=0}^{\infty} k(n) \frac{l + (c-l)^+}{k} \sum_{s=0}^{l+(c-l)^+ - 1} \frac{1}{l + (c-l)^+} z^{m+(l+(c-l)^+ - c)n + s}. \quad (29) \end{aligned}$$

Hence, from (27), (28) and (29) we have

$$W_a(z) = \frac{\sum_{m=0}^{\infty} p_m \sum_{k=1}^{\infty} W_a(z, k|m)}{1 - D}. \quad (30)$$

5 Numerical Results

In this section we investigate the effect of the partial buffer sharing policy on system performance. For numerical computations we consider the input traffic having the following probability generating function.

$$\sum_{k=1}^{N_h}\sum_{l=0}^{N_l} P\{\bar{B}^{(H)} = k, \bar{B}^{(L)} = l\} z^k w^l = z[t_h z + 1 - t_h]^{N_h - 1}[t_l w + 1 - t_l]^{N_l}$$

where N_h and N_l are finite positive integers and $0 < t_h, t_l < 1$. In this case, we can apply the numerical algorithm presented in subsection 3.3 to investigate the system performance. Let ρ be the offered load, defined by

$$\rho = (1 - b(0,0))[1 + t_h(N_h - 1) + t_l N_l].$$

Similarly we define ρ_h and ρ_l by

$$\rho_h = (1 - b(0,0))[1 + t_h(N_h - 1)], \qquad \rho_l = (1 - b(0,0)) t_l N_l.$$

In the first experiment, we vary the threshold value and examine the behavior of the waiting time distribution. Here we use $\rho = 0.85, \rho_h = 0.5, b(0,0) = 0.6, p = 0.5$ and $N_h = N_l = 10$. The results are given in figure 2. In figure 2, it is shown that the threshold value plays an important role in the waiting time distribution and as we expected the tail behaviors of the waiting time distribution become heavier as we increase the value of threshold. To check the effect of the value of p we vary the value of p from 0.5 to two different values: 0.3 and 1.0. When $p = 1.0$, arrivals are independent. In the experiment we use the same values for the other parameters as given above, and we plot the complementary waiting time distributions as we change the value of p in figure 3. As shown in the figure, when the arrivals are independent, the partial buffer sharing policy achieves significant improvement in the waiting time. On the other hand, when the arrivals are the DAR(1) arrivals, the partial buffer sharing policy also improve the performance in the waiting time, but the improvement in the waiting time is relatively less significant than in the case of independent arrivals. In addition, we find that the improvement becomes less as the arrivals have stronger time correlations. These results show that the threshold value of the partial buffer sharing policy should be larger in the case of correlated arrivals than in the case of independent or less correlated arrivals.

In our last experiment, we fix the ratio of the high priority packets and the low priority packets and change the total offered load ρ to examine the changes in the waiting time distribution. Here, we use $\rho_h/\rho = 0.7, b_0 = 0.6, p = 0.5, T = 50$ and $N_h = N_l = 10$. In figure 4 we change the total offered load ρ from 0.6 to 0.95 and the tail behaviors of the waiting time are plotted. As in previous observation, the partial buffer sharing policy achieves significant improvement in waiting time.

6 Conclusion

In the paper, we analyzed an infinite buffer multi-server queueing system with the partial buffer sharing policy, where the packet arrivals are according to a discrete autoregressive process of order 1 (DAR(1)) and either high or low priority is given to each packet. We obtained the waiting time distribution of a packet stored in the buffer in the steady state, based on the GI/GI/1 queueing theory. Some numerical results showed that the partial buffer sharing policy, when applied to the DAR(1) process, significantly decreases the waiting time at the expense of the packet loss of low priority. Since the DAR(1) process is a good candidate for modelling VBR-coded teleconference video traffic and the partial buffer sharing policy is also a good candidate for an overload control mechanism in the network, our analysis is useful in designing B-ISDN networks.

Acknowledgements

This work was supported by Korea Research Foundation Grant (KRF-2001-042-D00009). The authors would like to thank Dr. D. H. Han for his careful reading and comments on the paper. They also thank the anonymous reviewer for his valuable comments on the earlier version of the paper.

References

[1] M. M. Asrin, A Performance Analysis of the ATM Multiplexer with Selective Discarding of the Cells During Congestion, *Performance Evaluation* (2001); 45: 257–285.

[2] O.J. Boxma and V.I. Lotov, On a Class of One-dimensional Random Walks, CWI report, BS-R9537, 1995.

[3] J. Chang and C. Wu, The Effect of Prioritization on the Behavior of a Concentrator Under an Accept, Otherwise Reject Strategy, *IEEE Transactions on Communications* (1990); 38(7): 1031–1039.

[4] A. K. Choudhury and E.L. Hahne, Dynamic Queue Length Thresholds for Shared-Memory Packet Switches, *IEEE/ACM Transactions on Networking* (1998); 6(2): 130–140.

[5] T. Hou and A.K. Wong, Queueing Analysis for ATM Switching of Mixed Continuous-Bit-Rate and Bursty Traffic, *INFOCOM '90* (1990); 660-667.

[6] C. G. Kang and H. H. Tan, Queueing Analysis of Explicit Priority Assignment Partial Buffer Sharing Schemes for ATM Networks, *Proceedings fo INFOCOM '93* (1993); 810–819.

[7] H. Kröner, G. Hébuterne, P. Boyer and A. Gravey, Priority Management in ATM Switching Nodes, *IEEE Journal on Selected Areas in Communications* (1991); 9(3): 418–427.

[8] D. M. Lucantoni, S.P. Parekh, Selective Cell Discarding Mechanisms for a B-ISDN Congestion Control Architecture, *Proceedings of the Seventh ITC Seminar*, Morristown, NJ, October, paper 10.3, 1990.

[9] A. Elwalid, D. Heyman, T.V. Laksman, D. Mitra, and A. Weiss, Fundamental Bounds and Approximations for ATM Multiplexers with Applications to Video Teleconferencing, *IEEE Journal of Selected Areas in Communications* (1995); 13(6): 1004–1016.

[10] G. Galassi, G. Rigolio, L. Fratta, Bandwidth Assignment in Prioritized ATM Networks, *Proceedings of Globecom '90* (1990); 852–856.

[11] W.K. Grassmann and J.L. Jain, Numerical Solutions of the Waiting Time Distribution and Idle Time Distribution of the Arithmatic GI/G/1 Queue, *Operations Research* (1989); 37: 141–150.

[12] G.U. Hwang and K. Sohraby, On the Exact Analysis of a Discrete-time Queueing System with Autoregressive Inputs, *Queueing Systems* (2003); 43(1-2): 29–41.

[13] G.U. Hwang, B.D. Choi and J.-K. Kim, The Waiting Time Analysis of a Queue with A Discrete Autoregressive Model of Order 1, *Journal of Applied Probability* (2002); 39(3): 619–629.

[14] M.F. Neuts, A Queueing Model for a Storage Buffer in which the Arrival Rate is Controlled by a Switch with a Random Delay, *Performance Evaluation* (1985); 5: 243–256.

[15] H. Takagi, Analysis of a Finite-capacity M/G/1 Queue with a Resume Level, *Performance Evaluation* (1985); 5: 197–203.

[16] P. Yegani, Performance Models for ATM Switching of Mixed Continuous-Bit-Rate and Bursty Traffic with Threshold-Based Discarding, *ICC '92* (1992); 1621–1627.

[17] N. Yin, S. Li and T.E. Stern, Congestion Control for Packet Voice by Selective Packet Discarding, *IEEE Transactions on Communications* (1990); 38(5): 674–683.

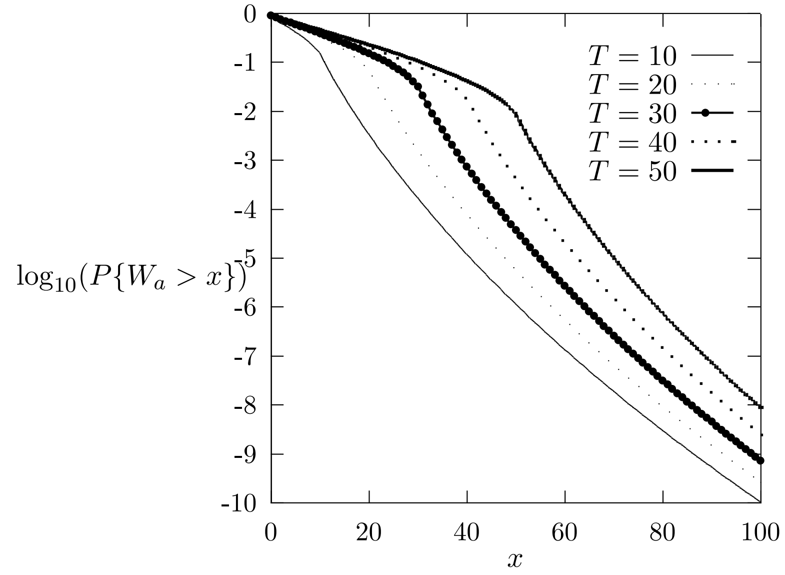

Figure 2: The effect of Threshold value T

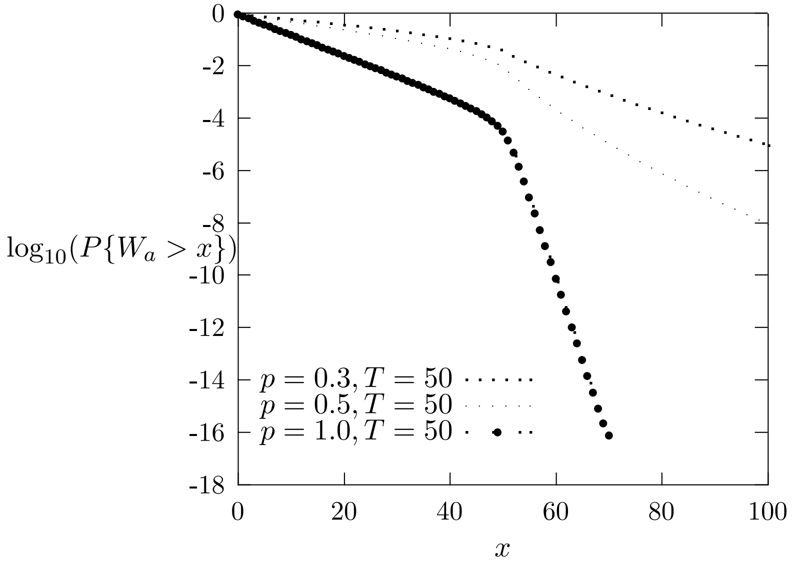

Figure 3: The effect of Correlations in Arrivals

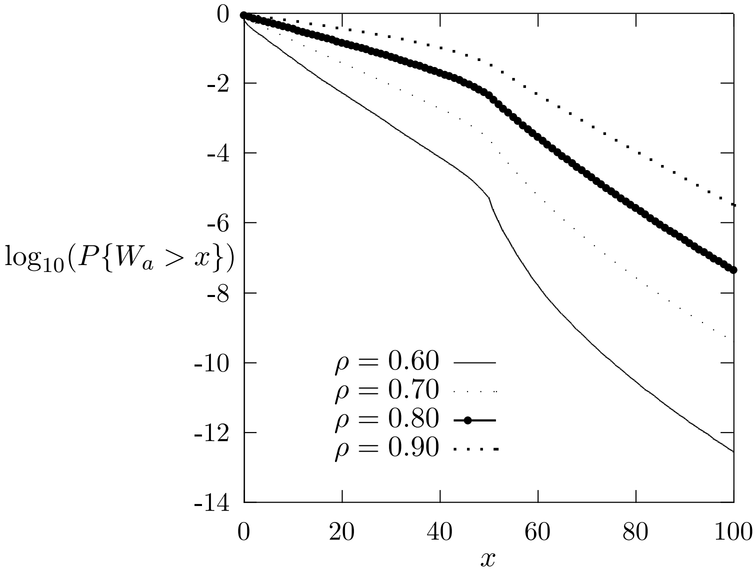

Figure 4: The effect of Change in ρ