Adaptive Bandwidth Allocation Method for Long Range Dependence Traffic

Bong Joo Kim and Gang Uk Hwang*

Division of Applied Mathematics and Telecommunication Program
Korea Advanced Institute of Science and Technology
373-1 Guseong-dong, Yuseong-gu, Taejeon, 305-701, South Korea
KimBongJoo@kaist.ac.kr
guhwang@amath.kaist.ac.kr
http://queue.kaist.ac.kr/~guhwang

Abstract. In this paper, we propose a new method to allocate bandwidth adaptively according to the amount of input traffic volume for a long range dependent traffic requiring Quality of Service (QoS). In the proposed method, we divide the input process, which is modelled by an $M/G/\infty$ input process, into two sub-processes, called a long time scale process and a short time scale process. For the long time scale process we estimate the required bandwidth using the linear prediction. Since the long time scale process varies (relatively) slowly, the required bandwidth doesn't need to be estimated frequently. On the other hand, for the short time scale process, we use the large deviation theory to estimate the effective bandwidth of the short time scale process based on the required QoS of the input traffic. By doing this we can capture the short time scale fluctuation by a buffer and the long time scale fluctuation by increasing or decreasing the bandwidth adaptively. Through simulations we verify that our proposed method performs well to satisfy the required QoS.

1 Introduction

Several traffic measurement studies in recent years have shown the existence of long-range dependence(LRD) and self-similarity in network traffic such as Ethernet LANs[6,11], variable bit rate(VBR) video traffic[2,7], Web traffic[5], and WAN traffic[15]. In addition, many analytical studies have shown that LRD network traffic can have a detrimental impact on network performance and pointed out the need of revisiting various issues of performance analysis and network design. One of the practical impacts of LRD is that buffers at switches and multiplexers should be significantly larger than those predicted by traditional queueing analysis and simulations to meet the required quality of service (QoS). This requirement for large buffers can be explained by the Noah effect and the

* This work was supported by Korea Research Foundation Grant funded by Korea Government (MOEHRD, Basic Research Promotion Fund) (KRF-2005-003-C00022).

G. Parr, D. Malone, and M. Ó Foghlú (Eds.): IPOM 2006, LNCS 4268, pp. 13–24, 2006.

Joseph effect. However, such a significantly large buffer requirement causes the inefficient use of the network resources.

To solve this problem, many studies recommend that the buffer be kept small while the link bandwidth is to be increased, and most studies, e.g. [3,9,16,17] in the open literature have been focusing on the estimation of the required bandwidth for LRD traffic based on the buffer overflow probability. However, a deterministic bandwidth allocation strategy seems not to be quite effective for LRD traffic. For example, if the traffic volume in a certain (relatively long) period of time is less than the estimated bandwidth, which can happen for LRD traffic with nonnegligible probability, then the network wastes the bandwidth for as much period of time. Therefore, it would be more effective if we can adaptively change the bandwidth allocated for LRD traffic, which is the motivation of this study.

In this paper, we propose a new method to allocate bandwidth adaptively according to the amount of input traffic volume for a long range dependent traffic requiring quality of service (QoS). We consider a discrete time queueing system with an $M/G/\infty$ input process[10,12,13,14] to model the system with LRD traffic. The $M/G/\infty$ input process is the busy server process of a discrete time infinite server system fed by Poisson arrivals of rate λ (customers/slot) with general service times. To mimic LRD traffic the discretized Pareto distribution is considered as the service time distribution.

In the proposed method, We divide the input process into two sub-processes: a long time scale process and a short time scale process. For the long time scale process we estimate the required bandwidth using the minimum mean square error(MMSE) linear prediction. Since the long time scale process varies (relatively) slowly, the required bandwidth doesn't need to be estimated frequently. On the other hand, for the short time scale process, since the required bandwidth for the long time scale process is fully captured by the linear prediction, we use the large deviation theory [4] to estimate the effective bandwidth of the short time scale process based on the required QoS of the input traffic. By doing this we can capture the short time scale fluctuation by a buffer and the long time scale fluctuation by increasing or decreasing the bandwidth adaptively. Hence, the total bandwidth allocated for the $M/G/\infty$ input process is obtained by adding the effective bandwidth of the short time scale process to the estimated bandwidth of the long time scale process.

To check the performance of our proposed method, we simulate a queueing system with the $M/G/\infty$ input process under various conditions and our simulation results show that the proposed method performs well. Further discussion on the proposed method will be given later.

The organization of this paper is as follows: In section 2, we introduce a stationary $M/G/\infty$ input process and investigate how to model LRD traffic with the stationary $M/G/\infty$ input process. In section 3, we propose our adaptive bandwidth allocation method for LRD traffic. In section 4, we present numerical results to validate our method and further discussions are also provided. Finally, we have conclusions in section 5.

2 The Stationary $M/G/\infty$ Input Process

An $M/G/\infty$ input process is the busy server process of a discrete-time infinite server system fed by Poisson arrivals with general service time. This process has been known to be useful to model LRD traffic because it is stable under multiplexing and flexible in capturing positive dependencies over a wide range of time scales[10,12,13,14]. To describe the $M/G/\infty$ input process, we assume that the time axis is divided into slots of equal size and consider a discrete time $M/G/\infty$ queueing system as follows: The customer arrival process is according to a Poisson process with rate λ, and β_{t+1} new customers arrive into the system during time slot $[t, t+1)$ for $t = 0, 1, \cdots$. Since the number of servers is infinite, customer j, $j = 1, ..., \beta_{t+1}$, immediately occupies a new server in the system after its arrival and begins its service from slot $[t+1, t+2)$, with service time $\sigma_{t+1,j}$ (slots). We assume that $\{\sigma_{t,j}, t = 1, 2, \cdots; j = 1, 2, \cdots \beta_t\}$ are i.i.d. and σ denotes a generic r.v. for $\sigma_{t,j}$.

Let $X(t)$, $t \geq 0$ denote the number of busy servers during time slot $[t, t+1)$, in other words, the number of customers still present in the $M/G/\infty$ system during time slot $[t, t+1)$. We assume that the system starts with $X(0)$ initial customers at time 0. Then, the $M/G/\infty$ input process is the busy server process $\{X(t), t = 0, 1, \cdots\}$ of the $M/G/\infty$ system described above. When $X(t)$ is used for traffic modelling, $X(t)$ is considered as the number of packets arriving during slot $[t, t+1)$.

When we assume that the initial number $X(0)$ of customers is a Poisson r.v. with parameter $\lambda E[\sigma]$, and that $\{\sigma_{0,j}, j = 1, 2, \cdots, X(0)\}$ are i.i.d. $\mathbb{N}$-valued r.v.s distributed according to the forward recurrence time $\hat{\sigma}$ associated with σ, i.e., the p.m.f. of $\hat{\sigma}$ is given by $p[\hat{\sigma} = r] \triangleq \frac{P[\sigma \geq r]}{E[\sigma]}, r = 1, 2, \cdots$, it can be shown that the resulting $M/G/\infty$ input process is stationary[10,13]. In addition, we can show that the covariance structure of $\{X(t), n = 0, 1, \cdots\}$ is given by [13]

$$\Gamma(k) \triangleq cov[X(t), X(t+k)] = \lambda E\left[(\sigma - k)^+\right], \quad t, k = 0, 1, \cdots, \tag{1}$$

which is further reduced as [10,13,18]

$$\Gamma(k) = \lambda \sum_{i=0}^{\infty} P\left[(\sigma - k)^+ > i\right] = \lambda \sum_{i=k+1}^{\infty} P[\sigma \geqslant i] = \lambda E[\sigma] P[\hat{\sigma} > k]. \tag{2}$$

In addition, the ACF (autocorrelation function) $\rho(k)$ of the stationary $M/G/\infty$ input process is obtained from (1) and (2) as follows:

$$\rho(k) \triangleq \frac{\Gamma(k)}{\Gamma(0)} = P[\hat{\sigma} > k], \quad k = 0, 1, \cdots.$$

Next, we use the stationary $M/G/\infty$ input process to model LRD traffic. To do this, we consider a discretized Pareto distribution as the service time σ of a customer in the corresponding $M/G/\infty$ queueing system. The discretized Pareto distribution is defined by

$$P[\sigma = i] \triangleq P[i \leq Y < i+1] \tag{3}$$

where the random variable Y has the Pareto distribution with the shape parameter $1 < \gamma < 2$ and the location parameter $\delta(> 0)$, given by

$$P[Y \leq x] = \begin{cases} 1 - (\frac{\delta}{x})^{\gamma} & \text{if } x > \delta, \\ 0 & \text{otherwise.} \end{cases}$$

Then, it can be easily checked that the stationary $M/G/\infty$ input process with discretized Pareto service times given above exhibits a long range dependence by using (1),(2) and (3). We omit the detailed proof in this paper due to the limitation of space. Hence, we use the the stationary $M/G/\infty$ input process with discretized Pareto service times to mathematically model LRD traffic.

3 Adaptive Bandwidth Allocation Method

In this section, we propose a new adaptive bandwidth allocation (ABA) method for LRD traffic requiring quality of service (QoS). In our ABA method, we assume that the LRD traffic is modelled by a stationary $M/G/\infty$ input process with discretized Pareto service times. To compute the bandwidth allocated for LRD traffic, we decompose the $M/G/\infty$ input traffic $X(t)$ into two components called a short time scale process, denoted by $X_s(t)$, and a long time scale process, denoted by $X_l(t)$ as follows: To decompose the input LRD traffic, a threshold value T is given *a priori*. Then, $X_s(t)$ is the number of busy servers at slot $[t-1, t]$ which became active by arrivals whose service times are less than or equal to a given threshold T, and $X_l(t)$ is the number of busy servers at slot $[t-1, t]$ which became active by arrivals whose service time is greater than T.

Due to the independence decomposition property of a Poisson process with respect to a random selection, we see that the Poisson arrivals with service times less than or equal to T generate $X_s(t)$ and accordingly, $X_s(t)$ is an $M/G/\infty$ input process with arrival rate $\lambda P[\sigma \leq T]$ and service time with p.m.f $\frac{P[\sigma=k]}{P[\sigma \leq T]}, 1 \leq k \leq T$. Similarly, $X_l(t)$ is also an $M/G/\infty$ input process with arrival rate $\lambda P[\sigma \geq T+1]$ and service time with p.m.f $\frac{P[\sigma=k]}{P[\sigma \geq T+1]}, k \geq T+1$. In addition, we see that both processes $X_l(t)$ and $X_s(t)$ are independent. Then, by adjusting the initial numbers and service times of customers for both processes, we make both processes stationary. From the definitions of $X_s(t)$ and $X_l(t)$, we see that $X_s(t)$ captures the short time fluctuation in the input LRD traffic and $X_l(t)$ captures the long time fluctuation in the input LRD traffic.

Our next step is to compute the effective bandwidths for $X_s(t)$ and $X_l(t)$ based on the given QoS requirement of the input LRD traffic. Since the service times of customers generating the short time scale process $X_s(t)$ are bounded by T, we can easily show that $X_s(t)$ is a short range dependence process. Hence, we use the large deviation theory [4] to compute the effective bandwidth of $X_s(t)$ based on the QoS requirement of the input LRD traffic. We will give the details in subsection 3.1. For the long time scale process $X_l(t)$, since the tail behavior of the service times of customers generating $X_l(t)$ is the same as the discretized Pareto service times, we see that $X_l(t)$ is a LRD traffic. However, since the

service times of customers generating $X_l(t)$ are relatively long, the process $X_l(t)$ is changing relatively slowly, so that we may think that $X_l(t)$ is almost constant during each time period of fixed length, called the basic allocation period. In addition, since the required QoS is already considered in the computation of the effective bandwidth of $X_s(t)$, we predict the amount of $X_l(t)$ based on the previous history of $X_l(t)$ at the first slot of each basic allocation period and allocate the same amount of bandwidth as the prediction for $X_l(t)$ during the basic allocation period. We will use the minimum mean square error (MMSE) linear predictor to predict the amount of $X_l(t)$, which will be given in detail in subsection 3.2. Finally, we will describe how to allocate the bandwidth for the LRD input traffic adaptively based on the effective bandwidth of $X_s(t)$ and the estimated bandwidth of $X_l(t)$ in subsection 3.3.

3.1 The Effective Bandwidth for $X_s(t)$

In this subsection, we estimate the effective bandwidth for the short time scale process $X_s(t)$. To do this, we assume that the system has a buffer of size b and should guarantee the overflow probability less than $e^{-\xi b}$ for the input LRD process. As mentioned above, since the effective bandwidth function of $X_l(t)$ is predicted by a MMSE linear predictor which can not consider the QoS requirement, we consider the QoS requirement of the input LRD process in the computation of the effective bandwidth of $X_s(t)$, which is given in the following theorem:

Theorem 1. *When b is the buffer size of the system, the effective bandwidth C_s for the required overflow probability $e^{-\xi b}$ is given by*

$$C_s = \sum_{n=1}^{T} \lambda P\left[\sigma = n\right] \frac{e^{\xi n} - 1}{\xi}.$$

Proof: First, note that the arrival process of customers generating $X_s(t)$ can be decomposed into T independent Poisson processes with rate $\lambda_n = \lambda P[\sigma = n]$. For convenience, the Poisson process with rate λ_n is called the λ_n-Poisson process. Let $B_{n,i}$ be the number of customers arriving in $[i-1, i)$ and $A_n(t)$ be the total amount of traffic arrived in $[0, t)$ for the λ_n-Poisson process. Then, observing that each customer of the λ_n-Poisson process eventually generates n packets (since the service time of the customer is always n), for sufficiently large t we get

$$A_n^{(l)}(t) \stackrel{\Delta}{=} \tilde{B}_0 + \sum_{i=1}^{t-n} B_{n,i} \cdot n \le A_n(t) \le \tilde{B}_0 + \sum_{i=1}^{t} B_{n,i} \cdot n \stackrel{\Delta}{=} A_n^{(u)}(t), \qquad (4)$$

where $\tilde{B}_0$ denotes the amount of traffic due to the initial customers of the λ_n-Poisson process. Note that $B_{n,i}$ are i.i.d poisson random variables with parameter λ_n. From the first inequality of (4) we get

$$E[e^{\theta A_n(t)}] \ge E[e^{\theta \sum_{i=1}^{t-n} B_{n,i} \cdot n}] = E[\Pi_{i=1}^{t-n} e^{\theta B_{n,i} \cdot n}] = \Pi_{i=1}^{t-n} E[e^{\theta B_{n,i} \cdot n}].$$

By taking the logarithm we obtain,

$$\log E[e^{\theta A_n(t)}] \geq \sum_{i=1}^{t-n} \log E[e^{\theta \cdot n B_{n,i}}] = (t-n)\lambda_n(e^{\theta \cdot n} - 1)$$

since B_n is the poisson r.v. with parameter λ_n. Then it follows that

$$\lim_{t\to\infty} \frac{1}{t} \log E[e^{\theta A_n(t)}] \geq \lambda_n(e^{\theta \cdot n} - 1).$$

Similarly, from the second inequality of (4) we can show that $\lim_{t\to\infty} \frac{1}{t} \log E[e^{\theta A_n(t)}] \leq \lambda_n(e^{\theta \cdot n} - 1)$. Hence, the Gärtner-Ellis limit [4] of the λ_n-Poisson process is given by

$$\lim_{t\to\infty} \frac{1}{t} \log E[e^{\theta A_n(t)}] = \lambda_n(e^{\theta \cdot n} - 1).$$

Now, since $X_s(t)$ is the superposition of the λ_n-Poisson processes for $1 \leq n \leq T$, the Gärtner-Ellis limit of $X_s(t)$ is given by $\Lambda(\theta) = \sum_{n=1}^{T} \lambda_n(e^{\theta \cdot n} - 1)$.

Therefore, when b is buffer size, the effective bandwidth C_s for $X_s(t)$ with the overflow probability $e^{-\xi b}$ can be calculated as follows [4]:

$$C_s = \frac{\Lambda(\xi)}{\xi} = \sum_{n=1}^{T} \lambda P[\sigma = n] \frac{e^{\xi n} - 1}{\xi},$$

■

3.2 The Effective Bandwidth for $X_l(t)$

In this subsection, we estimate the effective bandwidth for the long time scale process $X_l(t)$. Even though we can't extract $X_l(t)$ from $X(t)$ directly, we can achieve our purpose by introducing a new process $Z(t)$ defined by

$$\begin{aligned} Z(t) &\triangleq X(t) - E[X_s(t)] = X_l(t) + X_s(t) - E[X_s(t)] \\ &= X_l(t) + \eta(t), \end{aligned}$$

where $\eta(t) = X_s(t) - E[X_s(t)]$. The process $\eta(t)$ is viewed as a noise process which is not i.i.d., but $E[\eta(t)] = 0$. Note that $X_l(t)$ and $\eta(t)$ are independent because $X_l(t)$ and $X_s(t)$ are independent. In addition, if we know the threshold value T and the input traffic parameters λ, γ and δ, we can get $Z(t)$ from $X(t)$.

Next, we consider the p^{th} order linear predictor which can estimate $\hat{X}_l(t+1)$ of $X_l(t+1)$ using a linear combination of the current and previous values of $X_l(t)$ as follows:

$$\hat{X}_l(t+1) \triangleq \sum_{i=0}^{p-1} \omega(i) Z(t-i), \quad \sum_i \omega(i) = 1, \tag{5}$$

where the coefficients $\omega(i)$ can be obtained by an induced linear system given below. The optimal linear predictor in the mean square sense is such that minimizes the mean square error $\psi = E\left[\left(X_l(t+1) - \hat{X}_l(t+1)\right)^2\right]$. Note that ψ is

represented by a function of the vector $\boldsymbol{\omega} = (\omega(0), \omega(1), \cdots, \omega(p-1))$. So, the vector $\boldsymbol{\omega}$ that minimizes ψ is found by taking the gradient, setting it equal to zero and then solving it for $\boldsymbol{\omega}$. This results in

$$E[X_l(t+1)Z(t-j)] = E\left[\hat{X}_l(t+1)Z(t-j)\right]. \tag{6}$$

From the facts that $X_l(t)$ and $\eta(t)$ are independent and $E[\eta(t)] = 0$ for all t, the left hand side of (6) is computed as follows:

$$\begin{aligned} E[X_l(t+1)Z(t-j)] &= E[X_l(t+1)(X_l(t-j) + \eta(t-j))] \\ &= r(j+1), \quad j = 0, 1, \cdots, p-1, \end{aligned} \tag{7}$$

where $r(k) = E[X_l(t)X_l(t+k)]$. Similarly, the right hand side of (6) is computed as follows:

$$\begin{aligned} E\left[\hat{X}_l(t+1)Z(t-j)\right] &= E\Big[\sum_{i=0}^{p-1}\omega(i)Z(t-i)Z(t-j)\Big] \\ &= \sum_{i=0}^{p-1}\omega(i)E\big[X_l(t-i)X_l(t-j)\big] + \sum_{i=0}^{p-1}\omega(i)E\big[\eta(t-i)\eta(t-j)\big] \\ &= \sum_{i=0}^{p-1}\omega(i)\big[r(j-i) + s(j-i)\big], \quad 0 \le j \le p-1, \end{aligned} \tag{8}$$

where $s(k) = E[\eta(t)\eta(t+k)]$. Combining (6),(7) and (8) yields

$$r(j+1) = \sum_{i=0}^{p-1}\omega(i)[r(j-i) + s(j-i)], \quad j = 0, 1, \cdots, p-1.$$

In matrix form, this leads to the following linear system, so called Weiner-Hopf linear equation [1,8] whose solution gives the optimum filter coefficients $\omega(0), \cdots, \omega(p-1)$:

$$\begin{bmatrix} r(1) \\ r(2) \\ r(3) \\ \vdots \\ r(p) \end{bmatrix} = \begin{bmatrix} r(0)+s(0) & r(1)+s(1) & \cdots & r(p-1)+s(p-1) \\ r(1)+s(1) & r(0)+s(0) & \cdots & r(p-2)+s(p-2) \\ r(2)+s(2) & r(1)+s(1) & \cdots & r(p-3)+s(p-3) \\ \vdots & \vdots & \ddots & \cdots \\ r(p-1)+s(p-1) & r(p-2)+s(p-2) & \cdots & r(0)+s(0) \end{bmatrix} \begin{bmatrix} \omega(0) \\ \omega(1) \\ \omega(2) \\ \vdots \\ \omega(p-1) \end{bmatrix}. \tag{9}$$

It remains to compute $r(k)$ and $r(k) + s(k)$ to get $\omega(i), 0 \le i \le p-1$, which result in the optimal linear predictor. We need the following theorem:

Theorem 2. *$r(k)$ and $r(k) + s(k)$ are given by*

$$\begin{aligned} r(k) &= \Gamma^*(k) + E^2[X_l(t)], \\ r(k) + s(k) &= \Gamma(k) + E^2[X_l(t)], \end{aligned}$$

where $\Gamma^*(k)$ is autocovariance of lag k for $X_l(t)$ and given by

$$\Gamma^*(k)=\begin{cases}\lambda\left[E\left[\sigma\right]-k+\sum_{t=1}^{k}(k-t)P\left[\sigma=t\right]\right], \textit{if } k\geq T,\\ \lambda\left[P\left[\sigma\geq T+1\right](T-k)+E\left[\sigma\right]-T+\sum_{t=1}^{T}(T-t)P\left[\sigma=t\right]\right], \textit{if } k\leq T-1.\end{cases} \tag{10}$$

and $\Gamma(k)$ is autocovariance of lag k for $X(t)$, given in (2).

Proof: To compute $r(k)$ we only need to compute $\Gamma^*(k)$. Since a long time scale process $X_l(t)$ is an $M/G/\infty$ input process with arrival rate $\lambda P[\sigma \geq T+1]$ and service time σ^* with p.m.f $\frac{P[\sigma=k]}{P[\sigma\geq T+1]}$, $\quad k \geq T+1$, with the help of (2) $\Gamma^*(k)$ can be expressed as follows:

$$\Gamma^*(k) = \lambda P[\sigma \geq T+1]E[\sigma^*]P[\hat{\sigma}^* > k], \tag{11}$$

where $E[\sigma^*]$ is the mean of service times in a long time scale process $X_l(t)$ and $\hat{\sigma}^*$ is the forward recurrence time associated with σ^*. Observe that $P[\hat{\sigma}^* > k]$ can be calculated as follows:

$$\begin{aligned}P[\hat{\sigma}^* > k] &= \sum_{i=k+1}^{\infty} P[\hat{\sigma}^* = i] = \sum_{i=k+1}^{\infty} \frac{P[\sigma^* \geq i]}{E[\sigma^*]}\\ &= \begin{cases}\frac{1}{E[\sigma^*]}\frac{1}{P[\sigma\geq T+1]}\left[E\left[\sigma\right]-k+\sum_{t=1}^{k}(k-t)P\left[\sigma=t\right]\right], \text{if } k\geq T,\\ \frac{1}{E[\sigma^*]}\left\{(T-k)+\frac{1}{P[\sigma\geq T+1]}\left[E\left[\sigma\right]-T+\sum_{t=1}^{T}(T-t)P\left[\sigma=t\right]\right]\right\}, \text{if } k\leq T-1.\end{cases}\end{aligned} \tag{12}$$

Then combining (11) and (12) we show that $\Gamma^*(k)$ is given by (10).

Next, $r(k)+s(k)$ can be computed as follows:

$$\begin{aligned}r(k)+s(k) &= E\left[X_l(t)X_l(t+k)\right] + E\left[\eta(t)\eta(t+k)\right]\\ &= E\left[\{X_l(t)+\eta(t)\}\{X_l(t+k)+\eta(t+k)\}\right]\\ &\quad \text{since } E[\eta(t)]=0 \text{ for all } t, \text{ and } \eta(t) \text{ and } X_l(t) \text{ are independent}\\ &= E\left[\{X(t)-E\left[X_s(t)\right]\}\{X(t+k)-E\left[X_s(t+k)\right]\}\right]\\ &\quad \text{by the definition of } \eta(t)\\ &= E\left[X(t)X(t+k)\right]-2E[X(t)]E[X_s(t)]+E^2\left[X_s(t)\right]\\ &\quad \text{by the staionarity of } X(t) \text{ and } X_s(t)\\ &= \Gamma(k)+E^2\left[X(t)\right]-2E[X(t)]E[X_s(t)]+E^2\left[X_s(t)\right]\\ &= \Gamma(k)+\left(E[X(t)]-E[X_s(t)]\right)^2\\ &= \Gamma(k)+E^2[X_l(t)],\end{aligned}$$

where $\Gamma(k)$ is autocovariance of lag k for $X(t)$, given in (2). ∎

Consequently if we have the parameters λ, γ and δ of the $M/G/\infty$ input traffic, which can be obtained from the previously measured data or experience, the coefficients $\{\omega(i)\}_{i=0}^{p-1}$ can be determined with the help of Theorem 2 and (3.2). Then, by substituting $\omega(i)$ values into equation (5), we can obtain $\hat{X}_l(t)$ as an estimate of the long time scale process $X_l(t)$.

3.3 Adaptive Bandwidth Allocation Strategy

In this subsection, we describe how to allocate the bandwidth for a system with a buffer of size b and the $M/G/\infty$ input traffic requiring overflow probability less than $e^{-\xi b}$. To allocate the bandwidth for the $M/G/\infty$ input process, we first compute the effective bandwidth C_s of $X_s(t)$ by using Theorem 1. Next, we divide the time axis into *basic allocation periods* of equal size p (slots). At the beginning epoch, say time t, of each basic allocation period we compute the effective bandwidth $\hat{X}_l(t)$ of $X_l(t)$ by using Theorem 2 and (5). Then since the $M/G/\infty$ input traffic $X(t)$ satisfies $X(t) = X_s(t) + X_l(t)$, the bandwidth allocated for the $M/G/\infty$ input traffic is $C_s + \hat{X}_l(t)$ during the current basic allocation period. This procedure is performed continuously for each basic allocation period. By doing this we can capture the short time scale fluctuation by a buffer and the long time scale fluctuation by increasing or decreasing the bandwidth. Furthermore, we only need to estimate the required bandwidth for every period of length p, which reduces the implementation complexity.

4 Numerical Studies

To evaluate the performance of our proposed method, we simulate a number of queueing systems with different parameter values. In simulations since the allocated bandwidth for the $M/G/\infty$ input traffic in our ABA method is $C_s + \hat{X}_l(t)$ for each basic allocation period, the evolution equation of the buffer content process $q(t)$ is given as

$$q(t+1) = \min\Big(\max\big(q(t) + X(t) - C_s - \hat{X}_l(t) \ , \ 0\big) \ , \ b\Big)$$

where $\hat{X}_l(t)$ is updated for each basic allocation period.

To check if the proposed ABA method works well, we consider an $M/G/\infty$ input process with parameters $\lambda = 0.4, \gamma = 1.18$ and $\delta = 0.9153$, and the target overflow probability is given by 10^{-3}. We simulate the queueing process as given above and check the overflow probability of the system. The results for $T = p = 70$ with various values of b are given in Table 1. In the table, we give the mean values and confidence intervals for five sample paths. As seen in the table, the resulting overflow probabilities are very close to our target overflow probability 10^{-3}.

Next, we change the values of T and p to see the effect of T and p on performance. We first fix the value of T and change the value of p, and our experiment show that when p is equal to T, our ABA method performs well. We omit the experiment results due to the limitation of space. So, we propose to use the same value for T and p in our ABA method for simplicity. Now to investigate the effect of T we change the value of T from 10 to 120 and the results are given in Table 2. In this experiment, we use $\lambda = 0.4, \gamma = 1.18, \delta = 0.9153$ for the $M/G/\infty$ input process and the buffer size $b = 150$. As seen in the table, our ABA method performs well when the value of $T(= p)$ is neither small nor

Table 1. Overflow Probability (O.P.) with buffer size b : $\lambda = 0.4$, $\gamma = 1.18$, $\delta = 0.9153$

T value	$T = 70$				
p value	$p = 70$				
b value	$b = 110$	$b = 130$	$b = 150$	$b = 170$	$b = 190$
O.P.	$4.039e^{-4}$	$1.144e^{-3}$	$1.161e^{-3}$	$1.134e^{-3}$	$9.985e^{-4}$
Confidence Interval(C.I.)	$(2.429e^{-4}, 5.649e^{-4})$	$(8.999e^{-4}, 1.388e^{-3})$	$(7.896e^{-4}, 1.533e^{-3})$	$(8.369e^{-4}, 1.431e^{-3})$	$(5.765e^{-4}, 1.420e^{-3})$

Table 2. Overflow Probability (O.P.) : $\lambda = 0.4$, $\gamma = 1.18$, $\delta = 0.9153$, $b = 150$

T value	$T = 10$	$T = 30$	$T = 50$	$T = 70$	$T = 100$	$T = 120$
p value	$p = 10$	$p = 30$	$p = 50$	$p = 70$	$p = 100$	$p = 120$
O.P.	0.000	$2.823e^{-5}$	$8.627e^{-4}$	$1.161e^{-3}$	$8.711e^{-4}$	$5.381e^{-5}$
Confidence Interval(C.I.)	$(0.000, 0.000)$	$(1.036e^{-5}, 4.610e^{-5})$	$(6.829e^{-4}, 1.096e^{-3})$	$(7.896e^{-4}, 1.533e^{-3})$	$(6.119e^{-4}, 1.130e^{-3})$	$(1.815e^{-5}, 8.947e^{-5})$

Table 3. Overflow Probability (O.P.) : $\lambda = 0.4$, $\gamma = 1.5$, $\delta = 0.9153$, $b = 150$

T value	$T = 30$	$T = 50$	$T = 70$	$T = 100$	$T = 120$	$T = 140$
p value	$p = 30$	$p = 50$	$p = 70$	$p = 100$	$p = 120$	$p = 140$
O.P.	0.000	$3.755e^{-5}$	$2.834e^{-4}$	$9.939e^{-4}$	$8.002e^{-4}$	$6.167e^{-5}$
Confidence Interval(C.I.)	$(0.000, 0.000)$	$(9.728e^{-6}, 6.537e^{-5})$	$(8.732e^{-5}, 4.794e^{-4})$	$(5.853e^{-4}, 1.403e^{-3})$	$(5.643e^{-4}, 1.036e^{-3})$	$(6.415e^{-6}, 1.169e^{-4})$

Table 4. Overflow Probability (O.P.) : $\lambda = 0.4$, $\gamma = 1.9$, $\delta = 0.9153$, $b = 150$

T value	$T = 70$	$T = 100$	$T = 120$	$T = 140$	$T = 160$	$T = 180$
p value	$p = 70$	$p = 100$	$p = 120$	$p = 140$	$p = 160$	$p = 180$
O.P.	$6.541e^{-6}$	$2.707e^{-4}$	$9.360e^{-4}$	$9.722e^{-4}$	$9.999e^{-4}$	$5.698e^{-4}$
Confidence Interval(C.I.)	$(3.064e^{-6}, 1.002e^{-5})$	$(1.035e^{-4}, 4.379e^{-4})$	$(6.927e^{-4}, 1.179e^{-3})$	$(6.587e^{-4}, 1.286e^{-3})$	$(3.174e^{-4}, 1.682e^{-3})$	$(9.767e^{-5}, 1.042e^{-3})$

large. For instance, when T is in the range $[50, 100]$ in our experiment, our ABA method performs well in this case.

Finally, we investigate the effect of the input traffic parameters on performance. To do this, we use $\lambda = 0.4, \delta = 0.9153, b = 150$, but the scale parameter γ is changed from $\gamma = 1.5$ to $\gamma = 1.9$. Note that the long range dependence largely depends on the scale parameter γ, and as the value of γ is getting close to 1, the autocorrelation of the input traffic is getting stronger. The results are given in Tables 3 and 4. As seen in Tables 3 and 4, our ABA method performs well except for very small or large values of T. In addition, we see that when γ is close to 1, our ABA method performs well even under moderately small values of

T. Noting that the effective bandwidth C_s of $X_s(t)$ in Theorem 1 overestimates the required effective bandwidth of $X_s(t)$, when we use large values of T, the use of moderately small values of T is important to implement our ABA method in practice. Hence, we conclude that our ABA method is suitable for the real traffic which exhibits strong correlation.

5 Conclusions

In this paper, we considered a long range dependence traffic which is modelled by an $M/G/\infty$ input process and proposed a new adaptive bandwidth allocation (ABA) method for the long range dependence traffic. In the proposed method, we divide the input process into two sub-processes, a long time scale process and a short time scale process. For the long time scale process we estimate the required bandwidth using the MMSE linear prediction method. On the other hand, for the short time scale process we estimate the effective bandwidth based on the required QoS of the input process. We verified the effectiveness of our proposed method through simulations. Our simulation studies showed that our ABA method is suitable for the real traffic which exhibits strong correlation.

References

1. A. Adas: Supporting real time VBR video using dynamic reservation based on linear prediction. Proc. of IEEE Infocom. (1996) 1476-1483
2. J. Beran, R. Sherman, M. S. Taqqu, W. Willinger: Long-range dependence in variable bit-rate video traffic. IEEE Trans. Commun. **43** (1995) 1566-1579
3. S. Bodamer, J. Charzinski: Evaluation of effective bandwidth schemes for self-similar traffic. Proc. of the 13th ITC Specialist seminar on IP Measurement, Modeling and Management, Monterey, CA, Sep. (2000) 21-1-21-10
4. Cheng-Shang Chang: Performance Guarantees in Communication Networks. Springer (1999) 291-376
5. M. E. Crovella, A. Bestavros: Self-similarity in World Wide Web traffic: Evidence and possible causes. IEEE/ACM Trans. Networking **5** no. 6 (1997) 835-846 ; Proc. ACM SIGMETRICS '96, Philadelphia, May (1996)
6. H. J. Fowler, W. E. Leland: Local area network traffic characteristics with implications for broadband network congestion management. IEEE J. Select. Areas Commun. **9** (1991) 1139-1149
7. M. W. Garrett, W. Willinger: Analysis, modeling, and generation of self-similar VBR video traffic. Proc. SIGCOMM '94 Conf. Sept. (1994) 269-280
8. R. G. Garroppo, S. Giordano, S. Miduri, M. Pagano, F. Russo: Statistical multiplxing of self-similar VBR videoconferencing traffic. Proc. of IEEE Infocom. (1997) 1756-1760
9. A. Karasaridis, D. Hatzinakos: Bandwidth allocation bounds for α-stable self-similar Internet traffic models. IEEE Signal Processing Workshop on Higher-Order Statistics, Ceasarea, Israel, June (1999)
10. M. M. Krunz, A. M. Makowski: Modeling video traffic using $M/G/\infty$ input processes: A compromise between Markovian and LRD Models. IEEE J. Select. Areas Commun. **16** no. 5 (1998) 733-748

11. W. E. Leland, M. S. Taqqu, W. Willinger, D. V. Wilson: On the self-similar nature of Ethernet traffic (extended version). IEEE/ACM Trans. Networking **2** no. 1 Feb. (1994) 1-15
12. A. M. Makowski, M. Parulekar: Buffer asymptotics for $M/G/\infty$ input processes. In: K. Park, W. Willinger (eds.): Self-similar network traffic and performance evaluation. John Wiley & Sons (2002) 215-248
13. M. Parulekar: Buffer engineering for $M/G/\infty$ input processes. Ph.D. dissertation, Univ. Maryland, College Park, Aug. (1999)
14. M. Parulekar, A. M. Makowski: $M/G/\infty$ input processes: a versatile class of models for traffic network. Proc. of IEEE Infocom. (1997) 1452-1459
15. V. Paxson, S. Floyd: Wide area traffic: The failure of Poisson modeling. IEEE/ACM Trans. Networking **3** no. 3 (1993) 226-244
16. F. R. Perlingeiro, L. L. Lee: An effective bandwidth allocation approach for self-similar traffic in a single ATM connection. Proc. of IEEE Globecom. (1999) 1490-1494
17. C. Stathis, B. Maglaris: Modelling the self-similar behavior of network traffic. Computer Networks **34** (2000) 37-47
18. K. P. Tsoukatos, A. M. Makowski: Heavy traffic limits associated with $M/G/\infty$ input processes. Queueing Systems **34** (2000) 101-130