Effective Bandwidth for a Single Server Queueing System with Fractional Brownian Input

SUNGGON KIM, SEUNG YEOB NAM, DAN KEUN SUNG

ABSTRACT

The traffic patterns of today's IP networks exhibit two important properties: self-similarity and long-range dependence. The fractional Brownian motion is widely used for representing the traffic model with the properties. We consider a single server fluid queueing system with input process of a fractional Brownian motion type. Formulas for effective bandwidth are derived in a single source and multiple sources cases.

Keywords. Effective bandwidth; Fractional Brownian Motion; Fluid Queueing System

1. Introduction

The traffic patterns of today's IP networks have been known to exhibit self-similarity and long-range dependence [1,9,17]. Neither of them can be modeled using conventional Markovian models. Self-similarity or statistical self-similarity implies that fluctuation of traffic rates shows similar patterns throughout several time scales. Long-range dependence means that the correlation decays more slowly than that of conventional traffic models. These phenomena are not mutually independent. Rather, they are indispensably related. For applications of self-similarity and long-range dependence in other fields, readers can refer to [20].

Many attempts to model traffic traces with the above properties have been made [4, 8, 13, 15, 19]. Among them, a fractional Brownian motion(fBm) model is widely used. It has been shown empirically [1, 9, 17] that the holding times of a session, such as ftp, telnet, and http have heavy-tailed distributions. In other words, if G is the distribution function of the holding time, then G is said to be heavy-tailed if $\lim_{x\to\infty}(1-G(x+y))/(1-G(x))=1$, $y\geq 0$. Thus, the input traffic process can be modeled as the aggregation of a large number of on-off sources, where the length of a on-period has a heavy-tailed distribution. Many authors pointed out that long-range dependency in internet traffic can be explained by using the long-range dependency of on-off input models. In [22], it was shown that as the number of on-off sources grows infinitely large, the input process of the on-off model converges to an fBm when suitably normalized. Thus, the fBm model can be used as a traffic model when the aggregation level is sufficiently high. Since an fBm model is uniquely represented by only three parameters, it is a parsimonious one.

Effective bandwidth is defined as the minimum service rate to guarantee a required Quality of Service (QoS). Many studies have focused on Markovian type input traffic models [3,5]. The methods used in the studies can not be applied to obtain the effective bandwidth for a queueing system with an fBm or long-range dependent input process. There are also some other definitions for effective bandwidth [7,21]. Examples of QoS include packet/cell loss probability,

¹Department of Electrical Engineering and Computer Science, KAIST, Daejeon, 305-701, KOREA

mean delay, and delay jitter. Loss probability has been considered traditionally in communication networks. Recent network environment, which is required to support applications such as Internet Telephony, video on demand, and video conferencing, may require specific mean delay and loss probability values. For general input processes, however, it is not easy to obtain the loss probability and the mean delay analytically in a finite buffered queueing system. Thus, the overflow probability and the mean delay in the corresponding infinite buffered queueing system are generally obtained as the approximate values of the loss probability and the mean delay in the finite queueing system.

We, in this paper, consider a single server fluid queueing system with a service rate of C. The input process is modeled as an fBm model, which will be presented in the next section explicitly. Let Q denote the queue length in steady state. For a given buffer threshold b, overflow probability constraint L, and mean delay constraint d, the required QoS is given by

$$\Pr\{Q > b\} \le L \quad \text{and} \quad \mathbb{E}[Q] \le Cd.$$
 (1.1)

Then, the effective bandwidth e(b, L, d) is defined as the minimum value such that

$$C \ge e(b, L, d) \Rightarrow \Pr\{Q > b\} \le L \text{ and } \mathbb{E}[Q] \le Cd.$$
 (1.2)

We propose a scheme to obtain the effective bandwidth using a distribution function F for both single and multiple sources in Sections 3, where F is the distribution function of queue length in a special case. A method for the numerical evaluation of F is given in Section 4.

2. Fractional Brownian Input Process

A continuous-time stochastic process $\{Y(t), t \geq 0\}$ is self-similar with a Hurst parameter H(H > 0) iff Y(at) and $a^H Y(t)$ have identical finite-dimensional distributions for all a > 0, i.e., for all finite set of positive real numbers $\{t_1, t_2, \ldots, t_n\}$,

$${Y(at_1), Y(at_2), \dots, Y(at_n)} \stackrel{d}{=} {a^H Y(t_1), a^H Y(t_2), \dots, a^H Y(t_n)},$$

where $\stackrel{d}{=}$ represents the same in distribution. If $\{Y(t), t \geq 0\}$ has stationary increments, then we can construct a stationary increment process $\{Z_n, n \geq 0\}$ defined as

$$Z_n = Y((n+1)\tau) - Y(n\tau), \quad n = 0, 1, 2, \dots$$

where τ is a positive constant. Then, the autocorrelation function $\gamma(\cdot)$ of $\{Z_n, n \geq 0\}$ is given by

$$\gamma(k) = \frac{1}{2} \{ (k+1)^{2H} - 2k^{2H} + (k-1)^{2H} \}, \quad k = 1, 2, \dots$$
 (2.1)

Note that $\gamma(k)$ does not depend on τ . Suppose that a stochastic process has stationary increments and the autocorrelation function of the incremental process is the same as Eqn. (2.1). Then, it is said to be second-order self-similar with Hurst parameter H. For the case that H is in (1/2, 1], it is long-range dependent in the sense that

$$\sum_{k=0}^{\infty} \gamma(k) = \infty,$$

since $\gamma(k) = O(k^{2H-2})$ as k goes to infinite.

A fractional Brownian motion $\{Y(t), t \geq 0\}$ of Hurst parameter H(0 < H < 1) is a zero mean Gaussian process with covariance function

$$Cov(Y(t_1), Y(t_2)) = \frac{\sigma^2}{2} \left\{ |t_1|^{2H} + |t_2|^{2H} - |t_1 - t_2|^{2H} \right\}.$$
 (2.2)

When $\sigma = 1$, it is called the standard fractional Brownian motion. A fractional Brownian motion is a generalized model of the Brownian motion whose Hurst parameter is 1/2. Mandelbrot and Van Ness [11] showed that fractional Brownian motions are self-similar process with stationary increments.

A traffic model using the fractional Brownian motion (fBm model) was introduced by Norros [13]. Let A(t) be the traffic amount during time [0,t] and let $\{B_H(t), t \geq 0\}$ be a standard fractional Brownian motion with Hurst parameter H. Then, the fBm model is

$$A(t) = mt + \sigma B_H(t). \tag{2.3}$$

Since $\{A(t), t \geq 0\}$ is just a linear transformation of the $\{B_H(t), t \geq 0\}$, the autocorrelation function $\gamma(k)$ of $\{A((n+1)\tau) - A(n\tau), n \geq 0\}$ is given by Eqn. (2.1). Thus, $\{A(t), t \geq 0\}$ is second-order self-similar with Hurst parameter H. If the value of H is in (1/2, 1], then it is also long-range dependent. The fBm model can be generated using the generation algorithms [2, 16], which are based on the fast fourier transform and thus their complexities are $O(n \log(n))$.

3. Effective Bandwidth for Single and Multiple Sources

We consider a single server queueing system in which the service rate is C and the input process is given by the fBm model (2.3). Let $\alpha = C - m$. We assume that $\alpha > 0$ for the stability of the queueing system. Then, from the Reich's formula [18], the queue length in stationary state, $Q_{\alpha,\sigma}$, is given by

$$Q_{\alpha,\sigma} = \sup_{t \ge 0} (A(t) - Ct)$$

=
$$\sup_{t > 0} (\sigma B_H(t) - \alpha t).$$
 (3.1)

Using the above representation, we obtain the following theorem:

THEOREM 3.1. Let $\{B_H(t), t \geq 0\}$ be a standard fractional Brownian motion with Hurst parameter H and let α and σ be positive real numbers. If we define a random variable $Q_{\alpha,\sigma}$ as $Q_{\alpha,\sigma} = \sup_{t>0} (\sigma B_H(t) - \alpha t)$, then it follows that

$$Q_{\alpha,\sigma} \simeq \frac{\sigma^{1/(1-H)}}{\sigma^{H/(1-H)}} Q_{1,1},$$

where \simeq represents the same in distribution.

Now we can obtain a form of the effective bandwidth. Let Q^* denote $Q_{1,1}$ and let p=1/H-1 for simplicity. The above theorem implies that

$$\Pr\{Q_{\alpha,\sigma} > b\} = \Pr\{Q^* > \frac{\alpha^{H/(1-H)}}{\sigma^{1/(1-H)}}b\}.$$
(3.2)

Let $F(\cdot)$ be the distribution function of Q^* and let b_L is the (1-L) quantile of F, i.e., $F(b_L) = 1-L$. Note that Q^* , F and b_L depend on H. From the definitions of F and b_L , it follows that

$$\Pr\{Q^* > \frac{\alpha^{H/(1-H)}}{\sigma^{1/(1-H)}}b\} \le L \Leftrightarrow \frac{\alpha^{H/(1-H)}}{\sigma^{1/(1-H)}}b \ge b_L.$$

By Eqn. (3.2) and the above equation, we obtain that

$$\Pr\{Q_{\alpha,\sigma} > b\} \le L \Leftrightarrow \sigma^{1/H} \left(\frac{b_L}{b}\right)^p \le \alpha.$$

Since $\alpha = C - m$, the minimum service rate e_L satisfying that $\Pr\{Q_{\alpha,\sigma} > b\} \leq L$ is calculated as

$$e_L = m + \sigma^{1/H} \left(\frac{b_L}{b}\right)^p. \tag{3.3}$$

The minimum service rate e_d satisfying that $\mathbb{E}[Q_{\alpha,\sigma}] \leq Cd$ is calculated as follows. Theorem 3.1 implies that

$$\mathbb{E}[Q_{\alpha,\sigma}] \le Cd \iff \mathbb{E}[Q^*] \le \frac{\alpha^{H/(1-H)}}{\sigma^{1/(1-H)}} Cd, \tag{3.4}$$

and it is converted to

$$\mathbb{E}[Q_{\alpha,\sigma}] \leq Cd \Leftrightarrow \sigma^{1/H} \frac{(\mathbb{E}[Q^*]/d)^p}{C^p} \leq \alpha.$$

Substituting $\alpha = C - m$ in the above equation, we can see that e_d is a solution of the following nonlinear equation:

$$x = m + \sigma^{1/H} \frac{\left(\mathbb{E}[Q^*]/d\right)^p}{\tau^p}.$$
 (3.5)

If we let $w = \sigma^{1/H}(\mathbb{E}[Q^*]/d)^p$, then e_d is the x-value of the intersection point of an increasing function

$$y = x$$
,

and a decreasing function

$$y = m + \frac{w}{x^p}.$$

The intersection point exists and it is unique, which implies the existence and uniqueness of e_d .

THEOREM 3.2. Suppose that we have a single server queueing system with input process $A(t) = mt + \sigma B_H(t)$, where $B_H(t)$ is a standard fBm with Hurst parameter H. Let e be the effective bandwidth for the QoS requirement that the overflow probability $\Pr\{Q > b\}$ is less than or equal to L and the mean delay is less than or equal to d. Then,

$$e = \max\{m + \sigma^{1/H} \left(\frac{b_L}{b}\right)^p, e_d\},$$

where e_d is the unique solution of the equation $x=m+\sigma^{1/H}(\mathbb{E}[Q^*]/d)^p/x^p$.

Now, we consider the case that there exist multiple sources with the same Hurst parameter H. Then, the aggregated input process is also an fBm with the parameter H. It is natural to assume that the Hurst parameters of the sources can be different from each other. However, the aggregate input process in this case is not an fBm, but just an asymptotically self-similar process with the Hurst parameter of the maximum value among the Hurst parameters of each source. This makes it difficult to analyze the aggregated input process. This problem remains for further study.

Let J be the number of classes and n_j be the number of sources of class j. The input vector is defined as $\mathbf{n} = (n_1, n_2, \dots, n_J)$. Let $A_{ij}(t)$ denote the input process for the i-th source of the j-th class. We assume that

$$A_{ij}(t) = m_j t + \sigma_j B_H^{ij}(t),$$

where $B_H^{ij}(t)$ is a standard fractional Brownian motion with Hurst parameter H, m_j is the mean input rate of j-th class sources, and σ_j^2 is its variance. The aggregated input process A(t) can be written by

$$A(t) = \sum_{j=1}^{J} \sum_{i=1}^{n_j} A_{ij}(t)$$

= $(\mathbf{n} \cdot \mathbf{m})t + (\mathbf{n} \cdot \sigma^2)^{\frac{1}{2}} B_H(t),$

where $\mathbf{m} = (m_1, m_2, \dots, m_J)$, $\sigma^2 = (\sigma_1^2, \sigma_2^2, \dots, \sigma_J^2)$, and $\mathbf{n} \cdot \mathbf{m}$ is the inner product of \mathbf{n} and \mathbf{m} . The above form of the aggregated input process is the same as Eqn. (2.3), in which $m = \mathbf{n} \cdot \mathbf{m}$ and $\sigma = (\mathbf{n} \cdot \sigma^2)^{1/2}$. Then, we can apply Theorem 3.2. Under the same QoS requirement as Eqn. (1.1), the effective bandwidth $e(\mathbf{n})$ is given by

$$e(\mathbf{n}) = \max\{e_L(\mathbf{n}), e_d(\mathbf{n})\},\tag{3.6}$$

where $e_L(\mathbf{n}) = \mathbf{n} \cdot \mathbf{m} + (\mathbf{n} \cdot \sigma^2)^{1/2H} (b_L/b)^p$, and $e_d(\mathbf{n})$ is the unique solution of the equation

$$x = \mathbf{n} \cdot \mathbf{m} + (\mathbf{n} \cdot \sigma^2)^{1/2H} \frac{(\mathbb{E}[Q^*]/d)^p}{r^p}.$$

To investigate the effect of multiplexing gain, we consider the case that n identical sources given by Eqn. (2.3) are multiplexed. For fair comparison, we let the buffer threshold be nb, where b is the buffer threshold in a single source case. Then, the required QoS is

$$\Pr\{Q > nb\} \le L$$
 and $\mathbb{E}[Q] \le Cd$.

The effective bandwidth for the above QoS can be obtained by Eqn. (3.6).

$$e(n) = \max\{e_L(n), e_d(n)\},$$
 (3.7)

where $e_L(n) = nm + n^{1-1/2H} \sigma^{1/H} (b_L/b)^p$ and $e_d(n)$ is the unique solution of the equation

$$x = nm + n^{1/2H} \sigma^{1/H} \frac{(\mathbb{E}[Q^*]/d)^p}{x^p}.$$
 (3.8)

We define an index of multiplexing gain as follows:

$$I(n) = \frac{ne(1)}{e(n)}.$$

If I(n) > 1, then we say that there is a multiplexing gain. As the index I(n) increases, the multiplexing gain increases. From Eqn. (3.7), it follows that

$$I(n) = \frac{\max\{m + \sigma^{\frac{1}{H}}(b_L/b)^p, e_d\}}{\max\{m + n^{-\frac{1}{2H}}\sigma^{\frac{1}{H}}(b_L/b)^p, e_d(n)/n\}},$$
(3.9)

where $e_d = e_d(1)$. The following theorem says that that there is a multiplexing gain for all positive integer $n \ge 2$ and also gives some properties of I(n).

THEOREM 3.3. The multiplexing gain index I(n), n = 1, 2, ... is strictly-increasing and its upper bound is

$$\lim_{n\to\infty}I(n)=\frac{\max\{m+\sigma^{\frac{1}{H}}(b_L/b)^p,e_d\}}{m}.$$

4. Numerical Evaluation of the Distribution F_H

From Theorem 3.2, it suffices to know the value of b_L and $\mathbb{E}[Q^*]$ in order to find the effective bandwidths for a given QoS requirement (1.1) under the input traffic process of the fBm model (2.3). Since $b_L = F^{-1}(1-L)$ and $\mathbb{E}[Q^*] = \int_0^\infty (1-F(x)) dx$, finding the effective bandwidth can be converted into evaluating the distribution function F(y), which is a real valued function of two variables, H and y. We use $F_H(y)$ instead of F(y) to clarify the dependence on H. To our best knowledge, there is no explicit formula for $F_H(y)$.

The Reich's formula, $Q^* = \sup_{t \geq 0} (B_H(t) - t)$ can be used for evaluating $F_H(y)$. Suppose that we simulate the queue length process with a sufficiently long input trace of length T and we do this n times. Let n_f be the number of times such that $\sup_{0 \leq t \leq T} (B_H(t) - t)$ is less than y. If we assume that such an event corresponds to the event that $\sup_{t \geq 0} (B_H(t) - t)$ is less than y with a high probability, then we may estimate $F_H^c(y)$ by $1 - n_f/n$. However, it is not easy to know how large T is sufficient for a reliable estimation. Moreover, if it is possible, the required value of T is very huge for large values of T. Thus, we need a concept of importance sampling, which enables a fast simulation. Michana [12] proposed an importance sampling method for an input process of fractional Brownian motion.

Consider the following zero-mean Gaussian process

$$M(t) = \int_0^t w(t,s) dB_H(s),$$

where

$$w(t,s) = \begin{cases} c_1 s^{1/2-H} (t-s)^{1/2-H}, & s \in (0,t) \\ 0, & s \notin (0,t) \end{cases}$$

and $c_1 = [H(2H-1)B(3/2-H,H-1/2)]^{-1}$ and $B(\cdot,\cdot)$ is the beta function. Norros et al. [14] proved that the zero-mean Gaussian process $\{M(t), t \geq 0\}$ has independent increments and

$$\mathbb{E}[M^2(t)] = c_2^2 t^{2-2H},$$

where $c_2 = [H(2H-1)(2-2H)B(H-1/2,2-2H)]^{-1/2}$. Thus, it is a martingale, i.e., $\mathbb{E}[M(t)|\mathcal{F}_s] = M(s), t > s$, where \mathcal{F}_s is the set of all possible sample paths up to time s. They also proved that the stochastic exponent of M(t),

$$\mathcal{E}_t(M) = \exp\{M(t) - \frac{1}{2}\mathbb{E}[M^2(t)]\}$$

is also a martingale. Using some properties of $\mathcal{E}_t(M)$, Michna [12] obtained the following theorem.

Theorem 4.1 (Michna) Let $\tau_a = \inf\{t | \sup_{t>0} B_H(t) + at > y\}$ and let $\tau = \tau_{-1}$. Then,

$$\Pr\{\tau < \infty\} = \mathbb{E}[\mathcal{E}_{\tau_a}\{-(1+a)M\}].$$

Suppose that we generate n sample paths of $B_H^{(i)}(t) + at$ and obtain the stopping times $\tau_a^{(i)}$ for each i, i = 1, 2, ..., n. Since $\Pr\{Q^* > y\} = \Pr\{\tau < \infty\}$, the above theorem implies that a naive estimator of $F_H^c(y)$ is given by

$$\hat{F}_{H}^{c}(y) = \frac{1}{n} \sum_{i=1}^{n} \exp\{-(1+a)M(\tau_{a}^{(i)}) - \frac{1}{2}(1+a)^{2}c_{2}^{2}(\tau_{a}^{(i)})^{2-2H}\}. \tag{4.1}$$

In real applications, we generate a sample path $B_H(t)$ in discrete manner. We first generate the incremental process $B_H((t+1)\cdot\delta)-B_H(t\cdot\delta)$, $t=0,1,2,\ldots$ in each unit time interval of length δ , and use it to generate a sample path. Let $F_{H,\delta}$ be the stationary distribution of the queue length in the corresponding queueing system, where the inputs of size $B_H((t+1)\cdot\delta)-B_H(t\cdot\delta)$ occur discretely with interval δ , instead of the continuous process $B_H(t)$ and let $\hat{F}_{H,\delta}^c(y)$ be the estimator of $F_{H,\delta}^c(y)$ by Eqn. (4.1). Suppose that each sample path is generated independently. Then, $\mathcal{E}_{\tau_0^{(i)}}\{-(1+a)M^{(i)}\}, i=1,2,\ldots,n$ are independent and identically distributed. Thus, for sufficiently large n, the estimator $\hat{F}_{H,\delta}^c(y)$ is asymptotically normal, i.e., we can assume that

$$\hat{F}_{H,\delta}^c(y) \sim N(F_{H,\delta}^c(y), \frac{\epsilon^2}{n-1}),$$

where ϵ is the sample standard deviation of $\mathcal{E}_{\tau_a}(-(1+a)M)$. Clearly, $F_{H,\delta}^c(\cdot)$ converges to $F_H^c(\cdot)$ as δ decreases to zero. One can choose various values of a. Question is which value of a is the most efficient. It may be dependent on H.

REFERENCES

- M. E. Crovella and A. Bestavros, Self-similarity in World Wide Web traffic: evidence and possible causes, IEEE/ACM Trans. Networking 5(1997) 835–846.
- (2) C. R. Dietrich and G. N. Newsam, Fast and Exact Simuation of Stationary Gaussian Processes through Circulant Embedding of the Covariance Matrix, SIAM, J. Sci. Comput. 18(1997) 1088-1107.
- (3) A. Elwalid and D. Mitra, Effective Bandwidth of general Markovian traffic sources and admission control of high speed networks, IEEE/ACM Trans. Networking 1(1993) 323-343.
- (4) J. R. Gallardo, D. Makrakis, and L. Orozco-Barbosa, Use of α-stable self-similar stochastic processes for modeling traffic in brodband networks, Performance Evaluation 40(2000) 71– 98
- (5) R. Guérin, H. Ahmadi, and M. Naghshineh, Equivalent Capacity and Its Application to Bandwidth Allocation in High-speed Networks, IEEE J. Select. Areas Commun. 9(1991) 968-981.
- (6) J. R. M. Hosking, Fractional differencing, Biometrica 68(1981) 165-176.
- (7) F. P. Kelly, Notes on Effective Bandwidths, in: F. P. Kelly, S. Zachary and I. Ziedins, (Ed.), Stochastic Networks: Theory and Applications, Royal Statistical Society Lecture Notes Series, vol 4, Oxford University Press, 1996, pp. 141–168.
- (8) S. Kim, J. Y. Lee, and D. K. Sung, A shifted gamma distribution model for Long-range dependent Internet traffic, IEEE Commun. Letters 7(2003) 124–126.
- (9) W. E. Leland, M. S. Taqqu, W. Willinger, and D. V. Wilson, On the self-similar nature of Ethernet traffic, IEEE/ACM Trans. Networking 2(1994) 1-15.
- (10) A. Makowski and M. Parulekar, Buffer Asymptotics for M/G/∞ Input Processes, in: K. Park and W. Willinger, (Ed.), Self-Similar Network Traffic and Performance Evaluation, John Wiley & Sons, 2000, pp. 1–38.

- (11) B. B. Mandelbrot and J. W. van Ness, Fractional Brownian motions, fractional noises and applications, SIAM Review 10(1968) 422-437.
- (12) Z. Michna, On the tail probability and first passage times for fractional Brownian motion, Math. Methods Oper. Res. 49(1999) 335–354.
- (13) I. Norros, A storage model with self-similar input, Queueing Systems 16(1994), 387-396.
- (14) I. Norros, E. Valkeila, J. Virtamo, An elementary approach to a Girsanov formula and other analytical results on fractional Brownian motions, Bernoulli 5(1999) 571–587.
- (15) M. Parulekar and A. Makowski, M/G/∞ Input Process: a versatile class of models for network traffic, Proceedings of IEEE INFOCOM '97, pp. 419-426, 1997.
- (16) V. Paxson, Fast approximate synthesis of fractional Gaussian noise for generating self-similar network traffic, Computer Communication Review 27(1997) 5-18.
- (17) V. Paxson and S. Floyd, Wide-area traffic: the failure of Poisson modelling, IEEE/ACM Trans. Networking 3(1995) 226–244.
- (18) E. Reich, On the integrodifferential equation of Takács (I), Ann. Math. Statist. 29(1958) 563-570.
- (19) R. H. Reidi, M. S. Course, V. J. Ribeiro, and R. G. Baranuik, A multifractal wavelet model with application to network traffic, IEEE Tran. Information Theory 45(1999) 992-1018.
- (20) M. S. Taqqu, A bibliographical guide to self-similar process and long-range dependence, in: E. Eberline and M. S. Taqqu, (Ed.), Dependence in Probability and Statistics, Birkhäuser, 1986, pp. 137–162.
- (21) G. de Veciana, C. Courcoubetis, and J. Wanland, Decoupling bandwidths for networks: A decomposition approach to resource management for networks, Proceedings of IEEE INFOCOM '94, pp. 466–474, 1994.
- (22) W. Willinger, M. S. Taqqu, R. Shermann, and D. V. Wilson, Self-Similarity Through High-Variability: Statistical analysis of Ethernet LAN Traffic at the Source Level, IEEE/ACM Trans. Networking (1997) 71–86.