ELSEVIER

Information Processing Letters 74 (2000) 163–173

Information Processing Letters

www.elsevier.com/locate/ipl

Statistical delay analysis on an ATM switch with self-similar input traffic ☆

Joseph Kee-Yin Ng [a,*], Shibin Song [b,1], Wei Zhao [c,2]

[a] *Department of Computer Science, Hong Kong Baptist University, Kowloon, Hong Kong*
[b] *Department of Mathematics, ZhongShan University, GuangZhou, 510275, China*
[c] *Department of Computer Science, Texas A&M University, College Station, TX 77843-3112, USA*

Received 30 September 1998; received in revised form 20 August 1999
Communicated by F.Y.L. Chin

Keywords: Real-time systems; Performance evaluation; ATM delay analysis; Statistical delay bound

1. Introduction

ATM networks being a connection-oriented technology are providing service for real-time communication. Before two hosts can communicate, a connection has to be established between them. After a connection is established, messages are divided into fixed size packets called cells and are being sent from a source host to a destination host. Real-time communication service requires the underlying network to provide performance guarantees for the on-time delivery of cells. There are two types of performance guarantees: deterministic and statistical guarantees. While a deterministic guarantee provides an absolute bound on the worst case cell delay, a statistical guarantee provides a probabilistic bound on the worst case cell delay. This statistical guarantee is more suitable than the deterministic guarantee for the connection admission control of an ATM network, especially for soft real-time applications.

The delay analysis on ATM networks have been studied by many researchers [1,5,7–9,11,14–16,20–24,26–29,34,37,38]. Some recent works of ours [24, 26–29] have given a framework for the deterministic delay computation. We derive the delay bound by utilizing the inverse function of the arrival and service functions during a server's busy period. The method is proven to be simple and efficient as compared to the other methods proposed in the literature. Many papers have shown that the networks traffic, including the variable bit rate (VBR) video traffic, appears to be statistically self-similar ("fractal") [3,10,13,19]. We found out that if the input traffic can be described by a self-similar fractional Brownian motion (FBM) model, we can determine the statistical delay bound on the worst case cell delay. Hence, in this paper, we use a self-similar stochastic process (FBM) to characterize the arrival of the ATM traffic, and extending from our previous work on the deterministic delay guarantee, we provide methods for determining the statistical delay bound for the cell delay for an ATM switch with different kind of output port schedulers.

☆ The work reported in this paper was supported in part by the RGC Earmarked Research Grant under RGC/97-98/54, and by the FRG under FRG/96-97/II-103.

* Corresponding author. Email: jng@comp.hkbu.edu.hk.

[1] Email: ssong@comp.hkbu.edu.hk. Shibin Song is a visiting research scholar in the Department of Computer Science at Hong Kong Baptist University.

[2] Email: zhao@cs.tamu.edu.

PII: S0020-0190(00)00043-0

As for related work, there are relatively few delay analysis on the self-similar traffic. Norros [32], Mayer and Silvster [23], and Tsybakov [38] only obtained delay approximation for an ATM switch with a FIFO scheduling server in the output port. Our proposed methods can provide the upper delay bounds, and are applicable not only to the FIFO server but to other scheduling servers such as the static priority driven server, the Earliest Deadline First (EDF) server, and the Generalized Processor Sharing (GPS) server.

The rest of the paper is organized as follows: Section 2 describes the connections and the ATM model. Section 3 summarizes the deterministic delay bound for the worst case cell delay from our previous works. Section 4 describes the self-similar traffic model and the envelope process for the arrival traffic. Section 5 presents the statistical delay analysis for the worst case cell delay. In Section 6, we give an algorithm for computing the statistical delay bound for the worst case cell delay and a case study based on an ordinary LAN traffic to show the effectiveness of our statistical delay guarantee based on the FBM model as compare to our delay guarantee based on a two-piece linear maximal arrival function, as well as the actual cell delay determined by the LAN traffic trace. And lastly, we conclude the paper in Section 7.

2. Connections and ATM model

2.1. Connections

As mentioned before, ATM is a connection-oriented packet-switched technology. Before two hosts begin to communicate, a connection has to be set up between them. The term *connection* is then often referred as a stream of messages sent from a source host to destination host. In ATM networks, messages from individual connections are divided into fixed size packets called *cells*.

A real-time connection has a stringent deadline constraint on the delay of its cells. The admission control for a real-time connections is: when a new connection request arrives, the network must determine if all the cells within the connection can meet its deadline constraint and without violating the guarantees already provided to the currently active connections.

2.2. ATM switch model

An ATM switch itself consists of input ports, switching fabric, and output ports. A cell that arrives at an input port of a switch is transported by the switching fabric to an output port where the cell is then transmitted along the physical link associated with the output port.

The scheduling policy at an output port controller of an ATM switch determines the order of the cells (from different connections) being transmitted. Typical scheduling policies adopted by the ATM switches are either FIFO or priority driven. However, there are also other scheduling policies such as Earliest Deadline First (EDF) and Generalized Processor Sharing (GPS) [34].

2.3. Connection-server graph

To simplify the delay analysis, the network components mentioned above are abstracted as servers. Thus, an ATM network can be model as a connection-server graph in which the nodes are the servers.

Servers are classified into two categories: constant servers and variable servers [7,35,36]. A constant server offers a constant delay to each cell that traverse through it. A variable server, on the other hand, offers different delay to each cell. Physical links and switching fabric are constant servers. Furthermore, since the function of an input port is to de-multiplex and impose a constant time delay to each cell, therefore, an input port is also a constant server. The output port, on the other hand, is a multiplexor thus it is considered as a variable server and the delay suffered by a cell in this server should depend on the queue length in the buffer and the scheduling policy.

Note that the constant servers serving a connection only add a fixed amount of delay to its cells and do not change its traffic characteristics. Hence we can subtract the appropriate constant delays encountered by each of the connection from its deadline. We, therefore, eliminate all the constant servers from further consideration and focus only on the variable servers. We can view a connection as being served by a variable server only. Fig. 1 shows how we can construct a connection-server graph from an ATM switch having four real-time connections passing through it.

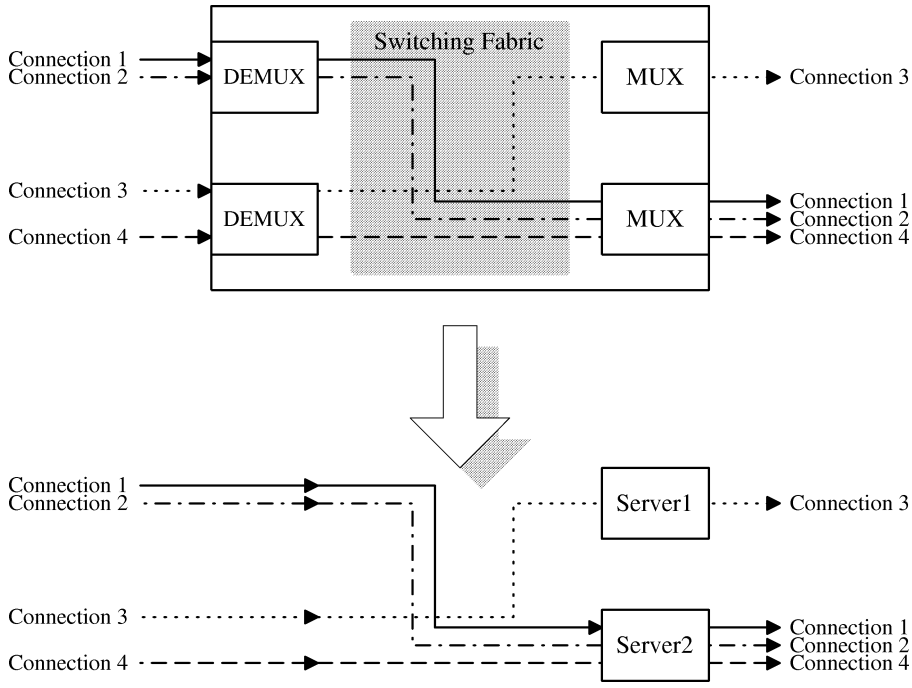

Fig. 1. Connection-server graph.

3. Deterministic performance guarantee

Throughout this paper,we assume a discrete time model [8], where the time slots are numbered starting at $0, 1, 2, \ldots$ and so on. We based on our previous works on the deterministic delay bound and provide our methods for the statistical delay analysis in follow sections. Consider a server system with some connections and assumes the server serves C cells in one time slot and every cell has the same service time. We also assume that the cells of a connection arrive only on the start of each time slot. Let $R_i[t]$ be the number of cells arrived from connection M_i at slot t, then

$$F_i(t, I) = \sum_{j=t+1}^{t+I} R_i[j]$$

is the number of cells arrived during the time interval of $(t, t + I]$. Suppose in any time interval of length I, the maximum number of cells from connection M_i that can arrive at the server is upper-bounded by a constraint function $F_i(I)$, then, for any $t \geqslant 0$

$$F_i(t, I) \leqslant F_i(I). \tag{1}$$

Given this relation, $F_i(I)$ is called the *maximum arrival function* for connection M_i. Because of this, we can select $F_i(I)$ to be an increasing, right-continue and sub-additive function [5,21]. Furthermore, since our delay computation utilized the inverse functions of the arrival and the server functions, we define the inverse function as follows:

Let $G(x)$ be a non-decreasing function. $G^{-1}(y)$, the inverse function of $G(x)$ is given as

$$G^{-1}(y) = \inf\{x \mid G(x) \geqslant y\}. \tag{2}$$

We also define a *busy period* of a connection in a server as the time interval during which any extra cells from the connection arrived at the server must be waiting for the service. Obviously, we are only interested in the busy period. We further state that an instant time t is the *starting point* of a busy period if at $t - 1$, the queue of the server is empty and starting at t, it is not.

Definition 3.1. $S_i(t, I)$ is the service function for connection M_i during the time interval $[t, t + I)$ in a busy period starting at time t. That is, $S_i(t, I)$ specifies

the number of cells from connection M_i that can be served (i.e., transmitted) by a server during the time interval between t and $t + I$.

Let $F_i(I)$ be the maximum arrival function for connection M_i, and $C_i(I)$ be the lower bound of the service function for M_i. $C_i(I)$ is defined as

$$C_i(I) \leqslant S_i(t, I), \tag{3}$$

for any busy period starting at time t with a time interval of length I.

Theorem 3.1. *Given $F_i(I)$ and $C_i(I)$, the upper bound of $F_i(t, I)$ and the lower bound of $S_i(t, I)$, respectively, the worst case delay bound for the cells from connection M_i is*

$$d_i \leqslant \max_{c \geqslant 0} \{C_i^{-1}(c) - F_i^{-1}(c)\}, \tag{4}$$

where $F_i^{-1}(c)$ and $C_i^{-1}(c)$ are the inverse functions of $F_i(I)$ and $C_i(I)$, respectively.

For the proof of the theorem, please refer to the technical report [25].

Definition 3.2. The input of a server is greedy if all arrival functions are always at their maximum from the starting point of the busy period. That means

$$F_j(t, I) = F_j(I), \tag{5}$$

for all j where t is the starting point of the busy period.

Definition 3.3. $S_i(I)$ is the service function for connection M_i when the input of the server is greedy starting at time zero, the beginning of a busy period.

For FIFO scheduling server, Priority driven scheduling server, Leaky bucket server and GPS server [34], ..., etc., the worst case delay occurs when the input of the server is greedy. We have proven these results for a regulated priority driven server and a regulated FIFO server in [24,26,27]. For all the scheduling servers mentioned above, the worst case delay can be computed by the following theorem.

Theorem 3.2. *If the worst case delay for a connection M_i occurs when the input of the server is greedy, then the worst case delay is*

$$d_i = \max_c \big(S_i^{-1}(c) - F_i^{-1}(c)\big), \tag{6}$$

where $F_i^{-1}(c)$ and $S_i^{-1}(c)$ are the inverse functions of $F_i(I)$ and $S_i(I)$, respectively.

For the proof of the theorem, please refer to the technical report [25].

4. The arrival process model and the envelope process

In recent years, many studies [3,4,6,10,13,19,38] have pointed out that network traffic in high-speed networks are best described by traffic models possessing long-range dependency. Consider the network traffic to be a stochastic process. Statistical analysis has shown that the kind of traffic is self-similar with a surprising accuracy. Norros [32] has given a Fractional Brownian Motion (FBM) model to represent the arrival process. FBM plays the same fundamental role among self-similar process that a standard Brownian motion does among processes with independent increments [3,4,6,33]. Erramilli et al. [10] has validated the FBM model by real traffic data. On the other hand, Konstantopoulos [17] has proven that under their assumptions, FBM is the limit of point processes with long-range dependency.

In this paper, we use the FBM model to describe the input traffic as a self-similar input process (see [23,32, 33]). Denote $Z(t)$ a normalized fractional Brownian motion process with a self-similarity parameter (Hurst parameter), H, where $H \in [\frac{1}{2}, 1)$. Following Norros' work [32], the arrival traffic process $A(t)$ is given as

$$A(t) = at + \sigma Z(t), \tag{7}$$

where a is the mean input rate, $\sigma^2 > 0$ is the variance. $A(t)$ represents the number of cells that arrived at time interval $(0, t]$.

Restrict a traffic process $A(t)$ to be a stochastic process of discrete time $t \in \{0, 1, 2, \ldots\}$. $A(t)$ has stationary increments. It can be proved that $A(t)$ is exactly a second-order self-similar process. [3]

Mayer [23] presented an envelope process $\hat{A}(t) = at + k\sigma t^H$ as the bound of $A(t)$ and derived the queue length and delay approximation for FIFO servers. The

[3] Since we used different notations from Norros' work [32], please refer to Appendix A for the details about the equivalency of Eq. (7) and Norros' equation.

delay approximation is statistical but the guarantee probability of the delay were not given in his work. In this paper, we extend our delay bounds derived from the deterministic envelope process in Section 3 to provide a statistical delay bound based on the self-similar FBM input traffic model. Furthermore, our methods are applicable to different kinds of typical servers and at the same time we can provide the guarantee probability.

We define the envelope process for the FBM traffic as follows:

Definition 4.1. $A(t)$ is a FBM arrival process. For a given ε, let $\hat{A}(t) = at + k\sigma t^H$ be the envelope process of the input $A(t)$ with error level ε if

$$P\{A(t) \leqslant \hat{A}(t),\ t = 1, 2, \ldots\} \geqslant 1 - \varepsilon. \tag{8}$$

This definition is different from Mayer's [23] in that the probability in our definition applies to all instances of t but Mayer's definition only specifies a probability for a given time t. By our FBM model, the probability of $\{A(t) \leqslant \hat{A}(t), t = 1, 2, \ldots\}$ is given as

$$\begin{aligned} &P\{A(t) \leqslant \hat{A}(t),\ t = 1, 2, \ldots\} \\ &\quad = P\{at + \sigma Z(t) \leqslant at + k\sigma t^H,\ t = 1, 2, \ldots\} \\ &\quad = P\{Z(t) \leqslant kt^H,\ t = 1, 2, \ldots\}. \end{aligned} \tag{9}$$

In order to determine the envelope process, we only need to find k for a given ε. However, since it is difficult to find k from Eq. (8) or Eq. (9), we provide two methods to find the value of k.

The simulation method

For the statistical guarantee, we only need to consider the busy periods and find a statistical bound for the length of the busy period. As the arrival process $A(t)$ is stationary, we assume the busy period starts at time 0. The length of the busy period B satisfies:

$$\begin{aligned} P\{B > \hat{b}\} &= P\{A(t) \geqslant Ct,\ t = 1, 2, \ldots, \hat{b}\} \\ &\leqslant P\{A(\hat{b}) \geqslant C\hat{b}\} \\ &= P\left\{Z(\hat{b}) \geqslant \frac{(C-a)\hat{b}}{\sigma}\right\} \\ &= P\left\{Z(1) \geqslant \frac{(C-a)}{\sigma \hat{b}^{H-1}}\right\}. \end{aligned} \tag{10}$$

Given a very small p ($\to 0$), such that

$$P\left\{Z(1) \geqslant \frac{(C-a)}{\sigma \hat{b}^{H-1}}\right\} = p$$

then

$$P\{B \leqslant \hat{b}\} = 1 - P\{B > \hat{b}\} \geqslant 1 - p.$$

Thus, we can find a bound $\hat{b}$ for the length of the busy period with a probability of not less than $1 - p$ ($\to 1$).

Since p is very small, we can use $P\{A(t) \leqslant \hat{A}(t),\ t = 1, 2, \ldots, \hat{b}\}$ to approximate $P\{A(t) \leqslant \hat{A}(t), t = 1, 2, \ldots\}$, hence, we have

$$\begin{aligned} &P\{A(t) \leqslant \hat{A}(t),\ t = 1, \ldots, \hat{b}\} \\ &\quad = P\{at + \sigma Z(t) \leqslant at + k\sigma t^H,\ t = 1, 2, \ldots, \hat{b}\} \\ &\quad = P\{Z(t) \leqslant kt^H,\ t = 1, 2, \ldots, \hat{b}\}. \end{aligned} \tag{11}$$

We use a simulation to find the value of k for a given ε. We first generate data according to $Z(t)$ with a specific H. With these data, we can compare them with kt^H where $t = 1, 2, \ldots, \hat{b}$ at different values of k where $k = 1, 2, \ldots$ and construct a table recording different values of k with its corresponding probability of $P\{Z(t) \leqslant kt^H,\ t = 1, 2, \ldots, \hat{b}\}$.

After constructing such a table, we can find the value of k at a given ε or probability $(1 - \varepsilon)$ by a simple table lookup and determine the envelope process $\hat{A}(t)$.

The approximation method

We present here a faster and simpler method for finding the envelope process $\hat{A}(t)$. Due to the fact that

$$\begin{aligned} &P\{A(t) \leqslant \hat{A}(t),\ t = 1, \ldots, \hat{b}\} \\ &\quad = P\left\{\bigcap_{t=1}^{\hat{b}} \{A(t) \leqslant \hat{A}(t)\}\right\} \\ &\quad = 1 - P\left\{\bigcup_{t=1}^{\hat{b}} \{A(t) > \hat{A}(t)\}\right\} \\ &\quad \geqslant 1 - \sum_{t=1}^{\hat{b}} P\{A(t) > \hat{A}(t)\} \\ &\quad = 1 - \sum_{t=1}^{\hat{b}} P\{Z(t) > kt^H\} \end{aligned}$$

$$= 1 - \sum_{t=1}^{\hat{b}} P\{Z(1) > k\}$$
$$= 1 - \hat{b}P\{Z(1) > k\}. \quad (12)$$

For a given ε, we can find k from $P\{Z(1) > k\} = \varepsilon/\hat{b}$, and by Eq. (12)

$$P\{A(t) \leqslant \hat{A}(t),\ t = 1, \ldots, \hat{b}\}$$
$$\geqslant 1 - \hat{b}P\{Z(1) > k\} = 1 - \varepsilon. \quad (13)$$

This method is simpler and faster but the envelope process is looser than that by the simulation method.

5. The statistical delay analysis

There are a few papers [1–4,6,10,17,22,23,32,38] that describe and analyze queuing systems with long range dependent inputs. However, these results only provide delay approximation that are based on the FIFO server. Our proposed methods, can provide the upper delay bounds, and are applicable not only to the FIFO server but to other scheduling servers such as the static priority driven server, the Earliest Deadline First (EDF) server, and the Generalized Processor Sharing (GPS) server.

In this section, we provide the statistical delay analysis for the ATM switch with some typical output port schedulers. Assuming a FBM traffic, the statistical delay bound for the cell delay of a connection can be obtained by each of the followings.

5.1. The FIFO server

For a FIFO server with m connections. Let $A_i(t)$ be a FBM process for the connection M_i, such that

$$A_i(t) = a_i t + \sigma_i Z(t), \quad (14)$$

where $Z(t)$ is a normalized FBM process with a self-similarity parameter H.

For a given ε_i, the envelope process of M_i is

$$\hat{A}_i(t) = a_i t + k_i \sigma_i t^{H_i} \quad (15)$$

such that

$$P\{A_i(t) \leqslant \hat{A}_i(t),\ t = 1, 2, \ldots\} \geqslant 1 - \varepsilon_i. \quad (16)$$

Theorem 5.1. *For a FIFO server with m connections and with the input of M_i being $A_i(t)$ and its envelope process being $\hat{A}_i(t)$, the probability for the worst case cell delay of M_i to exceed $\hat{d}_i$ is less than $1 - \sum_{j=1}^{m} \varepsilon_j$ in which $\hat{d}_i$ is given as*

$$\hat{d}_i = \max_t \left\{ \sum_{j=1}^{m} \hat{A}_j(t) - Ct \right\}, \quad (17)$$

where C is the constant service rate of the server.

Proof. For a FIFO server with m connections, if the envelope process $\hat{A}_i(t)$ is the maximum arrival function, then by Theorem 3.2, the worst case delay of a cell is bounded by $\hat{d}_i$:

$$\hat{d}_i = \max_t \left\{ \sum_{j=1}^{m} \hat{A}_j(t) - Ct \right\}.$$

Since $\hat{d}_i$ is the deterministic delay bound in the server for input $\hat{A}_i(t)$, in any busy period, if $A_j(t) \leqslant \hat{A}_j(t)$, for $j = 1, \ldots, m, \forall t > 0$, then the delay of the cell, d, in the busy period will not exceed $\hat{d}_i$. So we have,

$$\{A_j(t) \leqslant \hat{A}_j(t),\ j = 1, \ldots, m,\ t = 1, 2, \ldots\}$$
$$\subseteq \{d \leqslant \hat{d}_i\}. \quad (18)$$

Denote $Q_j = \{A_j(t) \leqslant \hat{A}_j(t),\ t = 1, 2, \ldots\}$, $j = 1, 2, \ldots, m$, by Eq. (16), we have $P\{\overline{Q_j}\} \leqslant \varepsilon_j$, hence,

$$P\{d \leqslant \hat{d}_i\}$$
$$\geqslant P\{A_j(t) \leqslant \hat{A}_j(t),\ t = 1, 2, \ldots,\ j = 1, \ldots, m\}$$
$$= P\left\{ \bigcap_{j=1}^{m} Q_j \right\} = 1 - P\left\{ \bigcup_{j=1}^{m} \overline{Q_j} \right\}$$
$$\geqslant 1 - \sum_{j=1}^{m} P\{\overline{Q_j}\} \geqslant 1 - \sum_{j=1}^{m} \varepsilon_j. \quad (19)$$

That means, the probability of a cell delay that will not exceed $\hat{d}_i$ is not less than $1 - \sum_{j=1}^{m} \varepsilon_j$. □

5.2. The static priority driven server

For a static priority driven server with m connections and suppose the priorities are assigned in the same order as the connection indices. Let $\hat{A}_i(t) = a_i t + k_i \sigma_i t^{H_i}$ be the envelope process of the input $A_i(t)$ with error level ε_i. Defining $\hat{d}_i$ as

$$\hat{d}_i = \max_c \{S_i^{-1}(c) - \hat{A}_i^{-1}(c)\}, \quad (20)$$

where $\hat{A}_i^{-1}(c)$, $S_i^{-1}(c)$ are the inverse functions of $\hat{A}_i(t)$, $S_i(t)$, respectively. And $S_i(t)$ is given as:

$$S_i(t) = \left\{ Ct - \sum_{j=1}^{i-1} \hat{A}_j(t) \right\}^+. \tag{21}$$

In this paper, for a function ψ, by ψ^+, we mean

$$\psi^+ = \begin{cases} \psi & \psi > 0, \\ 0 & \psi \leqslant 0. \end{cases} \tag{22}$$

We arrive at the following theorem, and the proof of this theorem is similar to that of Theorem 5.1.

Theorem 5.2. *For a static priority driven server with m connections, and with the input of M_i being $A_i(t)$ and its envelope process being $\hat{A}_i(t)$, the probability for the cell delay from M_i to exceed $\hat{d}_i$ is less than $1 - \sum_{j=1}^{i} \varepsilon_j$ where $\hat{d}_i$ is given as in Eq.* (20).

5.3. The EDF priority driven server

For an EDF server with m connections. Let $\hat{A}_i(t) = a_i t + k_i \sigma_i t^{H_i}$ be the envelope process of the input $A_i(t)$ with error level ε_i. Defining $\hat{d}_i$ as

$$\hat{d}_i = \max_c \left\{ S_i^{-1}(c) - \hat{A}_i^{-1}(c) \right\}, \tag{23}$$

where $\hat{A}_i^{-1}(c)$, $S_i^{-1}(c)$ are the inverse functions of $\hat{A}_i(t)$, $S_i(t)$, respectively, and $S_i(t)$, is given as

$$S_i(t) = \left\{ Ct - \sum_{j=1}^{i-1} \hat{A}_j(t - D_k) \right\}^+, \tag{24}$$

where D_k is the deadline for connection M_k. Thus we have the following theorem and the proof of this theorem is similar to that of Theorem 5.1.

Theorem 5.3. *For a EDF server with m connections, and with the input of M_i being $A_i(t)$ and its envelope process being $\hat{A}_i(t)$, the probability for the cell delay of M_i to exceed $\hat{d}_i$ is less than $1 - \sum_{j=1}^{m} \varepsilon_j$.*

5.4. The GPS server

For a GPS server with m connections, a parameter of ϕ_i is being assigned to connection M_i [34]. Let

$\hat{A}_i(t) = a_i t + k_i \sigma_i t^{H_i}$ be the envelope process of the input $A_i(t)$ with error level ε_i. Defining $\hat{d}_i$ as

$$\hat{d}_i = \max_c \left\{ C_i^{-1}(c) - \hat{A}_i^{-1}(c) \right\}, \tag{25}$$

where $\hat{A}_i^{-1}(c)$, $C_i^{-1}(c)$ are the inverse functions of $\hat{A}_i(t)$, $C_i(t)$, respectively and the lower bound of the service function for connection M_i, $C_i(t)$, is given as

$$C_i(t) = \frac{\phi_i}{\sum_{j=1}^{m} \phi_j} t. \tag{26}$$

We have the following theorem, and the proof of this theorem is similar to that of Theorem 5.1.

Theorem 5.4. *For a GPS server with m connections, and with the input of M_i being $A_i(t)$ and its envelope process being $\hat{A}_i(t)$, the probability for the cell delay of M_i to exceed $\hat{d}_i$ is less than $1 - \varepsilon_i$.*

6. An algorithm for computing the statistical delay bound

In Section 5, we obtained the statistical delay bound for each of the scheduling server in an ATM switch. In this section, we present an algorithm for computing these statistical delay bounds.

6.1. Estimation of the self-similar parameter

Under the Fractional Brownian Motion traffic model, $A(t) = at + \sigma Z(t)$, where a is the mean, σ^2 is the variance, H is the self-similar parameter of the input process and $Z(t)$ is a normalized Fractional Brownian Motion process. We present here how to estimate the self-similar process parameters as in [3].

Given the traffic observations $X_1, \ldots, X_n, \ldots$, where X_i is the number of cells arrived between the time interval $(i-1, i]$.

We define

$$\overline{X}_n = \frac{1}{n} \sum_{j=1}^{n} X_j, \tag{27}$$

$$s_n^2 = \frac{1}{n-1} \sum_{j=1}^{n} (X_j - \overline{X}_n)^2. \tag{28}$$

6.1.1. The variance plot method for estimating H

(1) Let k be an integer, $2 \leqslant k \leqslant n/2$. For a sufficient number (say m_k) for a sub-series of length k, calculate the sample means $\overline{X}_1(k), \overline{X}_2(k), \ldots, \overline{X}_{m_k}(k)$ and the overall mean

$$\overline{X}(k) = m_k^{-1} \sum_{j=1}^{m_k} \overline{X}_j(k). \tag{29}$$

(2) For each k, calculate the sample variance of the sample means $\overline{X}_j(k)$ where $j = 1, \ldots, m_k$

$$s^2(k) = (m_k - 1)^{-1} \sum_{k=1}^{m_k} \left(\overline{X}_j(k) - \overline{X}(k)\right)^2. \tag{30}$$

(3) Plot $\log s^2(k)$ against $\log k$. For large values of k, the points in this plot are expected to be scattered around a straight line with a negative slope of $2H - 2$, and therefore the estimator $\widehat{H}$ can be found.

6.1.2. The estimation of mean and variance

The estimators of the mean a and variance σ^2 is given as

$$\hat{a} = \overline{X}_n, \tag{31}$$

$$\hat{\sigma}^2 = \frac{n - n^{2\widehat{H}-1}}{n-1} s_n^2. \tag{32}$$

With these parameters, we can find the value for the envelope process of the input $A(t)$ with error level ε. Thus, we can compute the delay bounds for the connections as given by the methods from Sections 5.1–5.4 depending on the ATM output port scheduling server.

6.2. A case study on regular LAN traffic

Now that we have a way to compute the statistical delay bound for the cell delay in an ATM switch, we present here a case study based on regular LAN traffic for the comparison among our delay bounds, and the actual cell delay experienced at the ATM switch.

In this study, we assume an ATM switch with a FIFO scheduler as its output port controller. One hundred connections are fed through the switch and each of these connections is derived from a real packet trace with a network load of about 1.05 Mbps. This trace was done at the Bellcore Morristown Research and Engineering Facility, on the computing lab's LAN, which carried the local traffic as well as the traffic between Bellcore and the Internet on August 29, 1989. For details about the trace, please refer to [12,18, 19]. [4]

As for our previous delay bound for an ATM switch with a FIFO scheduling server, we assume the arrival traffic can be described by a two-piece linear function with parameters: ρ and β. With the given ρ and β, the worst case delay bound can be determined and can be found in Cruz's [5] and in our previous work [7, 8,24,26–28,30,31]. In choosing the parameter of ρ and β, we adopted Mayor's approach [23] in defining an envelope process such that $\rho t + \beta \geqslant \hat{A}(t),\ \forall t > 0$. With such an envelope process, ρ, and β should satisfy the following inequality, and hence, the delay bound assuming a two-piece linear maximal arrival function can be computed as given in these previous work.

$$(a - \rho)\left[\frac{k\sigma H}{\rho - a}\right]^{1/(1-H)} + k\sigma\left[\frac{k\sigma H}{\rho - a}\right]^{H/(1-H)} - \beta \leqslant 0. \tag{33}$$

As for our approach in finding the statistical delay bound, we can find the value for the envelope process for the input $A(t)$ with error level ε as described in the previous sections. The statistical delay bound can be obtained by the envelope process with a probability of ε that the cell delays will be outside the bound.

In order to show how tight and effective these delay bounds are, we also compute the actual delay for each cell in each busy period while the connections are passing through the ATM switch. We recorded the cell delay in each busy period and sort these cell delays in an ascending order so that we can compare the percentile of these cell delays with the statistical delay bound at the corresponding probability of $(1 - \varepsilon)$.

Fig. 2 shows the delay bound based on a two-piece linear maximal arrival function, the FBM statistical bound and the actual maximum cell delay for our 100-

[4] The traces are also available at http://ita.ee.lbl.gov/html/contrib/BC.html.

[5] Cruz's delay bound is the same as ours under a single connection and with a FIFO server. With schedulers other than the FIFO server and for multiple connections, our methods are found to be more efficient and effective (see [30,31]).

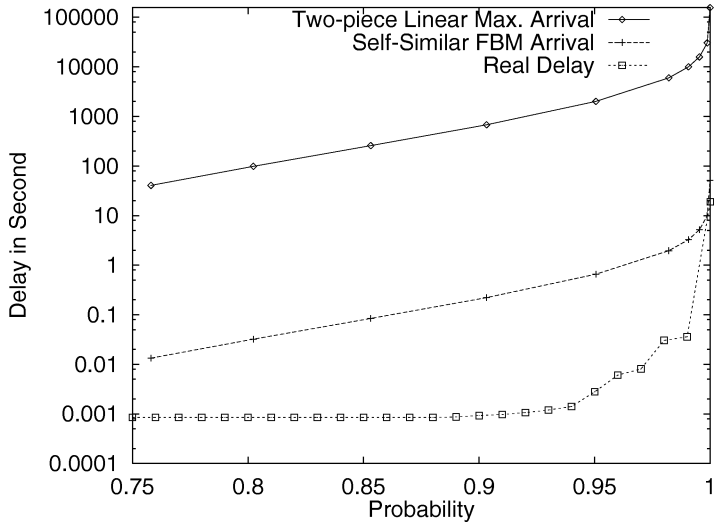

Fig. 2. Comparison among delay bounds and the real delays.

connections system with different probability guarantees. Fig. 2 clearly shows that the statistical delay bound is tighter than the delay bound based on a two-piece linear arrival function and that the statistical delay bound is an accurate statistical bound for the long-range dependent connections with respect to the actual cell delays.

7. Conclusion

As many researchers had proven and validated that many real-time network traffic appear to be statistically self-similar, we use a self-similar stochastic process to characterize the arrival traffic for an ATM switch. As a result, we can determine the statistical delay bound on the worst case cell delay. This statistical delay analysis was based on our previous work on the deterministic delay analysis on an ATM switch. With the statistical delay analysis, we provide methods to determine the statistical delay bound for the network traffic which is more suitable for the connection admission control for an ATM network with different output port controllers. Through our LAN traffic experiments, we show that the statistical bound is tighter than the delay bound derived by our previous method and stays close to the actual maximum cell delay in an ATM network. Not only showing that the statistical bound performs better, our methods are also applicable to different kinds of output port schedulers in the ATM switches other than the FIFO server.

Appendix A. Self-similar traffic equations

A.1. The definition of self-similar traffic model in Eq. (7) *in Section* 4 *is the same as the one given by* (2.2) *of Norros* [32]

In Norros [32], $A(t)$ is given by

$$A(t) = mt + \sqrt{am}Z(t), \quad t \in (-\infty, \infty)$$

and m, a are two different constants for a given $A(t)$, we just take a transform from (m, a) to $(\bar{a}, \sigma)$

$$\begin{cases} \bar{a} = m, \\ \sigma = \sqrt{am} \end{cases}$$

it has inverse transform

$$\begin{cases} m = \bar{a}, \\ a = \sigma^2/m \end{cases}$$

and we get an equivalent model of Eq. (2.2) in Norros [32]

$$A(t) = \bar{a}t + \sigma Z(t), \quad t \in (-\infty, \infty).$$

Replace $\bar{a}$ with a, we can see that this is the same model of Eq. (7) in Section 4.

A.2. About the average rate and the variance

In Norros [32], we have

$$var\big(A(t)\big) = var\big(\sqrt{am}Z(t)\big) = am \times var\big(Z(t)\big) = amt^{2H}$$

and so

$$var\big(A(1)\big) = am.$$

We can see that a is just a "variance coefficient", not a "variance" of $A(1)$, as stated by Norros [32].

A.3. Multiple input traffic case

Actually, in multiple connection case, i.e., there are $m(>1)$ connections, with input traffics $A_i(t)$, our model is more general than the one given by Norros [32].

In Norros [32]'s Proposition 2.2, $A_i(t)$ is defined by

$$A_i(t) = m_i t + \sqrt{am_i}Z_i(t), \quad t \in (-\infty, \infty),$$

where a is the same for different i's, the two coefficient m_i, $\sqrt{am_i}$ are known if we know m_i, they satisfy the relation $y = \sqrt{ax}$. So $A_i(t)$ is determined by m_i, $Z_i(t)$.

In Eq. (10), $A_i(t)$ is defined by

$$A_i(t) = a_i t + \sigma_i Z_i(t), \quad t \in (-\infty, \infty),$$

where the two coefficient a_i, σ_i have no relations, and $A_i(t)$ is determined by $a_i, \sigma_i, Z_i(t)$.

If we add a condition

$$\sigma_i = ca_i$$

we will get the model of Norros [32].

References

[1] V. Anantharam, On the sojourn time of sessions at an ATM buffer with long-range dependent input traffic, in: Proc. 34th Conference on Decision and Control, December 1995.

[2] V. Anantharam, Networks of queues with long-range dependent traffic streams, in: P. Glasserman (Ed.), Stochastic Networks, Stability and Rare Events, Springer, Berlin, 1996.

[3] J. Beran, Statistics for Long-Memory Processes, Chapman & Hall, New York, 1994.

[4] J. Beran, N. Terrin, Estimation of the long-memory parameter, based on a multivariate central limit theorem, J. Time Ser. Anal. 15 (1994) 269–278.

[5] C.S. Chang, Stability queue length, and delay of deterministic and stochastic queueing networks, IEEE Trans. Automat. Contr. 39 (5) (1994) 913–931.

[6] D.R. Cox, Long-range dependence, in: H.A. David, H.T. David (Eds.), An Appraisal, Proceedings 50th Anniversary Conference, Iowa State Statistical Laboratory, Iowa State University Press, 1984.

[7] R.L. Cruz, A calculus for network delay, Parts I & II, IEEE Trans. Inform. Theory 37 (1) (1991) 114–141.

[8] R.L. Cruz, Quality of service guarantees in virtual circuit switched networks, IEEE. J. Select. Areas Commun. 13 (6) (1995) 1048–1056.

[9] A. Ermedahl, H. Hansson, M. Sjodin, Response-time guarantees in ATM networks, in: Proc. 18th Real-Time Systems Symposium (RTSS'97), December 1997, pp. 274–284.

[10] A. Erramilli, O. Narayan, W. Willinger, Experimental queueing analysis with long-range dependence packet traffic, IEEE/ACM Trans. Networking 4 (1996) 209–223.

[11] V. Firoiu, J. Kurose, D. Towsley, Efficient admission control for EDF schedulers, in: Proc. IEEE INFOCOM'97, April 1997.

[12] H. Fowler, W. Leland, Local area network traffic characteristics, with implications for broadband network congestion management, IEEE JSAC 9 (1991) 1139–1149.

[13] M. Garrett, W. Willinger, Analysis modeling and generation of self-similar VBR video traffic, in: Proc. of ACM SIGCOMM, 1994.

[14] H. Hansson, M. Sjodin, K.W. Tindell, Guaranteeing real-time traffic through an ATM network, in: Proc. 30th Hawaii International Conference on System Sciences, Vol. 5, January 1997, pp. 44–53.

[15] E. Knightly, H. Zhang, Traffic characterization and switch utilization using deterministic bounding interval dependent traffic models, in: Proc. IEEE INFOCOM'95, 1995, pp. 1137–1145.

[16] E. Knightly, H-BIND: A new approach to providing statistical performance guarantees to VBR traffic, in: Proc. IEEE INFOCOM'96, 1996.

[17] T. Konstantopoulos, Fractional Brownian approximations of queueing networks, in: P. Glasserman (Ed.), Stochastic Networks, Stability and Rare Events, Springer, Berlin, 1996.

[18] W. Leland, D. Wilson, High time-resolution measurement and analysis of LAN traffic: Implications for LAN interconnection, in: Proc. IEEE INFOCOM'91, 1991, pp. 1360–1366.

[19] W. Leland, M. Taqqu, W. Willinger, D. Wilson, On the self-similar nature of Ethernet traffic (Extended version), IEEE/ACM Trans. Networking 2 (1) (1994) 1–15.

[20] C. Li, W. Zhao, R. Bettati, Delay computation and priority assignment for ATM networks with static priority scheduling, TAMU Tech. Report, Texas A&M University, College Station, TX, 1997.

[21] J. Liebeherr, D. Wrege, D. Ferrari, Exact admission control in networks with bounded delay services, IEEE/ACM Trans. Networking 4 (6) (1996) 885–901.

[22] N. Likhanov, B. Tsybakov, Analysis of an ATM buffer with self-similar ("fractal") input traffic, in: Proc. ACM SIGCOMM, 1995.

[23] G. Mayor, J. Silvster, Time scale analysis of an ATM queueing system with long-range dependent traffic, in: IEEE INFOCOM'97, 1997.

[24] J. Ng, S. Song, W. Zhao, Integrated delay analysis of regulated ATM switch, in: Proc. 18th IEEE Real-Time Systems Symposium (RTSS'97), December 1997, pp. 285–296.

[25] J. Ng, S. Song, W. Zhao, Integrated delay analysis of regulated ATM switch for real-time applications, Technical Report (JNG05-97.ps), Department of Computing Studies, Hong Kong Baptist University, May 1997. Also available at `http://www.comp.hkbu.edu.hk/~jng/Tech-Rpt/`.

[26] J. Ng, S. Song, W. Zhao, Integrated end-to-end delay analysis for regulated ATM networks, Real-Time Systems (1998) submitted.

[27] J. Ng, S. Song, C. Li, W. Zhao, A new method for integrated end-to-end delay analysis in ATM networks, J. Commun. Networks, to appear.

[28] J. Ng, S. Song, ATM delay analysis and efficient connection admission control on real-time traffic, IEICE Trans. Commun. E82-B (6) (June 1999) (IEICE, to appear).

[29] J. Ng, S. Song, C. Li, W. Zhao, Integrated end-to-end delay analysis in ATM networks, Technical Report (JNG07-98.ps), Department of Computer Science, Hong Kong Baptist University, July 1998. Also available at `http://www.comp.hkbu.edu.hk/~jng/Tech-Rpt/`.

[30] J. Ng, S. Song, B. Tang, Efficient delay computation methods for an ATM network with real-time video traffic, in: Proc. 20th IEEE Real-Time Systems Symposium (RTSS'99), December 1999.

[31] J. Ng, B. Tang, Performance comparison between different variations of D-BIND traffic constraint functions for MPEG video streams under an ATM network, Technical Report (JNG03-99.ps), Department of Computer Science, Hong Kong Baptist University, March 1999. Also available at `http://www.comp.hkbu.edu.hk/~jng/Tech-Rpt/`.

[32] I. Norros, A storage model with self-similar input, Queueing Systems 16 (1994) 387–396.

[33] I. Norros, On the use of fractional brownian motion, IEEE JASC 13 (1995) 953–962.

[34] A.K. Parekh, R.G. Gallager, A generalized processor sharing approach to flow control in integrated services networks: The single node case, IEEE/ACM Trans. Networking 1 (3) (1993) 344–357.

[35] A. Raha, S. Kamat, W. Zhao, Guaranteeing end-to-end deadlines in ATM networks, in: Proc. 15th IEEE ICDCS, June 1995.

[36] A. Raha, S. Kamat, W. Zhao, Admission control for hard real-time connections in ATM networks, in: Proc. IEEE INFOCOM'96, March 1996, pp. 111–119.

[37] H. Sariowan, L. Cruz, G. Polyzos, Scheduling for quality of service guarantees via service curves, in: Proc. ICCCN'95, Sept. 1995.

[38] B. Tsybakov, On self-similar traffic in ATM queues: Definitions, overflow probability bound, and cell delay distribution, IEEE/ACM Trans. Networking 5 (3) (1997) 397–408.

[39] D. Wrege, E. Knightly, H. Zhang, J. Liebeherr, Deterministic delay bounds for VBR video in packet-switching networks: fundamental limits and practical trade-offs, IEEE/ACM Trans. Networking 4 (3) (1996).