To be presented at IEEE ICC '95, Seattle, June, 1995

Fast Simulation for Self-Similar Traffic in ATM Networks

Changcheng Huang Michael Devetsikiotis Ioannis Lambadaris A. Roger Kaye
Department of Systems & Computer Engineering
Carleton University
Ottawa, Ontario K1S 5B6, Canada

ABSTRACT Recently *self-similar* (or *fractal*) stochastic processes were proposed as more accurate models of certain categories of traffic (e.g., Ethernet traffic, variable-bit-rate video) which will be transported in ATM networks.

Existing analytical results for the tail distribution of the waiting time in a single server queue based on Fractional Gaussian Noise and large deviation theory, are valid under a steady-state regime and for asymptotically large buffer size. However, predicted performance based on steady-state regimes may be overly pessimistic for practical applications. Theoretical approaches to obtain transient queueing behavior and queueing distributions for small buffer size become quickly intractable.

The approach we followed in this paper was based on fast simulation techniques for the study of certain rare vents such as cell losses with very small probability of occurrence. Our simulation experiments provide insight on transient behavior that is not possible to predict using current analytical results. Finally, they show good agreement with existing results when approaching steady-state.

I. INTRODUCTION

Recent extensive measurements of real traffic data, mainly at Bellcore [1], have led to the conclusion that Ethernet traffic cannot be sufficiently represented by traditional models, but instead can be more accurately matched by *self-similar* (*fractal*) models [2, 3]. More recently, variable-bit-rate (VBR) video traffic was also found to exhibit self-similar characteristics, similarly to LAN traffic [4].

A crucial feature of self-similar processes is that they exhibit *long range dependence* (LRD), that is, their autocorrelation function decays less than exponentially fast. This is in contrast to traditional stochastic models, all of which exhibit *short range dependence* (SRD), i.e., have an autocorrelation function that decays exponentially or faster. The serious implication for ATM network design is that, conclusions based on traditional models may not be applicable under the self-similar traffic scenario.

There have been, in general, only a few analytical results reported in this area, with the notable exception of [5] and [6], where asymptotic expressions for the steady-state waiting time in single-server queues were derived by generalizing large deviation theorems to include self-similar processes. Analytical work related to this subject can also be found in [7].

Results in [5, 6] deal with the steady-state asymptotics for a single-server queue under Fractional Gaussian Noise (FGN). While the self-similar property captures the burstiness of traffic at all time scales, realistic ATM networks are expected to have a limiting time scale. Therefore, predicted performance based on a steady-state regime may be overly pessimistic for practical applications. Furthermore, analytical approaches become quickly intractable.

Given the difficulties in analysis, simulation can play an important role in the study of network performance under self-similar traffic. While several approaches have been proposed for the synthetic generation of self-similar traffic traces (e.g., Hosking's method [8], Mandelbrot's *fast fractional Gaussian noise* approach [9], *nonlinear chaotic maps* [10]), they are, in general, efficient for generating only small numbers of relative long traces. Due to the long term dependent structure of self-similar traffic, accurate statistics can be obtained only from a large number of replications. This is especially true in ATM networks where one may want to simulate events that are *rare*, e.g., cell losses with probability $< 10^{-9}$. For this task, conventional simulation techniques can be extremely inefficient.

In this paper we present a fast simulation approach based on *importance sampling* (IS) and Hosking's method in [8]. Using this approach we simulate the transient queueing behavior of certain self-similar arrival processes, namely discrete-time FGN. We show that our transient results asymptotically approach the steady-state results in [5]. We verify experimentally the existence of a certain time scale at which the transient result is a good approximation for steady-state. Furthermore, we apply our approach to the simulation of the multiplexing effect under both homogeneous and heterogeneous traffic sources.

We focus on the following key issues in ATM network design: the *buffering gain*, i.e., the reduction in cell loss probability as the buffer size increases, and the *multiplexing gain*, i.e., the reduction in cell loss due to statistical smoothing when multiple bursty sources are aggregated. If we define the burstiness of self-similar traffic as the Hurst parameter [11], our results indicate that, the higher the burstiness, the lower the buffering gain, as predicted by large deviation results. Our results also agree with the predictions that, compared with SRD models, self-similar models show smaller buffering gains. On the other hand, perhaps contrary to common belief, our results indicate significant gains from multiplexing. These multiplexing gains increase with the burstiness (Hurst parameter) of the self-similar traffic.

In addition to these results, we show that multiplexing two heterogeneous self-similar sources, the steady-state behavior will be dominated by the burstier one, as predicted by large deviation theory. Therefore, when a process possesses both long range and short range dependence structures, e.g., the *fractional autoregressive integrated moving-average* (F-ARIMA) model, the steady-state will only reflect the contribution of long range dependence. This again emphasizes the need for transient in addition to steady-state analysis.

This paper is organized as follows: In Section II we present a

brief introduction to self-similar traffic models and the existing large deviations results. In Section III we describe the self-similar traffic model we use, namely discrete-time FGN, and Hosking's method for generating traces from it. In Section IV we develop an importance sampling technique for self-similar models. In Section V we present simulation results. Finally, in Section VI we summarize our conclusions and implications of this study for the design of ATM networks. In the Appendix, we generalize results in [5] to include multiplexing effects (required in order to compare with our simulation study).

II. SELF-SIMILAR TRAFFIC MODELS

A. DEFINITION OF SELF-SIMILARITY

Let $\mathbf{X} = \{X_k : k = 1, 2, \ldots\}$ be a *covariance stationary* stochastic process, that is, a process with constant mean $m = E[X_k]$, finite variance $\sigma^2 = E[(X_k - m)^2]$, and an autocorrelation function as follows:

$$r(k) \sim k^{-\beta} L(k), \text{ as } k \to \infty, \tag{1}$$

where $0 < \beta < 1$, and $L(k)$ is slowly varying at infinity, i.e., $\lim_{t\to\infty} L(tx)/L(t) = 1$, for every $x > 0$ [1]. For each $n = 1, 2, 3, \ldots$, let

$$X_k^{(n)} = 1/n(X_{kn} + X_{kn-1} + \cdots + X_{kn-(n-1)}),\ k = 1, 2, 3, \ldots \tag{2}$$

then the time series $\mathbf{X}^{(n)} = \{X_k^{(n)} : k = 1, 2, 3, \ldots\}$ is also a covariance stationary process. Let $r^{(n)}(k)$, $k = 1, 2, \ldots$, denote the corresponding autocorrelation function. If

$$r^{(n)}(k) = r(k), \text{ for all } n = 1, 2, 3, \ldots \text{ and } k = 1, 2, 3, \ldots \tag{3}$$

then the process $\mathbf{X}$ is called *exactly second-order self-similar* with Hurst parameter $H = 1 - \beta/2$. The process $\mathbf{X}$ is called *asymptotically second-order self-similar* with Hurst parameter $H = 1 - \beta/2$, if

$$r^{(n)}(1) \to 2^{1-\beta} - 1, \text{ as } n \to \infty, \tag{4}$$

$$r^{(n)}(k) \to 1/2\delta^2(k^{2-\beta}), \text{ as } n \to \infty\ (k = 2, 3, \ldots), \tag{5}$$

where $\delta^2(f(k)) = f(k+1) - 2f(k) + f(k-1)$.

Definitions of self-similar processes in a more general sense can be found in [2]. Intuitively, one of the most striking features of such processes is that their aggregated processes $\mathbf{X}^{(n)}$ possess a nondegenerate correlation structure as $n \to \infty$. An important recent development in traffic modeling is that Leland *et al.* [1] have found that Ethernet traffic satisfies (3), and Beran *et al.* [4] have shown that VBR video traffic also satisfies (3).

B. DEFINITION OF THE FGN PROCESS

While there are numerous stochastic models which exhibit the self-similar property, two of them, namely the exactly self-similar *fractional Gaussian noise* (FGN) and the asymptotically self-similar *fractional autoregressive integrated moving-average* (F-ARIMA) process, are the most commonly used. FGN can be viewed as a reasonable first approximation of more complex LRD processes, since it can be derived from a special type of central limit theorem applied to LRD processes. While we consider only FGN models in this paper, our approach can be easily extended to include F-ARIMA models. The advantage of F-ARIMA models is that they can model both long time dependence and short time dependence at the same time [12].

A fractional Gaussian noise process $\mathbf{X} = \{X_k : k = 1, 2, \ldots\}$ is a stationary Gaussian process with mean $m = E[X_k]$, variance $\sigma^2 = E[(X_k - m)^2]$, and autocorrelation function

$$r(k) = 1/2(|k+1|^{2H} - 2|k|^{2H} + |k-1|^{2H}),\ k = 1, 2, 3, \ldots \tag{6}$$

Therefore, if $1/2 < H < 1$, FGN is exactly second-order self-similar with Hurst parameter H.

C. LINDLEY EQUATION AND LARGE DEVIATIONS

Now consider a slotted-time single server queue with deterministic service rate μ and a FGN arrival process $\mathbf{X}$, with X_k representing the number of arriving cells within the kth time slot. Here, we assume X_k can take any real value. Let Q_k denote the size of the queue at time $k = 0, 1, \ldots$. Assuming $Q_0 = 0$, we have the following Lindley equation [13]:

$$Q_k = \langle Q_{k-1} + X_k - \mu \rangle^+ = \langle Q_{k-1} + Y_k \rangle^+, \text{ for } k = 1, 2, \ldots \tag{7}$$

where we define the process $\mathbf{Y} = \{Y_k : Y_k = X_k - \mu, k = 1, \ldots\}$ as *work load* process. Now define the *total work load process* $\mathbf{W}$ as $\{W_k : W_k = \sum_{i=1}^{k} Y_i,\ k = 1, 2, \ldots\}$. Then $\mathbf{W}$ is an stationary increment Gaussian process with mean $mk - \mu k$ and variance $\sigma^2 k^{2H}$. Therefore, since $\mathbf{X}$ is a stationary process, we have

$$\Pr(Q_k > b) = \Pr(\sup_{0 \le i \le k} W_i > b),\ for\ k = 0, 1, 2, \ldots \tag{8}$$

Duffield *et al.* [5] have shown the following steady-state, large deviation result:

$$\lim_{b\to\infty} b^{2(1-H)} \log \Pr(Q_\infty > b) = -c^{-2(1-H)}(c+\mu)^2/2 \tag{9}$$

where $c = \mu/H - \mu$ and $\mu > 0$. Therefore, in contrast to traditional SRD models, the steady-state queueing distribution decays asymptotically in a Weibull fashion rather than exponentially. Thus the performance predicted under FGN may be far worse than under traditional models.

Traditional models attempt to capture the burstiness of the traffic at different time scales by using complex hierarchical structures (such as Markov Modulated Arrival processes). In contrast, self-similar models capture long range dependence in a *parsimonious manner* which makes them extremely attractive from the standpoint of modeling realistic LRD traffic [1].

Results in [5] deal with the steady-state asymptotics for a single-server queue under FGN. While the self-similar property captures the burstiness of traffic at all time scales, realistic ATM networks are expected to have a limiting time scale. Therefore, predicted performance based on a steady-state regime may be inaccurate for practical applications. Furthermore, questions regarding the transient behavior, small buffer sizes, multiplexing effects, and, in general, the performance of ATM networks under LRD traffic, remain unanswered. In the following, we develop a simulation approach that can be used to answer the above questions.

III. GENERATION OF FGN TRACES

We briefly describe Hosking's generation procedure [8] in the following paragraphs.

For a FGN process $\mathbf{X}$ with $m = 0$, the conditional mean and variance of X_k, given the past values $x_{k-1}, x_{k-2}, \ldots, x_1$, may be written as [14]

$$m_k = \mathrm{E}(X_k | x_{k-1}, x_{k-2}, \ldots, x_1) = \sum_{j=2}^{k} \phi_{kj} x_{k-j} \tag{10}$$

$$v_k = \mathrm{Var}(X_k | x_{k-1}, x_{k-2}, \ldots, x_1) = \sigma^2 \prod_{j=2}^{k} (1 - \phi_{jj}^2) \tag{11}$$

Here ϕ_{jj} is the jth partial correlation coefficient of $\{X_k\}$ and the ϕ_{kj} are partial linear regression coefficients. For simulating

a sample $\{x_1, x_1, \ldots, x_{n-1}\}$ of size n from a FGN process, [8] describes the following algorithm:

1. Generate a starting value x_1 from a Gaussian distribution $N(0, v_1)$. Set $N_1 = 0$, $D_1 = 1$.

2. For $k = 2, \ldots, n-1$, calculate ϕ_{kj}, $j = 2, \ldots, k$, recursively via the equations

$$N_k = r(k) - \sum_{j=2}^{k-1} \phi_{k-1,j} r(k) \quad (12)$$

$$D_k = D_{k-1} - N_{k-1}^2 / D_{k-1} \quad (13)$$

$$\phi_{kk} = N_k / D_k \quad (14)$$

$$\phi_{kj} = \phi_{k-1,j} - \phi_{kk}\phi_{k-1,k-j} \quad j = 2, \ldots, k-1 \quad (15)$$

Calculate $m_k = \sum_{j=2}^{k} \phi_{kj} x_{k-j}$ and $v_k = (1-\phi_{kk}^2) v_{k-1}$. Generate x_k from the Gaussian distribution $N(m_k, v_k)$.

The above method is applicable to any Gaussian process as long as the correlation function $r(k)$ is known. However, the computational effort required increases approximately as $O(n^2)$ with the length of the trace, n.

Given the computational cost of trace generation, the number of replications required becomes crucial, especially when analyzing ATM networks where one may want to simulate events that are *rare*, e.g., cell losses with probability $< 10^{-9}$, or extremely long cell waiting times. In the following, we develop a fast simulation approach based on importance sampling, that makes Hosking's method applicable to quality-of-service (QoS) evaluation in ATM networks.

IV. IMPORTANCE SAMPLING FOR THE FGN PROCESS

A. IMPORTANCE SAMPLING THEORY

Let U be a random variable that has a probability density function $p(u)$ and consider estimating the probability P that U is in some set A, then $P = \int_{-\infty}^{\infty} I_A(t) p(t) dt = \mathrm{E}_p[I_A(U)]$, where $I_A(\cdot)$ is the indicator function of event A. Assume that $p'(u)$ is another density function. Assuming that $p(u) = 0$ whenever $p'(u) = 0$ (*absolute continuity* condition), we have $P = \int_{-\infty}^{\infty} I_A(t) \frac{p(t)}{p'(t)} p'(t) dx = \mathrm{E}_{p'}[I_A(U) \frac{p(U)}{p'(U)}] = \mathrm{E}_{p'}[I_A(U) L(U)]$ where $L(u) = p(u)/p'(u)$ is a *likelihood ratio* (*weight function*) and the notation p' denotes sampling from the density $p'(u)$. This equation suggests the following variance reduction estimation scheme which is called *importance sampling* (IS) (see [15] and references within): Draw N samples $u_1, \ldots, u_N$ using the density p'. Then, an unbiased estimate of P is given by $\hat{P}_N = \frac{1}{N} \sum_{n=1}^{N} I_A(u_n) L(u_n)$, i.e., P can be estimated by simulating a random variable with a different density and then unbiasing the output $I_A(u_n)$ by multiplying with the likelihood ratio. We call $p'(u)$ the *twisted density*. Since any density can be used as the twisted density, the question arising is which is the *optimal* twisted density, i.e., which is the density that minimizes the variance of $\hat{P}$. A variety of approaches, namely analytical, large deviation-based, and statistical have been proposed in order to choose $p'(u)$ ([15, 16, 17] and references within).

B. TWISTED DENSITY AND LIKELIHOOD RATIOS

Without loss of generality, we assume in the following that we want to simulate a queueing process with a FGN arrival process $\mathbf{X}$ as defined in Section II.B, with mean value $m = 0$. Define a new process $\mathbf{Y}' = \{Y'(k) : Y'(k) = X(k) + m^*, k = 1, \ldots\}$. It is easy to see that process $\mathbf{Y}'$, the *twisted work load process*, is a FGN process with mean m^*, and that its variance and correlation function are the same as for $\mathbf{X}$. Given a realization $(y'_1, \ldots, y'_{k-1})$ of process $\mathbf{Y}'$, the corresponding realization of process $\mathbf{X}$ satisfies $x_j = y'_j - m^*$, for $j = 1, 2, \ldots, k-1$. From equations (10)–(11),

$$\begin{aligned} \mathrm{E}_{Y'}(Y'_k | y'_{k-1}, \ldots, y'_1) &= m^* + \mathrm{E}_X(X_k | y'_{k-1} - m^*, \ldots, y'_1 - m^*) \\ &= m^* + \mathrm{E}_X(X_k | x_{k-1}, \ldots, x_1) = m^* + \sum_{j=2}^{k} \phi_{kj} x_{k-j} \\ &= m^* + \sum_{j=2}^{k} \phi_{kj}(y'_{k-j} - m^*) = m^* + m_{k,Y'} \end{aligned} \quad (16)$$

for $k = 2, 3, \ldots$, where $m_{k,Y'} \triangleq \sum_{j=2}^{k} \phi_{kj}(y'_{k-j} - m^*)$. Also from equations (10)–(11)

$$\mathrm{Var}_{Y'}(Y'_k | y'_{k-1}, \ldots, y'_1) = \mathrm{Var}_X(X_k | x_{k-1}, \ldots, x_1) \quad (17)$$

In IS simulation, we simulate a twisted work load process $\mathbf{Y}'$ instead of the work load process $\mathbf{Y}$. In order to calculate the required likelihood ratio, we let $(y'_1, \ldots, y'_{k-1})$ be also taken as a realization of the work load process $\mathbf{Y}$, as defined in Section II.C. Then, for $k = 2, 3, \ldots$, we have

$$\mathrm{E}_Y(Y_k | y'_{k-1}, \ldots, y'_1) = -\mu + \sum_{j=2}^{k} \phi_{kj}(y'_{k-j} + \mu) = -\mu + m_{k,Y} \quad (18)$$

where $m_{k,Y} \triangleq \sum_{j=2}^{k} \phi_{kj}(y'_{k-j} + \mu)$. We also have

$$\mathrm{Var}_Y(Y_k | y'_{k-1}, \ldots, y'_1) = \mathrm{Var}_{Y'}(Y'_k | y'_{k-1}, \ldots, y'_1) \quad (19)$$

The likelihood ratio up to time k is

$$\begin{aligned} L(k) &= \frac{f_Y(y'_1, \ldots, y'_k)}{f_{Y'}(y'_1, \ldots, y'_k)} \\ &= \frac{f_Y(y'_1) f_Y(y'_2|y'_1) \cdots f_Y(y'_k | y'_{k-1}, \ldots, y'_1)}{f_{Y'}(y'_1) f_{Y'}(y'_2|y'_1) \cdots f_{Y'}(y'_k | y'_{k-1}, \ldots, y'_1)} = \prod_{i=1}^{k} L_i \end{aligned} \quad (20)$$

where, for $i = 2, 3, \ldots, k$,

$$L_i = \frac{f_Y(y'_i | y'_{i-1}, \ldots, y'_1)}{f_{Y'}(y'_i | y'_{i-1}, \ldots, y'_1)}, \; L_1 = \frac{f_Y(y'_1)}{f_{Y'}(y'_1)} \quad (21)$$

Then, from equations (16) to (18), we have

$$L_i = \frac{e^{\theta_i y_i}}{M_i} \text{ for } i = 2, 3, \ldots, \; L_1 = e^{-\frac{2(m^*+\mu)x_0 + (m^*+\mu)^2}{2\sigma^2}} \quad (22)$$

where

$$\theta_i = -\frac{\mu - m_{i,Y} + m^* + m_{i,Y'}}{\sigma^2 \prod_{j=2}^{i}(1 - \phi_{jj}^2)} \quad (23)$$

and $M_i = e^{-\theta_i / 2(\mu - m_{i,Y} - m^* - m_{i,Y'})}$.

The probability $\Pr(Q_k > b)$ can be estimated by observing N iid replications of the realization $w_1^{(n)}, \ldots, w_k^{(n)}$ of $\mathbf{W}$, for $n = 1, \ldots, N$. Let $L^{(n)}$, $n = 1, \ldots, N$, denote the corresponding likelihood ratio for each replication. Then, we propose the following simulation procedure for estimating $\Pr(Q_k > b)$:

1. Initialize $i = 1, n = 1$;

2. Generate a sample point x_i by Hosking's method described in Section III;

3. Generate a sample point y'_i by the equation $y'_i = x_i + m^*$;

4. Generate a sample point w_i by replacing the process $\mathbf{Y}$ with the process $\mathbf{Y}'$ in the definition of total work load process;

5. If $w_i \leq b$ and $i < k$, then repeat from step 2 with $i = i+1$; otherwise continue with step 6;

6. If $w_i \leq b$ and $i = k$, set $I_n = 0$ and go to step 8; otherwise continue with step 7;

7. Set $I_n = 1$ and calculate $L^{(n)} = L(i)$ via equations (20) to (22);

8. If $n = N$ evaluate the estimate using $\hat{P} = \frac{1}{N}\sum_{n=1}^{N} I_n L^{(n)}$; otherwise set $n = n+1$, $i = 1$ and goto step 2.

C. OPTIMAL TWISTED MEAN VALUE

Based on the above description, we can apply IS by suitably modifying (twisting) the mean of the arrival process. However, an efficient method to obtain a favorable (or near-optimal) twisted mean remains to be devised. In this paper, we describe two such methods, namely a heuristic search and an approximate analytical approach. The heuristic search approach has been successfully applied to traditional (SRD) models (see [16] and references within), and will be briefly explained in Section V.

We now focus our attention on the approximate analytical approach. From equation (8), we have $\Pr(Q_k > b) > \max_{0\leq i\leq k} \Pr(W_i > b) \triangleq \mathrm{P}_{W,k}$. This approximation, which is an optimistic bound for $\Pr(Q_k > b)$, becomes quite accurate for any time k, when b is large. Furthermore, as time k grows large, it can be shown that there exists a value $k = k_s$ such that $\mathrm{P}_{W,\infty} \triangleq \max_{i\geq 0} \Pr(W_i > b) \simeq \Pr(W_{k_s} > b)$, where $k_s = \lceil b/c \rceil$, and c is defined in equation (9) [5]. Therefore, for $k > k_s$, $\mathrm{P}_{W,k} \simeq \Pr(W_{k_s} > b)$. Thus, loosely speaking, k_s is the time when the queueing state enters steady-state, and $\Pr(Q_\infty > b) \simeq \Pr(W_{k_s} > b)$. A very accurate approximate formula for calculating $\Pr(W_{k_s} > b)$ (i.e., the tail of a Gaussian distribution) was recommended in [18]. The above approximation procedures lead to quite accurate results, as our results in Section V indicate.

Since $\Pr(Q_\infty > b) \simeq \Pr(W_{k_s} > b)$, our approximate analytical approach consists of finding a near-optimal mean twisted value for $\Pr(W_{k_s} > b)$ and then applying that same twisted value to the simulation of $\Pr(Q_\infty > b)$. Since W_{k_s} is normally distributed with mean $-\mu k_s$ and variance $\sigma^2 k_s^{2H}$, a near-optimal twisted mean value can be readily obtained by minimizing the likelihood ratio $L(k)$ as suggested in [17]. Following this procedure we find a twisted mean value $m^*_{W,opt} \simeq c k_s$. Hence, a near-optimal twisted mean value for process $\mathbf{Y}$ can be found by $m^*_{opt} = m^*_{W,opt}/k_s \simeq c = \mu/H - \mu$. Furthermore, it is reasonable to assume that m^*_{opt} is also near-optimal for the estimation of the (transient) probability $\Pr(Q_k > b)$ when $k > k_s$.

V. NUMERICAL RESULTS

For IS simulation, the estimator $\hat{P}$ of the unknown probability $\Pr(Q_k > b)$ is a function of $(m, m^*, \mu, H, k, b, N, \sigma^2)$. Since our set-up is translation-invariant with respect to m, we assume $m = 0$ without loss of generality. We let σ be fixed at $\sigma = 1$, since as shown in the Appendix, by changing the number of multiplexed homogeneous sources L, we can observe the same effect as if scaling σ.

We divide our simulation experiments into two cases, one with $H = 0.7$, which represents less bursty traffic, and one with $H = 0.9$ representing more bursty traffic. In each case, we consequently discuss dependence on the twisted mean value m^*, on service rate μ, on stopping time k, on the buffer size b, and on the number L of multiplexed homogeneous sources. By homogeneous sources we mean sources which have the same Hurst parameter, H. In the final part, we simulate multiplexing two heterogeneous sources, one with $H = 0.7$ and one with $H = 0.9$. We also provide example values of the improvement factor of our IS technique over conventional MC simulation.

A. CASE I: $H = 0.7$

All simulations are based on 1000 iid replications, except in Fig. 2.

1. The dependence on m^*:

It is important to point out that the IS estimator of $\Pr(Q_k > b)$ is always *unbiased*, regardless of the value of m^*. However, the sample path properties as well as the variance of the IS estimator are dramatically affected by the choice of m^*. This is the basis for the heuristic search procedure for the optimal twisted mean value, described in [16]. Fig. 1 is an example of plotting the estimated $\Pr(Q_k > b)$, while Fig. 2 plots the normalized variance $\sigma^2_{\hat{P}}/\hat{P}^2$ of $\hat{P}$, both versus the twisted mean value m^*. The value corresponding to $m^* = -0.5$ is in fact the result of direct (conventional) Monte Carlo (MC) simulation. We can see that, as m^* increases, the normalized variance exhibits a clear "valley" around the most favorable values of m^*. This behavior, as well as the behavior of the estimated $\Pr(Q_k > b)$ versus m^*, is discussed in detail in [16] and the references therein. The minimum normalized variance appears around $m^* = 0.2$ which coincides with the approximate value m^*_{opt} of Section IV.C.

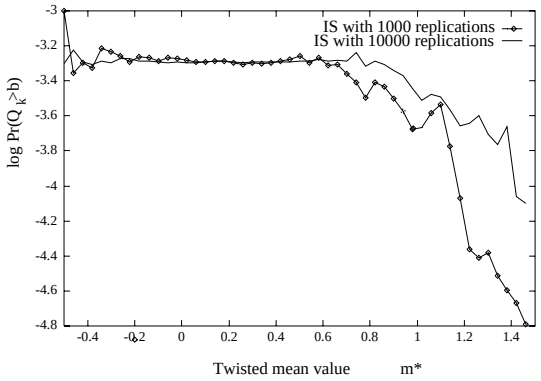

Figure 1: Estimated $\log \Pr(Q_\infty > b)$ versus the twisted mean value m^*. The Hurst parameter is $H = 0.7$.

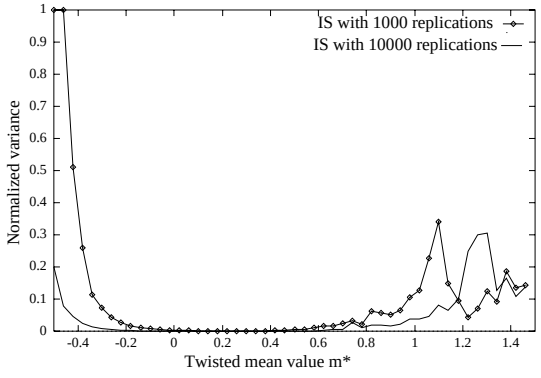

Figure 2: Normalized variance $\sigma^2_{\hat{P}}/\hat{P}^2$ of estimated $\log \Pr(Q_\infty > b)$ versus the twisted mean value m^*. The Hurst parameter is $H = 0.7$.

2. The dependence on μ:

Fig. 3 shows the estimated $\log \Pr(Q_\infty > b)$ versus the service rate μ. In all simulations, we apply the IS technique using the near-optimal twisted mean value of Section IV.C. Our simulation result is compared with the optimistic bound of Section IV.C.

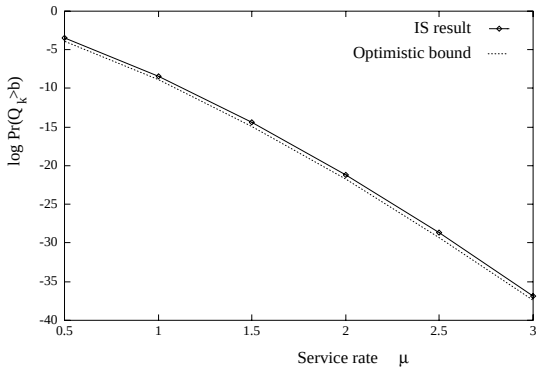

Figure 3: Estimated $\log \Pr(Q_\infty > b)$ versus the service rate μ. The Hurst parameter is $H = 0.7$.

3. The dependence on k:

Fig. 4 depicts the estimated $\log \Pr(Q_k > b)$ versus the stopping time k. The dependence of $\log \Pr(Q_k > b)$ on k reflects the transient nature of our experiments. The curves show how the queue approaches asymptotically the steady-state as k increases. In order to see how the time of entering steady-state depends on the buffer size b, in Fig. 4 we show results with different buffer sizes. For $b = 20$, we also show the direct MC simulation result in order to illustrate that the IS approach is in agreement with direct simulation. When b becomes larger, direct simulation becomes exceedingly long. In this case, IS simulation provides good results with only a minimal number of replications. Notice that the empirically observed times of entering steady-state are very close to the k_s predicted in Section IV.C, with $c = \mu/H - \mu$.

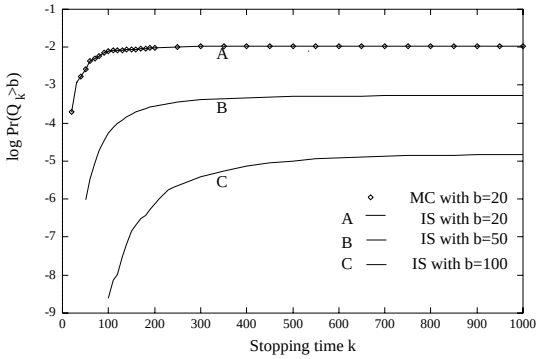

Figure 4: Estimated $\log \Pr(Q_k > b)$ versus stopping time k. The Hurst parameter is $H = 0.7$.

4. The dependence on b:

We simulate the dependence of $\log \Pr(Q_k > b)$ on b for two stopping times k: one is time k_s predicted in Section IV.C, and the other is $2 \times k_s$. We compare our simulation results with the large deviation result of equation (9) and the optimistic bound of Section IV.C, in Fig. 5. It can be seen that, with increasing stopping time, the results approach the large deviation bound, which is a steady-state result.

5. The dependence on L:

Fig. 6 shows the estimated $\log \Pr(Q_k > b)$ versus the number of homogeneous multiplexed sources L, for $H = 0.7$. Fig. 6 also depicts the optimistic bound of Section IV.C. The service rate is in fact $L \times \mu$ in order to maintain the same load on the queue. The multiplexing gain (i.e., reduction in $\Pr(Q_k > b)$ with increasing L) is evident.

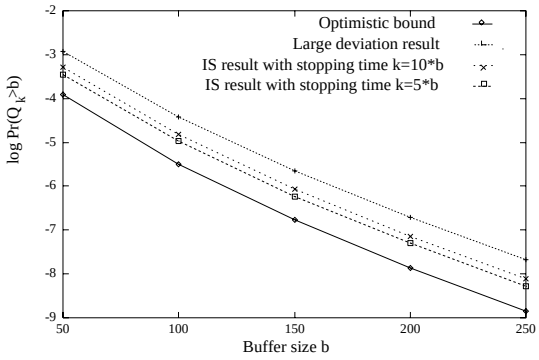

Figure 5: Estimated $\log \Pr(Q_k > b)$ versus the buffer size b. The Hurst parameter is $H = 0.7$.

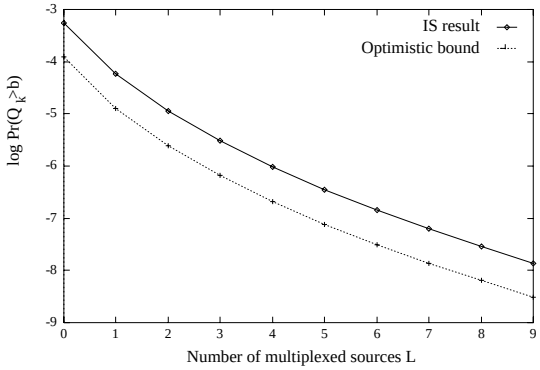

Figure 6: Estimated $\log \Pr(Q_k > b)$ versus the number of multiplexed sources L. The Hurst parameter is $H = 0.7$.

B. CASE II: $H = 0.9$

The simulation procedures are basically the same as for $H = 0.7$. Therefore, we only comment on those features which are different from previous experiments. All simulations are based on 1000 iid replications.

1. The dependence on μ:

Fig. 7 shows the estimated $\log \Pr(Q_k > b)$ versus the service rate μ, for $H = 0.9$. Comparing this result with Fig. 3, we see that increasing μ is more efficient for burstier sources.

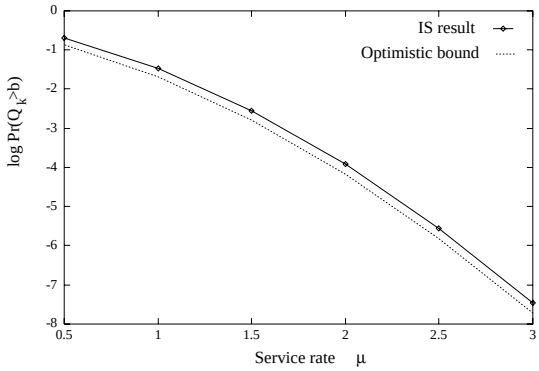

Figure 7: Estimated $\log \Pr(Q_k > b)$ versus the service rate μ. The Hurst parameter is $H = 0.9$.

2. The dependence on b:

Fig. 8 depicts the dependence of the estimated $\log \Pr(Q_k > b)$ on b, for $H = 0.9$. Comparing this result with Fig. 5, we find that increasing the buffer size is more efficient in reducing the overflow probability than for less bursty sources ($H = 0.7$), while always less efficient than for SRD models (estimated $\log \Pr(Q_k > b)$

decays less than exponentially fast).

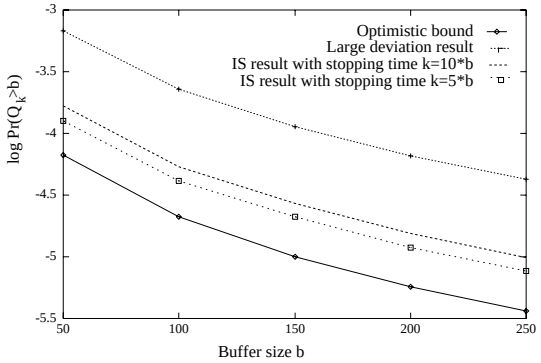

Figure 8: Estimated $\log \Pr(Q_k > b)$ versus the buffer size b. The Hurst parameter is $H = 0.9$.

3.The dependence on L:

Fig. 9 shows the estimated $\log \Pr(Q_k > b)$ versus the number of multiplexed sources L, for $H = 0.9$. Comparing Fig. 9 with Fig. 6, we see that increasing the number of multiplexed sources leads to higher gains (larger reductions in overflow probability) for burstier sources (higher values of H). We can easily check that the abovementioned dependency of $\Pr(Q_\infty > b)$ on μ, b, L is in agreement with the large deviation result (9).

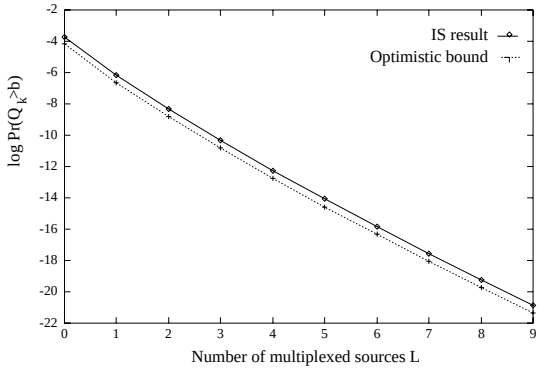

Figure 9: Estimated $\log \Pr(Q_k > b)$ versus the number of multiplexed sources L. The Hurst parameter is $H = 0.9$.

C. MULTIPLEXING HETEROGENEOUS SOURCES

Fig. 10 shows the result of multiplexing two self-similar sources, one with $H = 0.7$ and another with $H = 0.9$. As we aggregate the two arrival sources, we also increase the total service rate in order to maintain constant load, and observe the gain from increased buffer capacity. As shown in Fig. 10, the burstier source ($H = 0.9$) will dominate the queueing tail distribution, which agrees with the large deviation result in the Appendix.

D. IS IMPROVEMENT FACTOR

The speed-up or improvement factor of IS over conventional MC simulation denotes the relative decrease in the required number of replications in order to achieve the same statistical accuracy. Let $\sigma^2_{MC}(N)$ denote the estimator variance after N replications using conventional MC simulation. Furthermore, let $\sigma^2_{IS}(N)$ denote the estimator variance after N replications using IS simulation. Then the improvement factor is defined as $\sigma^2_{MC}(N)/\sigma^2_{IS}(N)$. Fig. 11 shows the estimated improvement factor versus buffer size, b, for Case I. ($H = 0.7$), and Case II. ($H = 0.9$), respectively.

We observe significant improvement factors for both cases.

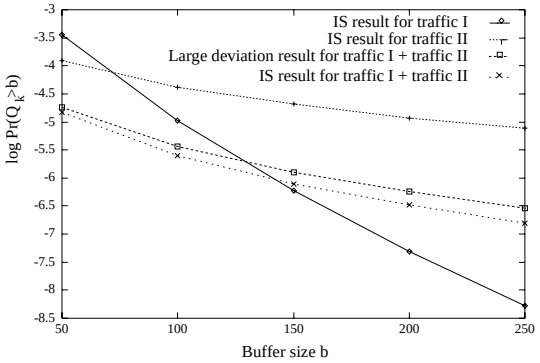

Figure 10: Estimated $\log \Pr(Q_k > b)$ versus the buffer size b (heterogeneous sources, one with $H = 0.7$, the other with $H = 0.9$). Each simulation is based on 1000 iid replications.

The improvement factor increases dramatically as the buffer size increases (i.e., as the overflow probability decreases), as is desirable.

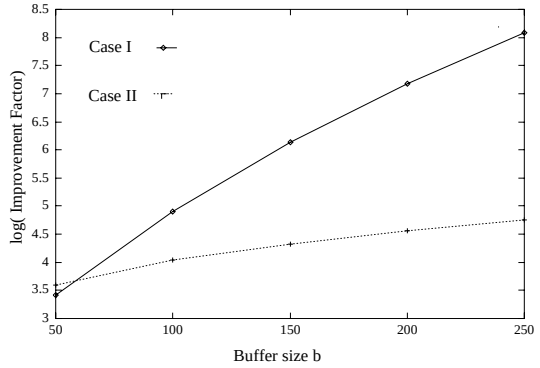

Figure 11: Estimated IS improvement factors over conventional MC simulation. Improvement factors denote the ratio of required number of replications for the same statistical accuracy, and are plotted here versus buffer size, b, for Case I. ($H = 0.7$), and Case II. ($H = 0.9$), respectively.

VI. CONCLUSIONS

Theoretical and simulation approaches applicable to traditional (short range dependent) traffic models may not be able applicable to self-similar processes due to their long range dependence. Predicted performance based on steady-state may be overly pessimistic for practical applications. Theoretical approaches to obtain transient queueing behavior and queueing distributions for small buffer size become quickly intractable.

In this paper, we have developed a fast simulation approach that can be used to simulate self-similar traffic in ATM networks efficiently. Using this approach, we have simulated the queueing and multiplexing behavior of self-similar processes in an ATM multiplexer, including the estimation of extremely low cell-loss probabilities. Our simulation experiments provide insight on transient behavior that is not possible to predict using existing analytical results. Finally, they show good agreement with existing results when asymptotically approaching steady-state.

VII. ACKNOWLEDGMENTS

This research was supported by grants from the Telecommunications Research Institute of Ontario and the National Science and Engineering Research Council of Canada.

VIII. APPENDIX

A. MULTIPLEXING HOMOGENEOUS SOURCES

Consider the aggregation of L independent FGN arrival processes $\mathbf{X}_i = \{X_{k,i},\ k = 1, 2, \ldots\}$, $i = 1, 2, \ldots, L$, with zero mean, unit variance and correlation function $r_i(k) = r(k)$, $k = 0, 1, \ldots$, where $r(k)$ is defined in equation (6). Then, the aggregate traffic $\mathbf{X}^{(L)} = \sum_{i=1}^{L} \mathbf{X}_i$ is again Gaussian, has zero mean, variance L and the same correlation function $r(k)$. Therefore, the aggregate traffic is also a FGN process. Thus the simulation procedures described in Section IV.B are directly applicable with $\sigma^2 = L$.

B. MULTIPLEXING HETEROGENEOUS SOURCES

First, we briefly summarize some important results that appear in [5] which are necessary for our results. Due to space limitations, we restrict our presentation to the very essentials leaving most of the algebraic manipulations to be checked by the interested reader. Furthermore, we urge the interested reader to consult [5] and in particular hypotheses 2.1 and 2.2 and associated theorems 2.1 and 2.2.

We now consider the aggregation of two independent FGN processes $\mathbf{X}_1$ and $\mathbf{X}_2$. We assume that $\mathbf{X}_1$ and $\mathbf{X}_2$ have zero mean and unit variance. Their corresponding correlation functions are defined as in (6) with $H = H_1$ for $\mathbf{X}_1$ and $H = H_2$ for $\mathbf{X}_2$. We assume $H_1 > H_2$ and the service rate to be μ. Then the mean of total work load process $\mathbf{W}$ is $-\mu\ k$, $k = 1, 2, \ldots$, and the variance is $k^{2H_1} + k^{2H_2}$. We can show the following lemma:

Lemma 1. Let $\mathbf{X}_i$, $i = 1, 2$, be two FGN traffic processes with zero mean, unit variance, and Hurst parameters H_i, $i = 1, 2$, respectively. Let $H_1 > H_2$ and $1/2 < H_i < 1$, $i = 1, 2$. Then the queue length process resulting from the aggregate FGN traffic satisfies:

$$\lim_{b \to \infty} b^{2(1-H_1)} \log \Pr(Q_\infty > b) = -c^{-2(1-H_1)}(c+\mu)^2/2 \qquad (24)$$

Proof: Define $a_k \triangleq k$, $v_k \triangleq \frac{k^2}{k^{2H_1}+k^{2H_2}}$, and $h_k \triangleq k^{2(1-H_1)}$. We first check the three parts of hypothesis 2.1 in [5]:

(i) It is easy to see that both a_k and v_k increase to infinity, and for all $\theta \in \mathbf{R}$

$$\lambda(\theta) = \lim_{k \to \infty} v_k^{-1} \log E e^{\theta v_k W_k / a_k} = \frac{\theta^2}{2} - \theta\mu \qquad (25)$$

(ii) It is also easy to check that $\lambda(\theta)$ is a smooth function and there exists $\theta > 0$ for which $\lambda(\theta) < 0$.

(iii) For each $c > 0$, we can show

$$g(c) = \lim_{k \to \infty} \frac{v(a^{-1}(k/c))}{h_k} = c^{2H_1 - 2} \qquad (26)$$

Therefore hypothesis 2.1 of [5] is satisfied, and we can easily get

$$\lambda^*(x) = \sup_{\theta \in \mathbf{R}} \{\theta x - \lambda(\theta)\} = \frac{(x+\mu)^2}{2} \qquad (27)$$

We now check condition (iii) in hypothesis 2.2 in [5] since conditions (i) and (ii) can be checked in a straightforward manner. We note that $v_k > k^{2-2H_1}/2$. Hence $e^{-\gamma v_k} < e^{-\gamma k^{(2-2H_1)}/2}$ *for* $\gamma > 0$

The remaining conditions (iii) and (iv) of hypothesis 2.1 in [5] follow after some algebra. Then by Theorem 2.1 and 2.2 in [5] our lemma follows.

Clearly, we have the same result as in equation (9) with $H = H_1$. This indicates that the steady-state tail distribution is dominated by the arrival process with the larger Hurst parameter.

References

[1] W. E. Leland, M. S. Taqqu, W. Willinger, and D. V. Wilson. On the Self-Similar Nature of Ethernet Traffic (Extended Version). *ACM/IEEE Transactions on Networking*, 2(1):1–15, Feb. 1994.

[2] B. B. Mandelbrot and J. W. Van Ness. Fractional Brownian Motions, Fractional Noises and Applications. *SIAM Review*, 10(4):422–437, 1968.

[3] B. B. Mandelbrot. *The Fractal Geometry of Nature*. Freeman, 1983.

[4] J. Beran, R. Sherman, M. S. Taqqu, and W. Willinger. Long-Range Dependence in Variable-Bit-Rate Video Traffic. To appear on *IEEE Transactions on Communications*, 1994.

[5] N. G. Duffield and N. O'Connell. Large Deviations and Overflow Probabilities for the General Single-Server Queue, with Applications. Technical Report DIAS-STP-93-30, Dublin Institute for Advanced Studies, 1993.

[6] I. Norros. A Storage Model with Self-Similar Input. *Queueing Systems*, 16:387 – 396, 1994.

[7] R. G. Addie and M. Zukerman. An Approximation for Performance Evaluation of Stationary Single Server Queues. In *Proc. IEEE INFOCOM '93*, 1993.

[8] J. R. M. Hosking. Modeling Persistence in Hydrological Time Series Using Fractional Differencing. *Water Resources Research*, 20(12):1898–1908, 1984.

[9] B. B. Mandelbrot. A Fast Fractional Gaussian Noise Generator. *Water Resources Research*, 7:543–553, 1971.

[10] A. Erramilli and R. P. Singh. The Application of Deterministic Chaotic Maps to Characterize Traffic in Broadband Packet Networks. In *Proc. 7th ITC Specialists Seminar*, 1990.

[11] H. E. Hurst. Long-Term Storage Capacity of Reservoirs. *Trans. of the Am. Soc. of Civil Eng.*, 116:770–799, 1951.

[12] J. R. M. Hosking. Fractional Differencing. *Biometrika*, 68(1):165–176, 1981.

[13] J. W. Cohen. *The Single Server Queue*. North-Holland, 1982.

[14] F. L. Ramsey. Characterization of the Partial Autocorrelation Function. *The Annals of Statistics*, 2(6):1296–1301, 1974.

[15] P. W. Glynn and D. L. Iglehart. Importance Sampling for Stochastic Simulations. *Management Science*, 35(11):1367–1392, Nov. 1989.

[16] M. Devetsikiotis and J. K. Townsend. Statistical Optimization of Dynamic Importance Sampling Parameters for Efficient Simulation of Communication Networks. *IEEE/ACM Trans. Networking*, 1(3), June 1993.

[17] P. Heidelberger. Fast Simulation of Rare Events in Queueing and Reliability Models. In *Proc. of Performance '93*, Rome, Italy, October 1993.

[18] P. O. Borjesson and C. E. W. Sundberg. Simple Approximations of the Error Function $Q(x)$ for Communications Applications. *IEEE Transactions on Communications*, pages 639–643, Mar. 1979.