FAST SIMULATION OF QUEUES WITH LONG-RANGE DEPENDENT TRAFFIC

C. Huang, M. Devetsikiotis, I. Lambadaris, and A. R. Kaye

Department of Systems & Computer Engineering
Carleton University
Ottawa, Ontario K1S 5B6, Canada

ABSTRACT

Self-similar stochastic processes have been proposed as more accurate models of certain categories of traffic (e.g., Ethernet traffic, variable-bit-rate video). Analytical and simulation approaches applicable to traditional traffic models may not be applicable to these categories of traffic due to their *long range dependence.*

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 sizes. Predicted performance based on steady-state regimes may be overly pessimistic for practical applications. Analytical approaches to obtain transient queueing behavior and queueing distributions for small buffer sizes become quickly intractable.

In this paper, we develop a fast simulation approach based on importance sampling that we use to simulate the queueing behavior of self-similar processes in a multiplexer, including the estimation of very low cell-loss probabilities. We describe both a simpler heuristic approach as well as a simulation approach inspired by *asymptotically efficient* simulation of general Gaussian processes. Our simulation experiments provide insight on transient behavior that is not possible to predict using current analytical results. Finally, our simulations show good agreement with existing results when approaching steady-state.

1 Introduction

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 possesses long-range dependent (LRD) characteristics that can be more accurately matched by *self-similar* models [2, 3].

More recently, variable-bit-rate (VBR) video traffic was also found to exhibit LRD characteristics, similarly to LAN traffic [4].

Both Ethernet and VBR video traffic streams exhibit *long range dependence*, that is, their autocorrelation function is non-summable and decays less than exponentially fast. This is in contrast to traditional stochastic models, all of which exhibit *short range dependence* (SRD), i.e., have a summable autocorrelation function. The serious implication for network design is that, conclusions based on traditional models may not be applicable to these traffic sources.

There have been only a few analytical queueing results reported in this area. In [5, 6] asymptotic expressions for the steady-state waiting time in single-server queues were derived by generalizing large deviation theorems to include LRD and 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) [2]. While strict self-similarity captures the burstiness of traffic at all time scales, realistic networks are expected to carry traffic that although long-range dependent, will still have a limiting time scale. Therefore, predicted performance based on a steady-state regime may be overly pessimistic for practical applications. Furthermore, questions regarding the transient behavior, small buffer sizes, multiplexing effects, and, in general, the performance of networks under LRD traffic, remain unanswered. For this purpose, analytical approaches become quickly intractable.

Given the difficulties in analysis, simulation can play an important role in the study of network performance under long-range dependent traffic modeled by self-similar processes. 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, 10], *aggregation of a large number of heavy-tail sources* [11], *nonlinear chaotic maps* [12, 13, 14], *random midpoint displacement* [15]), they are, in general, either approximate or exact but efficient for generating small numbers of relative long traces.

Due to the long-range dependence, accurate statistics can be obtained only from a large number of replications. This is especially true in broadband multimedia 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. Although Hosking's method is not the most efficient generation method, it is, however, exact, and by combining it with IS we illustrate how it can be made applicable to practical simulation studies.

Our transient results are consistent with the steady-state results in [5]. Furthermore, we verify experimentally the existence of a certain time scale at which the transient result is a good approximation for steady-state. Finally, we apply importance sampling to the simulation of the multiplexing effect under both homogeneous and heterogeneous traffic sources. For the case of multiplexing heterogeneous sources, we prove a Proposition stating that the traffic source with the higher Hurst parameter, H, will dominate the tail behavior of the queueing distribution (which we also verify by simulation).

We focus on the following key issues in 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 long-range dependent traffic as the Hurst parameter [16], 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, LRD models show smaller buffering gains. On the other hand, our results indicate significant gains from statistical multiplexing. These multiplexing gains appear to increase with the burstiness (Hurst parameter) of the LRD traffic.

In addition to these results, we show that when two heterogeneous LRD sources are statistically multiplexed, the steady-state behavior of a system with large buffer size will be dominated by the burstier process, as predicted by large deviation theory. Therefore, when a process possesses both long and short-range dependent components, e.g., the *fractional autoregressive integrated moving-average* (F-ARIMA) [17, 18] model, the steady-state behavior will only reflect the contribution of the long-range dependent component. This again emphasizes the need for transient in addition to steady-state analysis.

This paper is organized as follows: In Section 2 we present a brief introduction to self-similar traffic models and the existing large deviations results. In Section 3 we describe the self-similar traffic model we use, namely discrete-time FGN, and Hosking's method for generating LRD traffic traces from it. In Section 4 we develop an importance sampling technique for simulating self-similar processes. In Section 5 we present simulation results and we generalize results in [5] to include multiplexing effects (required in order to compare with our simulation study). Finally, in Section 6 we present our conclusions.

2 Self-Similar Traffic Models

2.1 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 $m = \mathrm{E}[X_k]$, finite variance $\sigma^2 = \mathrm{E}[(X_k - m)^2]$, and an autocorrelation function as follows:

$$r(k) = \frac{\mathrm{E}[(X_i - m)(X_{i+k} - m)]}{\sigma^2} \sim k^{-\beta} L(k)$$

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

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

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

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

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

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

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

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]. A crucial feature of such processes is that their aggregated processes $\mathbf{X}^{(n)}$ possess a nondegenerate correlation structure as $n \to \infty$. Mathematically, it was shown in [19] that the autocorrelations of general self-similar processes decay hyperbolically rather than exponentially fast, implying a nonsummable autocorrelation function (i.e., long range dependence).

An important recent development in traffic modeling is that Leland *et al.* [1] have found that Ethernet traffic satisfies (1), while Beran *et al.* [4] have shown that VBR video traffic satisfies (2) – (3).

2.2 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) [2] and the asymptotically self-similar *fractional autoregressive integrated moving-average* (F-ARIMA) process [17, 18], 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, as shown in [9]. While we consider only FGN models in this paper, our approach can be easily extended to include F-ARIMA models.

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

$$r(k) = 1/2(|k+1|^{2H} - 2|k|^{2H} + |k-1|^{2H}),\ k = \ldots, -1, 0, 1, \ldots \qquad (4)$$

Therefore, if $1/2 < H < 1$, FGN is exactly second-order self-similar with Hurst parameter H. Now define process $\mathbf{Z} = \{Z_k : k = 0, 1, 2, \ldots\}$ as

$$Z_k = \sum_{i=1}^{k} X_i, \text{ for } k = 1, 2, \ldots$$

Then $\mathbf{Z}$ is a stationary increment process with mean $\alpha(k) = km$, and variance $\gamma(k) = \sigma^2 k^{2H}$ (see also [2]).

2.3 Lindley Equation and Large Deviation Result

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, without loss of generality, we assume X_k can take any real value (another name for such processes used in the literature is the *netput* process) [20]. 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 [21]:

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

where the process $\mathbf{Y} = \{Y_k : Y_k = X_k - \mu, k = 1, 2 \ldots\}$ is called the *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 and reversible process (Hosking's method yields a stationary FGN process from the start), we have [21]

$$\Pr(Q_k > b) = \Pr(\max_{0 \le i \le k} W_i > b),\ for\ k = 0, 1, 2, \ldots \tag{5}$$

Duffield *et al.* [5] have shown the following steady-state, large deviation result (assuming $m = 0$):

$$\lim_{b \to \infty} b^{-2(1-H)} \log \Pr(Q_\infty > b) = -c^{-2(1-H)} (c + \mu)^2 / 2 \tag{6}$$

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. This is to be expected since FGN models capture the burstiness of traffic at all time scales, something that traditional Markovian models cannot exhibit.

From equation (5), we have, according to [6]

$$\Pr(Q_k > b) = \Pr(\max_{0 \le i \le k} W_i > b) > \max_{0 \le i \le k} \Pr(W_i > b) \triangleq \mathrm{P}_{W,k} \tag{7}$$

This approximation, which is an optimistic bound for $\Pr(Q_k > b)$, can be quite accurate for any time k, when b is large [6]. Since W_i is a Gaussian random variable with mean $(m - \mu)i$ and variance $\sigma^2 i^{2H}$ we can write

$$\mathrm{P}_{W,\infty} \triangleq \sup_{i \ge 0} \Pr(W_i > b) = \sup_{i \ge 0} \Phi \left(Y > \frac{b + (m - \mu) i}{\sigma^2 i^{2H}} \right)$$

where Y is the standard normal random variable and Φ its cumulative probability distribution function. Furthermore,

$$\arg\sup_{i\geq 0} \Phi\left(Y > \frac{b+(m-\mu)i}{\sigma^2 i^{2H}}\right) = \arg\inf_{i\geq 0} \frac{b+(m-\mu)i}{\sigma^2 i^{2H}} = k_s$$

Without loss of generality for $m = 0$ a straightforward calculation can show that

$$k_s = \lceil b/c \rceil \tag{8}$$

where $c = \frac{\mu}{H-\mu}$ as it is defined in equation (6) [5]. In brief, the derivation above shows that as time k grows larger, there exists a value $k = k_s$ such that

$$\mathrm{P}_{W,\infty} \overset{\triangle}{=} \sup_{i\geq 0} \Pr(W_i > b) \simeq \Pr(W_{k_s} > b)$$

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 [22]. The above approximation procedures lead to quite accurate results, as our results in Section 5 indicate.

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 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, questions regarding the transient behavior, small buffer sizes, multiplexing effects, and, in general, the performance of networks under LRD traffic, remain unanswered. In the following, we develop a simulation approach that can be used to answer the above questions. But before that, we need to introduce Hosking's method to generate synthetic FGN traces.

3 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 [23]

$$\begin{aligned} m_k &= \mathrm{E}(X_k | x_{k-1}, x_{k-2}, \ldots, x_1) = \sum_{j=2}^{k} \phi_{kj} x_{k-j+1} \text{ for } k \geq 2 && (9) \\ v_k &= \mathrm{Var}(X_k | x_{k-1}, x_{k-2}, \ldots, x_1) = \sigma^2 \prod_{j=2}^{k} (1 - \phi_{jj}^2) \text{ for } k \geq 2 && (10) \end{aligned}$$

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 $\{x_1, x_2, \ldots, x_{n-1}\}$ of size n from a FGN process, [8] describes the following algorithm:

1. Generate a starting value x_1 from a Gaussian distribution $N(0, \sigma^2)$. Set $N_1 = 0$, $D_1 = 1$,$v_1 = \sigma^2$.

2. Set $N_2 = r(1)$,$D_2 = D_1$,$\phi_{22} = \frac{N_2}{D_2}$,$m_2 = \phi_{22} x_1$ and $v_2 = (1 - \phi_{22}^2) v_1$, generate a value x_2 from a Gaussian distribution $N(m_2, v_2)$.

3. For $k = 3, \ldots, n-1$, calculate ϕ_{kj}, $j = 2, \ldots, k$, recursively via the equations

$$\begin{aligned} N_k &= r(k-1) - \sum_{j=2}^{k-1} \phi_{k-1,j} r(k-j) \\ D_k &= D_{k-1} - N_{k-1}^2 / D_{k-1} \\ \phi_{kk} &= N_k / D_k \\ \phi_{kj} &= \phi_{k-1,j} - \phi_{kk} \phi_{k-1,k-j+1} \quad j = 2, \ldots, k-1 \end{aligned}$$

Calculate $m_k = \sum_{j=2}^{k} \phi_{kj} x_{k-j+1}$ 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 well-defined Gaussian process as long as the correlation function $r(k)$ is known [8]. However, due to the recursive nature of Hosking's method, 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 broadband multimedia 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 such cases, using conventional Monte Carlo simulation, we may need to generate millions of traces by using Hosking's method, which is practically infeasible. In the following, we develop a fast simulation approach based on importance sampling, that makes Hosking's method applicable to quality-of-service evaluation in communication networks.

4 Importance Sampling for the FGN Process

4.1 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

$$P = \int_{-\infty}^{\infty} I_A(t)\frac{p(t)}{p'(t)}p'(t)dx = \mathrm{E}_{p'}[I_A(U)\frac{p(U)}{p'(U)}] = \mathrm{E}_{p'}[I_A(U)L(U)] \qquad (11)$$

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 [24] and references within): Draw N samples $u_1, \ldots, u_N$ using the density p'. Then, by equation (11), 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 *biased density.* Since any density can be used as the biased density, the question arising is how to choose a *favorable* biased density, i.e., a density that reduces the variance of $\hat{P}$.

Although the unconstrained *optimal* density is easy to describe, implementing it is not practically feasible because it represents a tautology (i.e., requires knowledge of P). Typically, the search for $p'(u)$ focuses on constrained or parametric sub-optimal solutions. When A is a rare event under density $p(u)$, one needs to choose a sampling density in order to make the event A more likely to occur. In doing this, one typically reduces the variance of the estimate $\hat{P}$. A general rule for choosing a favorable biased density is to make the likelihood ratio small on the set A. Importance sampling has been successfully applied to the simulation of various SRD processes. A variety of approaches, namely analytical, large deviation-based, and statistical have been proposed in order to choose $p'(u)$ ([24, 25, 26, 27] and references within).

4.2 Biased Density and Likelihood Ratios for FGN

In the following we simulate a queueing system with a FGN arrival process $\mathbf{X}$ as defined in Section 2.2, with mean value $m = 0$. Define a new process $\mathbf{Y}' = \{Y'(k) : Y'(k) = X(k) + m_k^*, k = 1, \ldots\}$. It is easy to see that process $\mathbf{Y}'$, which we call the *biased work load process*, is a FGN process with mean m_k^*, 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 $x_j = y_j' - m_j^*$, for $j = 1, 2, \ldots, k-1$. From equations (9)–(10),

$$\begin{aligned}
\mathrm{E}_{Y'}(Y_k' | y_{k-1}', \ldots, y_1') &= m_k^* + \mathrm{E}_X(X_k | y_{k-1}' - m_{k-1}^*, \ldots, y_1' - m_1^*) \\
&= m_k^* + \mathrm{E}_X(X_k | x_{k-1}, \ldots, x_1) \\
&= m_k^* + \sum_{j=2}^{k} \phi_{kj}(x_{k-j}) \\
&= m_k^* + \sum_{j=2}^{k} \phi_{kj}(y_{k-j}' - m_{k-j}^*) \\
&= m_k^* + m_{k,Y'} \;\; \text{for } k = 2, 3, \ldots \qquad (12)
\end{aligned}$$

where

$$m_{k,Y'} \triangleq \sum_{j=2}^{k} \phi_{kj}(y_{k-j}' - m_{k-j}^*)$$

Also from equations (9)–(10)

$$\mathrm{Var}_{Y'}(Y_k' | y_{k-1}', \ldots, y_1') = \mathrm{Var}_X(X_k | x_{k-1}, \ldots, x_1)$$

In IS simulation, we simulate a biased 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 2.3. Then,

$$\begin{aligned} \mathrm{E}_Y(Y_k | y'_{k-1}, \ldots, y'_1) &= -\mu + \sum_{j=2}^{k} \phi_{kj}(y'_{k-j} + \mu) \\ &= -\mu + m_{k,Y} \text{ for } k = 2, 3, \ldots \end{aligned} \tag{13}$$

where

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

We also have

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

The likelihood ratio up to time k is

$$\begin{aligned} L(k) &= \frac{f_Y(y'_1, \ldots, y'_k)}{f_{Y'}(y'_1, \ldots, y'_k)} \\ &= \frac{f_Y(y'_1) f_Y(y'_2 | y'_1) \cdots f_Y(y'_k | y'_{k-1}, \ldots, y'_1)}{f_{Y'}(y'_1) f_{Y'}(y'_2 | y'_1) \cdots f_{Y'}(y'_k | y'_{k-1}, \ldots, y'_1)} \\ &= \prod_{i=1}^{k} L_i \end{aligned} \tag{14}$$

where

$$\begin{aligned} L_1 &= \frac{f_Y(y'_1)}{f_{Y'}(y'_1)} \\ L_i &= \frac{f_Y(y'_i | y'_{i-1}, \ldots, y'_1)}{f_{Y'}(y'_i | y'_{i-1}, \ldots, y'_1)} \quad for\ i = 2, 3, \ldots, k \end{aligned}$$

Then, from equations (12) to (13), we have

$$L_i = \frac{e^{\theta_i y'_i}}{M_i}, \quad \text{for } i = 2, 3, \ldots$$

where

$$\begin{aligned} \theta_i &= -\frac{\mu - m_{i,Y} + m^*_i + m_{i,Y'}}{\sigma^2 \prod_{j=2}^{i}(1 - \phi^2_{jj})} \\ M_i &= e^{-\theta_i(\mu - m_{i,Y} - m^*_i - m_{i,Y'})/2} \end{aligned}$$

and

$$L_1 = e^{-\frac{2(m^*_1 + \mu) y'_1 + \mu^2 - m^{*2}_1}{2\sigma^2}} \tag{15}$$

The probability $\Pr(Q_k > b)$ can be estimated by observing N i.i.d. 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. 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 3;
3. Generate a sample point y_i' by the equation $y_i' = y_i + m_i^* + \mu = x_i + m_i^*$;
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 (14) to (15);
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.

4.3 Optimal Biased Mean Value

Based on the above description, we can apply IS by suitably modifying (*biasing*) the mean of the arrival process. However, an efficient method to obtain a favorable (or near-optimal) biased mean value remains to be devised. In this paper, we describe two such methods: a simpler, approximate analytic approach, supported by a heuristic search based on stochastic optimization; and an analytical method inspired by the Large Deviation-based "asymptotically efficient" biasing for Gaussian processes described in [28].

4.3.1 Approximate Approach

We first focus our attention on the approximate analytical approach. Since $\Pr(Q_\infty > b) \simeq \Pr(W_{k_s} > b)$, our approximate analytical approach consists of finding a near-optimal mean biased value for $\Pr(W_{k_s} > b)$ and then applying that same biased value to the simulation of $\Pr(Q_\infty > b)$. Since W_{k_s} is normally distributed with mean $-\mu k_s$ and variance $\sigma^2 k_s^{2H}$, the likelihood ratio at buffer size b will be

$$L(k_s, b) = \frac{e^{-\frac{(b+\mu k_s)^2}{2\sigma^2 k_s^{2H}}}}{e^{-\frac{(b-m_W^*)^2}{2\sigma^2 k_s^{2H}}}}$$

where m_W^* is the biased mean value. By minimizing the above likelihood ratio as suggested in [29, 27], we can find a near-optimal biased mean $m_{W,opt}^* \simeq b = ck_s$. Hence, a near-optimal *constant* biased mean value for process $\mathbf{Y}$ can be found as follows

$$m_{i,opt}^* = m_{opt}^* \simeq m_{W,opt}^*/k_s \simeq c = \mu/H - \mu, \qquad i = 1, 2, \ldots, k \tag{16}$$

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 $k < k_s$.

The stochastic search approach has been successfully applied to traditional (SRD) models (see [26] and references within), and will be briefly explained in Section 5. In the next section we will show, using numerical examples, that the results of the heuristic stochastic search and the approximate result described in the last paragraph are in very close agreement. Therefore, the above approximate value for $m_{i,opt}^* = m_{opt}^*$ can be used directly or provide a good initial estimate for the further search for a near-optimal biased mean value.

4.3.2 Analytical Approach Based on Large Deviations

Following the exposition and notation in [28] on "Systems with small Gaussian inputs," let $\{\mathbf{Y}_n,\ n = 1, 2, \ldots\}$ be a sequence of d-dimensional Gaussian vectors with distribution $F_n(d\mathbf{y})$, mean $\boldsymbol{\psi}$, and covariance $\sigma_n^2\,\mathbf{C}$, where $\sigma_n^2 = \sigma_0^2/n$ and $\mathbf{C}$ is positive definite.

The asymptotic log-moment generating function

$$\mu(\mathbf{s}) = \lim_{n\to\infty} \mu_n(\mathbf{s}) = \lim_{n\to\infty} \frac{1}{n} \log \mathrm{E}[\exp(\mathrm{n}\,\mathbf{s}\cdot\mathbf{Y_n})]$$

where $\mathbf{s} \in \Re^d$ and $\cdot$ denotes the Euclidean dot product, is in this case equal to $\mu(\mathbf{s}) = [\sigma_0^2 \mathbf{s}^T \mathbf{C}\mathbf{s}]/2 + \mathbf{s}^T\boldsymbol{\psi}$ (the superscript T denotes vector transpose). Furthermore, the Large Deviation rate function $I(\mathbf{y})$, defined as the *Legendre-Fenchel* transform of $\mu(\mathbf{s})$, is

$$I(\mathbf{y}) = \sup_{s\in\Re^d}\{\mathbf{s}\cdot\mathbf{y} - \mu(\mathbf{s})\} = \frac{(\mathbf{y}-\boldsymbol{\psi})^T\mathbf{C}^{-1}(\mathbf{y}-\boldsymbol{\psi})}{2\sigma_0^2}$$

Let the goal of the simulation be to estimate the probability $P_n = \Pr(\mathbf{Y}_n \in E) = \int 1_E(\mathbf{y})\,F_n(d\mathbf{y})$, where $1_E(\cdot)$ is the indicator function of the Borel set E.

As a special case of the discussion in [28], let E be defined as $\{g(\mathbf{Y}_n) \geq b\}$, where $g(\mathbf{Y}_n)$ is a linear combination of the elements of $\mathbf{Y}_n$, and specifically in this case, merely the *sum* of the elements of $\mathbf{Y}_n$ (therefore E is a half-space).

Then the hypotheses of *Theorem 1* in [28] hold and $\lim_{n\to\infty} \frac{1}{n} \log(P_n) = -I(E)$, where $I(E) = \inf_{\mathbf{t}\in E} I(\mathbf{t})$ (the *Cramér* transform of E). Furthermore, there exists a "dominating point" $\mathbf{v}$ which is also the unique "minimum rate point" (according to the definition in [28]) that satisfies $\mathbf{v} \in \partial E$ and $I(\mathbf{v}) = I(E) = \inf_{\mathbf{y}\in E} I(\mathbf{y})$.

Let the "biased" distribution used in the importance sampling simulation be given by

$$F_n(d\mathbf{y}) = \exp[n\,(\mathbf{s}\cdot\mathbf{y} - \mu_n(\mathbf{s}))]\, F_n(d\mathbf{y})$$

Then, it follows from [28] that

$$F_n^*(d\mathbf{y}) = F_n^{(\mathbf{v})}(d\mathbf{y}) = \exp[n\,(\mathbf{s_v}\cdot\mathbf{y} - \mu_n(\mathbf{s_v}))]\, F_n(d\mathbf{y})$$

where $\nabla\mu(\mathbf{s_v}) = \mathbf{v}$ (or $\nabla I(\mathbf{v}) = \mathbf{s_v}$) and $\mathbf{v}$ is the *dominating point*, is an "asymptotically efficient" sampling distribution. Clearly, $F_n^*(d\mathbf{y})$ is again Gaussian with the same covariance matrix $\sigma_n^2\mathbf{C}$ but with its mean *translated* to the point $\mathbf{v}$.

In our case, we estimate the probability

$$\Pr(Q_k > b) = \Pr(\max_{0\leq i\leq k} W_i > b) > \max_{0\leq i\leq k} \Pr(\sum_{i=1}^{k} Y_i > b)$$

according to (7), where $\{Y_i, i = 1, 2, \ldots\}$ is an FGN trace, with $\mathbf{C} = \{C_{ij}\} = \{r(i-j)\}$ (from (4)) and mean $\mathbf{m} = [m-\mu, m-\mu, \ldots, m-\mu]$ (μ is the service rate, and m has been assumed earlier to be equal to zero). Because of the *linear* form of $g(\cdot)$, we can easily apply quadratic programming (see for example [30]) to locate the dominating point $\mathbf{v}$ by calculating the minimum of $I(\mathbf{y})$ constrained on the boundary of E, that is by solving:

$$\min I(\mathbf{y})$$

$$\text{s.t.} \qquad \sum_{i=1}^{k} y_i = b$$

(where y_i are the elements of $\mathbf{y}$) leading to [30]:

$$\mathbf{v} = \mathbf{Ch}(\mathbf{h}^T\mathbf{Ch})^{-1}\,(b - (\mathbf{h}^T\mathbf{m})) + \mathbf{m}$$

where $\mathbf{h} = [1, 1, 1, \ldots, 1]^T$.

5 Numerical Results

For IS simulation, the estimator $\hat{P}$ of the unknown probability $\Pr(Q_k > b)$ depends on $m, m^*, \mu, H, k, b, N, \sigma^2$. 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. Furthermore, we let $m = 0$.

We focus in our simulation experiments on two types of traffic, one with $H = 0.7$, and one with $H = 0.9$ representing burstier traffic. In each case, we consequently discuss the dependence of $\Pr(Q_k > b)$ on the termination time k, the buffer size b, and on L, i.e., the number of multiplexed homogeneous sources minus one. By homogeneous sources we mean sources which have the same Hurst parameter. In the final part, we simulate multiplexing two heterogeneous sources, one with $H = 0.7$ and one with $H = 0.9$. We also provide representative values of the run-time improvement factor of our IS technique over conventional MC simulation.

5.1 The Choice of Biased Mean 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_i^*, $i = 1, 2, \ldots$. However, the sample path properties as well as the variance of the IS estimator are dramatically affected by the choice of m_i^*. This is the basis for the heuristic search procedure for the optimal biased mean value, described in [26]. Fig. 1 is an example of plotting the estimated $\Pr(Q_k > b)$, for $b = 50$, while Fig. 2 plots the normalized variance $\sigma_{\hat{P}}^2 / \hat{P}^2$ of $\hat{P}$ (the estimator of $\Pr(Q_k > b)$), both versus a *constant* biased mean value $m_i^* = m^*$. In this experiment we assume that $H = 0.7$ and $\mu = 0.5$. The estimate corresponding to $m^* = -0.5$ is the result of direct (conventional) Monte Carlo (MC) simulation. We can see that, as m^* increases, the normalized variance exhibits an obvious "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 [26] and the references therein. For Case I, the minimum normalized variance appears around $m^* = 0.21$ which coincides with the approximate value m_{opt}^* from equation (16) which for this example turns out to be 0.214. For Case II, a near-optimal value $m^* = 0.22$ was found in a similar way.

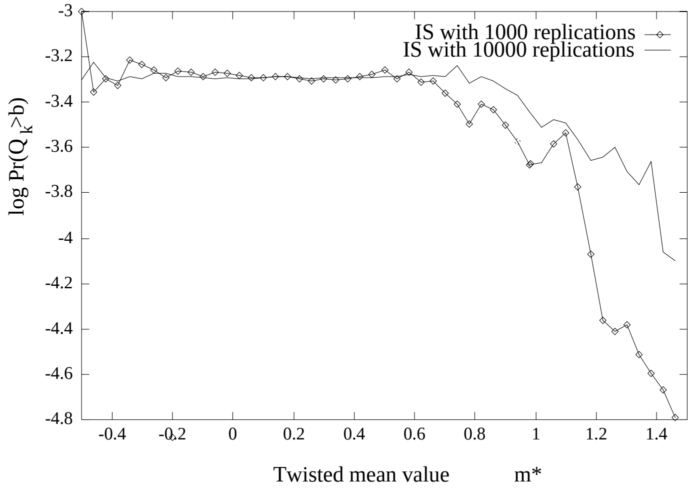

Figure 1: Estimated $\log \Pr(Q_\infty > b)$ versus the biased mean value m^*. The Hurst parameter is $H = 0.7$, $b = 50$, $\mu = 0.5$, and $k = 500$.

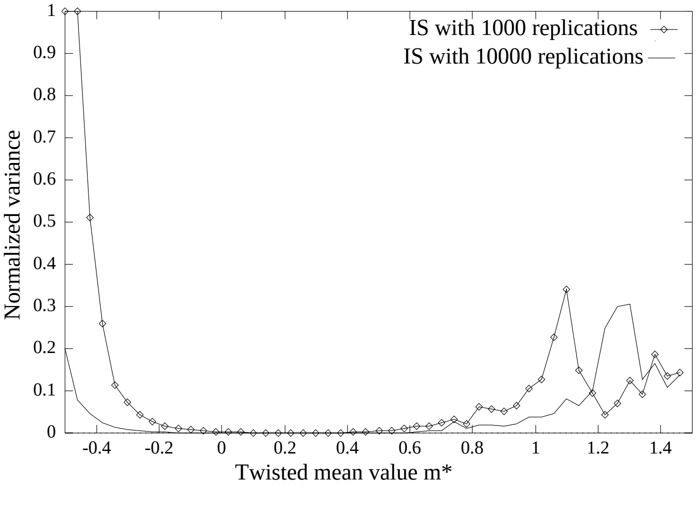

Figure 2: Normalized variance $\sigma^2_{\hat{P}}/\hat{P}^2$ of estimated $\log \Pr(Q_\infty > b)$ versus the biased mean value m^*. The Hurst parameter is $H = 0.7$, $b = 50$, $\mu = 0.5$, and $k = 500$.

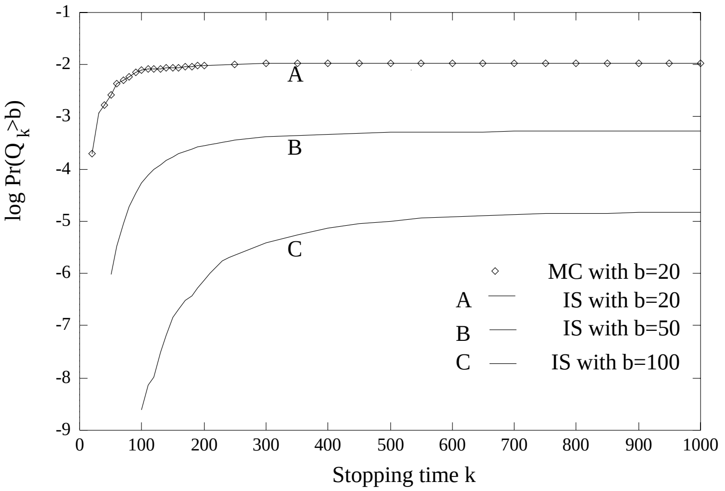

Figure 3: Estimated $\log \Pr(Q_k > b)$ versus termination time k. Each simulation is based on 1000 i.i.d. replications. The Hurst parameter is $H = 0.7$, $\mu = 0.5$, and $m^* = 0.21$.

5.2 Case I: $H = 0.7$

5.2.1 The dependence on the termination time k

Fig. 3 depicts the estimated $\log \Pr(Q_k > b)$ versus the termination time k. Each simulation is based on 1000 i.i.d. replications. The dependence of $\log \Pr(Q_k > b)$ on k reflects the transient nature of our experiments. The curves show how the queueing state 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. 3 we show results with different buffer sizes. For $b = 20$, we also show the direct MC simulation result in order to illustrate the agreement with the IS approach. With b increasing conventional MC simulation may become impractically long. In this case simulations based on IS may provide accurate results with a minimum number of independent replications (in our case 1000). The reader should also notice that the transient period of the system as observed from the simulation is very close to the k_s predicted by equation (8), with $c = \mu/H - \mu$.

5.2.2 The dependence on the buffer size b

We simulate the dependence of $\log \Pr(Q_k > b)$ on b for two termination times k: one is the time k_s as predicted by equation (8), and the other is $2 \times k_s$. We compare our simulation results with the large deviation result of equation (6)

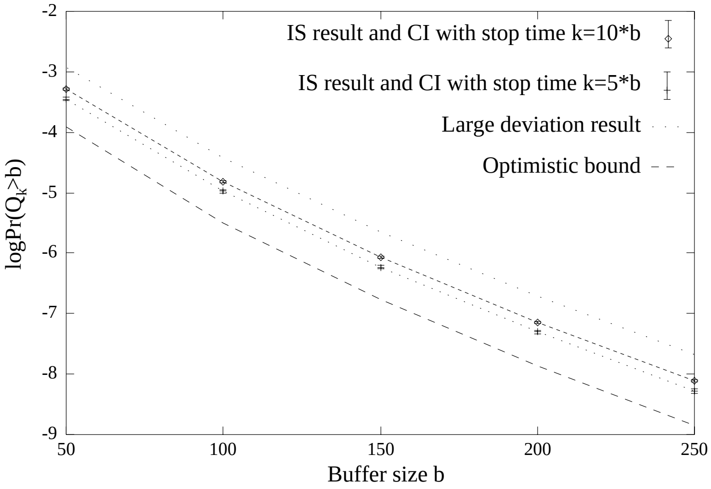

Figure 4: Estimated $\log \Pr(Q_k > b)$ versus the buffer size b and their corresponding confidence intervals. Each simulation is based on 1000 i.i.d. replications. The Hurst parameter is $H = 0.7$, $\mu = 0.5$, and $m^* = 0.21$.

and the optimistic bound of equation (7) in Fig. 4. Each simulation is based on 1000 i.i.d. replications. It can be seen that, with increasing termination time, the difference between the simulation results and the large deviation result is reduced. This behavior is indeed expected since the large deviation result is based on a steady-state regime while our simulation captures the transient behavior of the system.

5.2.3 The dependence on the number of multiplexed 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 (4). 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 4.2 are directly applicable with $\sigma^2 = L$.

Fig. 5 shows the estimated $\log \Pr(Q_k > b)$ versus L, the number of homogeneous multiplexed sources minus one, for $H = 0.7$. Each simulation is based on 1000 i.i.d. replications. Fig. 5 also depicts the optimistic bound of equation (7). The service rate is adjusted to be $L \times \mu$ 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.

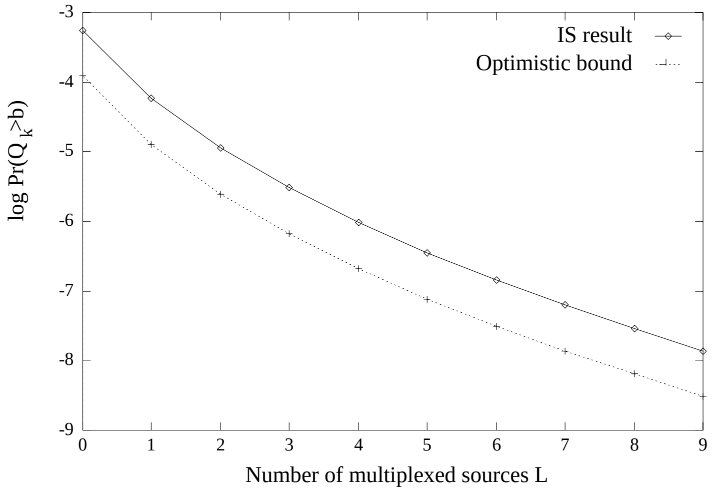

Figure 5: Estimated $\log \Pr(Q_k > b)$ versus the number of multiplexed sources L. Each simulation is based on 1000 i.i.d. replications. The Hurst parameter is $H = 0.7$, $\mu = 0.5L$, $b = 50(L+1)0.21$, $k = 500/(L+1)$.

5.3 Case II: $H = 0.9$

Since the simulation procedure is very similar to Case I, we only comment on those features which are different from previous experiments.

5.3.1 The dependence on the buffer size b

Fig. 6 depicts the dependence of the estimated $\log \Pr(Q_k > b)$ on b, for $H = 0.9$. Comparing this result with Fig. 4, we find that increasing the buffer size is less effective in reducing the overflow probability than for less bursty sources ($H = 0.7$), while always less effective when compared with SRD models (estimated $\Pr(Q_k > b)$ decays less than exponentially fast). This agrees with the theory of large deviations which predicts that $\Pr(Q_\infty > b) \approx de^{-ab^{2(1-H)}}$ for large b, where a, d are positive, slowly changing functions of b.

5.3.2 The dependence on the number of multiplexed sources

Fig. 7 shows the estimated $\log \Pr(Q_k > b)$ versus L, the number of multiplexed sources minus one, for $H = 0.9$. A comparison of Fig. 7 with Fig. 5 reveals that increasing the number of multiplexed sources leads to higher gains (larger reductions in overflow probability) for burstier sources (higher values of H). Using large deviation theory (but also from the optimistic bound of Section 4)

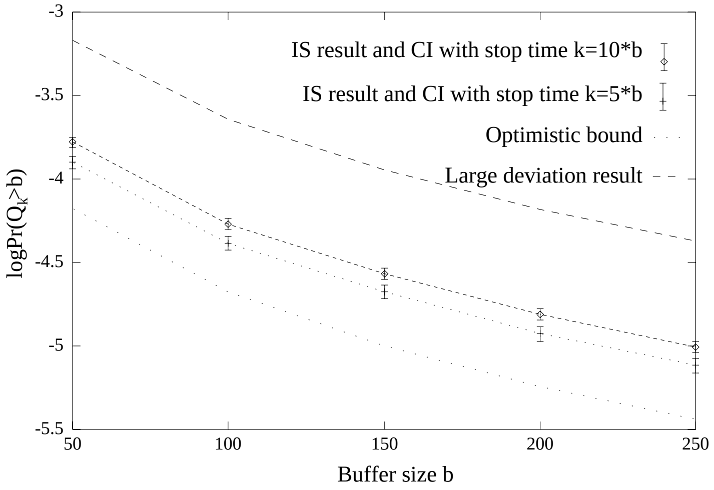

Figure 6: Estimated $\log \Pr(Q_k > b)$ versus the buffer size b and their corresponding confidence intervals. Each simulation is based on 1000 i.i.d. replications. The Hurst parameter is $H = 0.9$, $\mu = 2.0$, and $m^* = 0.22$.

we obtain that $\Pr(Q_\infty > b) \approx de^{-aL^{2H-1}}$ for large b, where a, d are positive, slowly changing functions of L. This fact is in agreement with our simulation results.

5.4 Case III: Multiplexing Heterogeneous Sources

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 (4) 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 proposition:

Proposition 1. Let $\mathbf{X}_i$, $i = 1, 2$, be two FGN traffic processes with zero mean, variances σ_i^2, 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:

$$\lim_{b \to \infty} \sigma_1^2 b^{-2(1-H_1)} \log \Pr(Q_\infty > b) = -c^{-2(1-H_1)}(c + \mu)^2/2$$

The proof is given in the Appendix.□

Clearly, we have the same result as in equation (6) with $H = H_1$. This indicates that the steady-state tail distribution is dominated by the arrival process with the larger Hurst parameter.

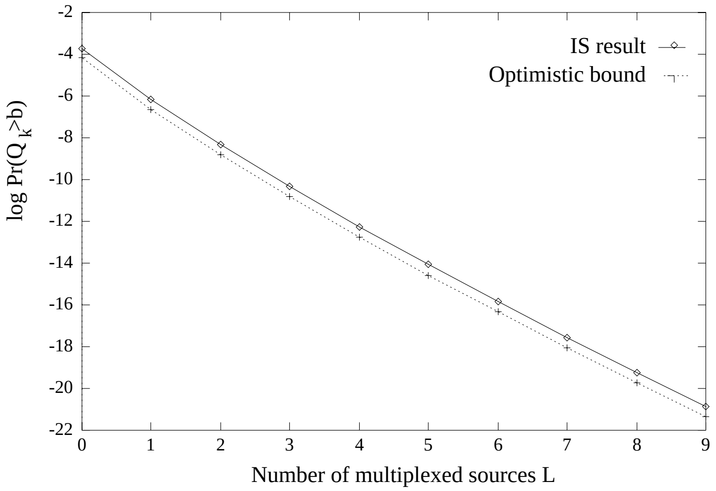

Figure 7: Estimated $\log \Pr(Q_k > b)$ versus the number of multiplexed sources L. Each simulation is based on 1000 i.i.d. replications. The Hurst parameter is $H = 0.9$, $\mu = 2.0L$, $b = 50(L+1)0.22$, $k = 1000/(L+1)$.

The simulation procedures for multiplexing heterogeneous sources are similar to the steps for a single source if we note that the aggregate process is still a Gaussian process and its mean, variance and autocorrelation function can be calculated from the corresponding values of individual sources. Therefore, we simulate by generating two independent FGN traces according to Hosking's method and aggregate them in order to calculate the total input during each time slot. While applying importance sampling, each FGN process is biased separately (different Hurst parameter values H imply different biasing values) and the likelihood ratio is taken as the *product* of the likelihood ratios corresponding to each trace (due to independence).

Fig. 8 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 accordingly the total service rate in order to maintain constant load, and observe the gain from increased buffer capacity. As shown in Fig. 8, the burstier source ($H = 0.9$) will dominate the queueing tail distribution, which agrees with Proposition 1 above.

5.5 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

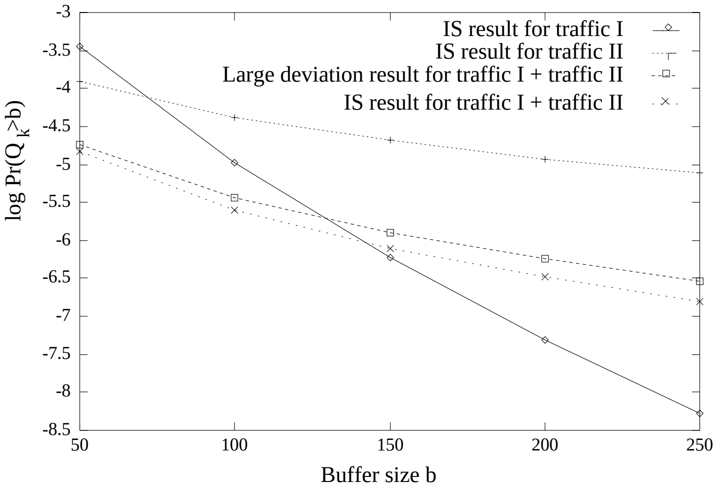

Figure 8: Estimated $\log \Pr(Q_k > b)$ versus the buffer size b (heterogeneous sources, one with $H = 0.7$, the other with $H = 0.9$). Each simulation is based on 1000 i.i.d. replications, $\mu = 2.5$, $m^* = \mu/(H_{II} - \mu)$, $k = b(H_{II} - \mu)/\mu$.

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 $\sigma^2_{MC}(N)/\sigma^2_{IS}(N)$.

Denote with P the probability $\Pr(Q_k > b)$ to be estimated using N i.i.d. replications. Then, $\sigma^2_{MC}(N) = P(1 - P)/N$. Since only an estimate $\hat{P}$ of P is known, we use the approximation $\sigma^2_{MC}(N) \simeq \hat{P}(1 - \hat{P})/N$. We also approximate the true $\sigma^2_{IS}(N)$ with a sample variance estimate. Fig. 9 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. The improvement factor increases dramatically as the buffer size increases (i.e., as the overflow probability decreases), a fact that clearly demonstrates the effectiveness of IS.

5.6 Results Based on Large Deviations

In order to illustrate the effectiveness of our second scheme that is based on large deviations and the notion of asymptotic efficiency [28], we present simulation results where the importance sampling parameter values $m_i, i = 1, 2, \ldots$ are calculated as in section 4.3.2 (called hereafter the "non-uniform"

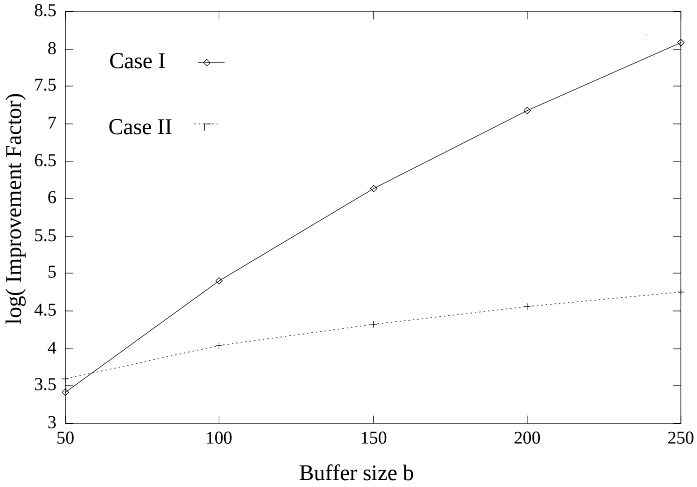

Figure 9: Estimated IS improvement factors over conventional MC simulation. Improvement factors denote the ratio of required number of replications for the same statistical accuracy, and are plotted here versus buffer size, b, for Case I ($H = 0.7$, $\mu = 0.5$, $m^* = 0.21$), and Case II ($H = 0.9$, $\mu = 2.0$, $m^* = 0.22$), respectively.

bias, as opposed to the simpler method followed earlier that we refer to as the "uniform" bias).

For all fast simulation cases, the number of independent replications was 1000 and the stop time was $k = 1000$. For the conventional Monte Carlo cases, the number of independent replications was 100,000 and the stop time was also $k = 1000$.

The service rates are $\mu = 2.0$ for the case of $H = 0.9$ and $\mu = 0.5$ for the case of $H = 0.7$. The purpose of these simulations was to estimate the *relative accuracy* of the non-uniform biasing scheme and compare it to that of the uniform biasing scheme and of conventional Monte Carlo simulation. For this purpose, we calculate the relative accuracy of each simulation run by estimating the quantity $\sqrt{\widehat{var}(\hat{P})}/\hat{P}$ (proportional to the half-width of the confidence interval normalized by the sample mean). The results are illustrated in Fig. 10 and Fig. 11, and they demonstrate both the consistently superior performance the non-uniform biasing scheme (as expected) but also the fact that the simpler to implement, uniform biasing performs also very well, as was illustrated already in the previous sections. On the other hand, the accuracy of conventional Monte Carlo simulation deteriorates very rapidly (approximately exponentially) with the buffer size, making conventional Monte Carlo completely impractical for buffer sizes $200 < b < 4000$ (at $b = 200$, the probability

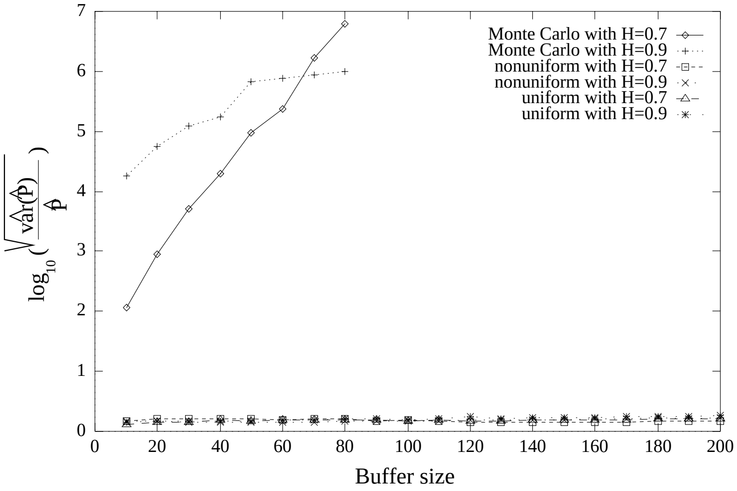

Figure 10: Plot of the logarithm of relative simulation accuracy, $\sqrt{\widehat{var}(\hat{P})}/\hat{P}$, versus buffer size ($b \leq 200$).

P has already fallen to 10^{-5} for $H = 0.9$, and 10^{-7} for $H = 0.7$).

Finally, our empirical plot of $\log(\hat{\mathrm{E}}[\hat{P}^2])/\log(\hat{P})$ is shown in Fig. 12–13, and, for the case of non-uniform biasing, appears to converge to 2.0 as the buffer size increases, indicating an approximately "asymptotic efficient" or "optimal" behavior [28, 31]. In contrast, for the case of conventional Monte Carlo simulation, the estimate of $\log(\hat{\mathrm{E}}[\hat{P}^2])/\log(\hat{P})$ remains almost constant at 1.0.

6 Conclusions

Analytical and simulation approaches applicable to traditional (short range dependent) traffic models may not be applicable to long-range dependent traffic modeled by self-similar processes. Furthermore, predicted performance based on a steady-state regime may be overly pessimistic for practical applications. However, analytical approaches to obtain transient queueing behavior and queueing distributions for small buffer sizes become quickly intractable.

In this paper, we have developed a fast simulation approach based on importance sampling that can be used to simulate long-range dependent, self-similar traffic in queues efficiently. Using this approach, we have simulated the queueing behavior of self-similar processes in a multiplexer, including the estimation of extremely low cell-loss probabilities. Our simulation experiments

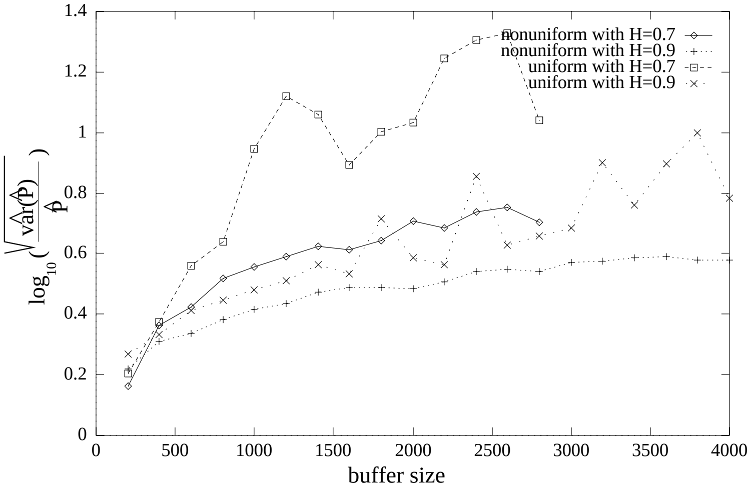

Figure 11: Plot of relative simulation accuracy, $\sqrt{\widehat{var}(\hat{P})}/\hat{P}$, versus buffer size ($200 \leq b \leq 4000$).

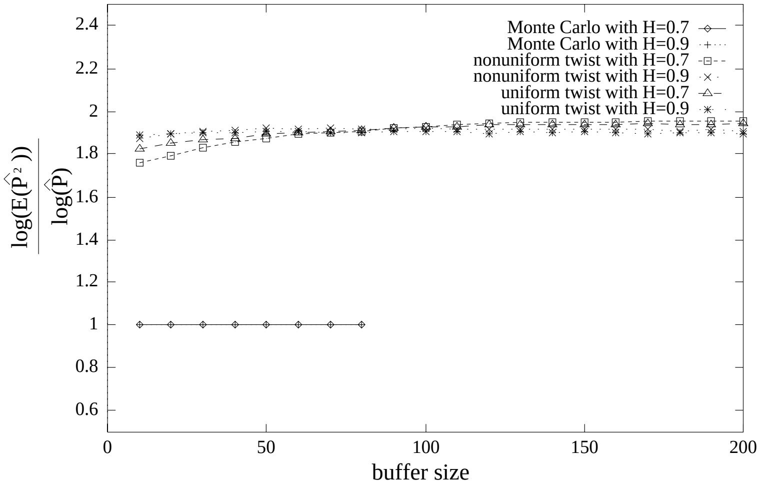

Figure 12: Plot of $\log(\hat{\mathrm{E}}[\hat{P}^2])/\log(\hat{P})$ versus buffer size ($b \leq 200$). The graph for the case of importance sampling converges to 2.0 as buffer size increases, indicating "asymptotically efficient" behavior, which is not the case for conventional Monte Carlo.

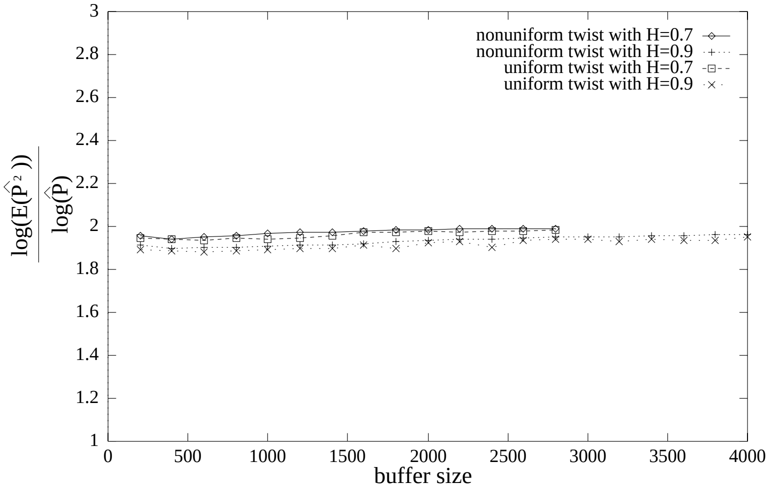

Figure 13: Plot of $\log(\hat{\mathrm{E}}[\hat{P}^2])/\log(\hat{P})$ versus buffer size ($200 \leq b \leq 4000$). The graph for the case of importance sampling converges to 2.0 as buffer size increases, indicating "asymptotically efficient" behavior.

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.

Acknowledgements

This work was supported by the Telecommunications Research Institute of Ontario (TRIO) and the Natural Sciences and Engineering Research Council of Canada (NSERC).

Appendix: Proof of Proposition 1

First, we briefly summarize some important results that appear in [5] which are necessary for our results. Due to space restrictions, we restrict our presentation to the very essentials leaving most of the algebraic manipulations to be checked by the interested reader. We start by the following two assumptions:

Hypothesis A [5]: *(i) There exist functions $a, v : \mathbf{Z}_+ \rightarrow \mathbf{R}_+$ that increase to infinity, such that for each $\theta \in \mathbf{R}$, the cumulant generating function defined as the limit*

$$\lambda(\theta) \triangleq \lim_{k\rightarrow\infty} v_k^{-1} \log E e^{\theta v_k W_k / a_k}$$

exists as an extended real number.

(ii) $\lambda(.)$ is essentially smooth, lower semi-continuous and there exists $\theta > 0$ for which $\lambda(\theta) < 0$. Note that λ is automatically convex.

(iii) There exists an increasing function $h : \mathbf{Z}_+ \to \mathbf{R}_+$ such that the limit

$$g(c) \triangleq \lim_{k\to\infty} \frac{v(a^{-1}(k/c))}{h_k}$$

exists for which $c > 0$, where

$$a^{-1}(x) \triangleq \sup\{s \in \mathbf{R}_+ : a(s) \leq x\}$$

Hypothesis B [5]: *There exists $d > 0$ such that*

(i)

$$\inf_{c>0} g(c)\lambda^*(c) = \inf_{c>d} g(c)\lambda^*(c) < \infty$$

(ii)

$$\lim_{k\to\infty} \inf_{c>d} \frac{\lambda^*(c)v_k}{h(ca_k)} = \inf_{c>d} \lambda^*(c)g(c)$$

(iii) for each $\gamma > 0$

$$\limsup_{b\to\infty} h_b^{-1} \log \sum_{k=[a^{-1}(b/d)]}^{\infty} e^{-\gamma v_k} \leq -\inf_{c>0} g(c)\lambda^*(c)$$

(iv)

$$\limsup_{b\to\infty} h_b^{-1} \log a^{-1}(b/d) = 0$$

where

$$\lambda^*(x) \triangleq \sup_{\theta\in\mathbf{R}} \{\theta x - \lambda(\theta)\}$$

Now, we have the following theorem [5]:

Theorem 1. Suppose that Hypotheses A and B are satisfied, then

$$\limsup_{b\to\infty} h_b^{-1} \log \Pr(Q > b) = -\inf_{c>0} g(c)\lambda^*(c)$$

Proof of Proposition 1: Define

$$\begin{aligned} a_k &\triangleq k \\ v_k &\triangleq \frac{k^2}{\sigma_1^2 k^{2H_1} + \sigma_2^2 k^{2H_2}} \\ h_k &\triangleq \frac{k^{2(1-H_1)}}{\sigma_1^2} \end{aligned}$$

We first check Hypothesis A:

(i) It is easy to see that both a_k and v_k increase to infinity, and

$$\begin{aligned}\lambda(\theta) &= \lim_{k\to\infty} v_k^{-1}\log Ee^{\theta v_k W_k/a_k} \\ &= \frac{\theta^2}{2} - \theta\mu \quad \text{for all} \quad \theta\in\mathbf{R}\end{aligned}$$

(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

$$\begin{aligned}g(c) &= \lim_{k\to\infty}\frac{v(a^{-1}(k/c))}{h_k} \\ &= c^{2H_1-2}\end{aligned}$$

Therefore Hypothesis A is satisfied, and we can easily get

$$\begin{aligned}\lambda^*(x) &= \sup_{\theta\in\mathbf{R}}\{\theta x - \lambda(\theta)\} \\ &= \frac{(x+\mu)^2}{2}\end{aligned}$$

We now check Hypothesis B: Conditions (i) and (ii) can be checked in a straightforward manner. To check conditions (iii) we note that $\exists K > 0$ such that $\forall k > K$

$$v_k > \frac{k^{2-2H_1}}{2\sigma_1^2}$$

Hence,

$$e^{-\gamma v_k} < e^{-\frac{\gamma k^{(2-2H_1)}}{2\sigma_1^2}} \quad \text{for} \quad \gamma > 0$$

Conditions (iii) and (iv) follow after some algebra. Then by Theorem 1, Proposition 1 is proved.□

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. *IEEE Trans. on Comm.*, 43(2/3/4):1566–1579, Feb./Mar./Apr. 1995.

[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. Studies on a Model for Connectionless Traffic, Based on Fractional Brownian Motion. In *Conf. on Applied Probability in Engineering, Computer and Communication Sciences INRIA/ORSA/TIMS/SMAI*, Paris, France, 1993.

[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 and J. R. Walls. Computer Experiments with Fractional Gaussian Noises. *Water Resources Research*, 5:228–267, 1969.

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

[11] M. S. Taqqu and J. B. Levy. Using Renewal Processes to Generate Long-Range Dependence and High Variability. In E. Eberlein and M. S. Taqqu, editors, *Probability and Statistics*, pages 137 – 165. Birkhauser, Basel, 1985.

[12] 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.

[13] A. Erramilli, R. P. Singh, and P. Pruthi. Chaotic Maps as Models of Packet Traffic. In *Proc. of* Int. Teletraffic Congress*, ITC '94*, pages 329 – 338, Antibes, France, June 1994.

[14] P. Prothi and A. Erramilli. Heavy-Tailed ON/OFF Source Behavior and Self-Similar Traffic. In *Proc. ICC '95*, pages 445 – 450, June 1995.

[15] W-C Lau, A. Erramilli, J. L. Wang, and W. Willinger. Self-Similar Traffic Generation: The Random Midpoint Displacement Algorithm and its Properties. In *Proc. ICC '95*, pages 466 – 472, June 1995.

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

[17] G. E. P. Box and G. M. Jenkins. *Time Series Analysis: Forecasting and Control.* Holden Day, San Fransisco, 2nd edition, 1976.

[18] C. W. J. Granger and R. Joyeux. An Introduction to Long-Memory Time Series Models and Fractional Differencing. *J. Time Series Anal.*, 1:15 – 29, 1980.

[19] D. R. Cox. Long-Range Dependence: A Review. In H. A. David and H. T. David, editors, *Statistics: An appraisal.* The Iowa State University Press, 1984.

[20] J. M. Harrison. *Brownian Motion and Stochastic Flow Systems.* Wiley, 1985.

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

[22] 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.

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

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

[25] J. A. Bucklew. *Large Deviation Techniques in Decision, Simulation, and Estimation.* John Wiley & Sons, 1990.

[26] 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.

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

[28] J. S. Sadowsky and J. A. Bucklew. On Large Deviation Theory and Asymptotically Efficient Monte Carlo Estimation. *IEEE Trans. Inform. Theory*, IT-36(3):579–588, May 1990.

[29] D. Lu and K. Yao. Estimation Variance Bounds of Importance Sampling Simulations in Digital Communication Systems. *IEEE Trans. Commun.*, COM-39(10):1413–1417, Oct. 1991.

[30] D. G. Luenberger. *Linear and Nonlinear Programming.* Addison-Wesley, 2nd edition, 1984.

[31] C. S. Chang, P. Heidelberger, S. Juneja, and P. Shahabuddin. Effective Bandwidth and Fast Simulation of ATM Intree Networks. *Perform. Eval.*, 20:45 – 65, 1994.