Fast Simulation of Packet Loss Rates in a Shared Buffer Communications Switch

Cheng-Shang Chang *
National Tsing Hua University
Taiwan

Philip Heidelberger †
ICASE
NASA Langley Research Center
Hampton, Virginia 23665

Perwez Shahabuddin
IBM T.J. Watson Research Center
P.O. Box 704
Yorktown Heights, New York 10598

September 3, 2001

Abstract

This paper describes an efficient technique for estimating, via simulation, the probability of buffer overflows in a queueing model that arises in the analysis of ATM (Asynchronous Transfer Mode) communication switches. There are multiple streams of (autocorrelated) traffic feeding the switch that has a buffer of finite capacity. Each stream is designated as either being of high or low priority. When the queue length reaches a certain threshold, only high priority packets are admitted to the switch's buffer. The problem is to estimate the loss rate of high priority packets. An asymptotically optimal importance sampling approach is developed for this rare event simulation problem. In this approach, the importance sampling is done in two distinct phases. In the first phase, an importance sampling change of measure is used to bring the queue length up to the threshold at which low priority packets get rejected. In the second phase, a different importance sampling change of measure is used to move the queue length from the threshold to the buffer capacity.

---

*Currently on a leave of absence from IBM Research.

†Currently on a sabbatical leave from IBM Research.

1 Introduction

This paper is concerned with efficient simulation techniques for estimating very low packet loss rates, e.g. less than 10^{-9}, in models of ATM (Asynchronous Transfer Mode) communications switches. Such low packet loss rates are deemed a requirement in ATM networks and there is currently great interest in quantitative methods for analyzing the performance of ATM models. Standard simulation of such rare events is known to take a prohibitive amount of time. Therefore, research has focused on how to apply a general technique called importance sampling [23, 25] so as to speed up the simulation of such rare events. A survey, with numerous references, on the application of importance sampling to rare event simulation of models in queueing and reliability theory is given in [26], while a survey on fast simulation of reliability models is given in [32]. Relevant references on the use of importance sampling in queueing models include [1, 2, 3, 17, 18, 19, 22, 28, 29, 33, 34, 35, 37].

In this paper, we consider simulation of a single switch, the architecture and control of which has been proposed for use in ATM networks [15, 16]. The switch operates in discrete time. It is fed by K ($K \geq 1$) independent streams of traffic. Each stream is designated as being either a high priority or low priority stream. The switch can service up to c ($c \geq 1$) packets in each unit of time. The switch has a buffer of size $B = B_1 + B_2$. Whenever the queue length (buffer contents) is less than B_1, then both high and low priority traffic streams are accepted by the switch. However, when the queue length is between B_1 and $B_1 + B_2$, low priority arrivals are rejected and only high priority arrivals are admitted to the buffer. When the queue length exceeds $B_1 + B_2$, both low and high priority packets are rejected. We will consider estimating the loss rate of high priority packets.

In order to model realistic situations, it is necessary to consider the case when the arrival streams contain autocorrelation, as may occur in the transfer of video data. Although somewhat more general arrival processes can be handled, for simplicity of notation, we will consider the case when each arrival stream is governed by a Markov chain. If the number of sources K is large, and if $B_1 + B_2$ is large, then exact analysis of this model is numerically infeasible due to the problem of state space size explosion. While a "fluid" approximation to this "cell" based model can be formulated and solved (if K is not too large) [15, 16], the difference between performance measures of the fluid and cell models is neither well understood nor easily quantified. (Indeed, one application of this paper is to facilitate such an understanding.)

Fast simulation of such a switch without priorities (i.e., all packets are high priority and $B_2 = 0$) was considered in [9, 10]. In those papers, the asymptotically optimal (as

$B \to \infty$) importance sampling change of measure was described. (Here, asymptotically optimal means that the variance of the estimate goes to zero at the fastest possible rate.) This change of measure is closely related to the Gärtner-Ellis theorem and the theory of large deviations for Markov additive processes [3, 6, 7, 13, 20, 30, 36]. The relationship between "effective bandwidths" (see, e.g., [8, 14, 21, 24, 27, 39]) and asymptotically optimal fast simulation is also explored in [9, 10]. (The theory of effective bandwidths deals with the rate at which large queue length probabilities decay.)

In this paper, we show how the results of [9, 10] can be extended to the switch with priorities. The extension is unusual in several respects. First, it requires piecing together two different importance sampling changes of measure: the first involving both high and low priority packets so as to bring the queue length up to level B_1, the second involving only the high priority packets so as to increase the queue length from B_1 to $B_1 + B_2$. Assuming that the queue is stable when fed with both high and low priority traffic, both of these events are rare and need to be simulated differently. Second, additional care has to be taken in this second phase, since once the system reaches B_1, it may hover there for some period before ascending to $B_1 + B_2$ (assuming it does reach the higher level). In this paper, we describe an asymptotically optimal (as both B_1 and $B_2 \to \infty$) simulation algorithm for this problem.

The rest of the paper is organized as follows. In Section 2, the general technique of importance sampling and its application to the single priority switch is reviewed. In Section 3 we describe how to extend this approach to the switch with priorities. Experimental results are given in Section 4. The paper is then summarized in Section 5.

2 Background

To demonstrate the difficulty involved in simulating rare events, consider a simple example of estimating the probability γ_B of an event $\mathcal{R}_\mathcal{B}$ that becomes rarer and rarer as $B \to \infty$, i.e., $\lim_{B\to\infty} \gamma_B = 0$. Suppose, for simplicity, that γ_B can be written as

$$\gamma_B = \int 1_{\{x \in \mathcal{R}_\mathcal{B}\}} p(x) dx = E_p[1_{\{X \in \mathcal{R}_\mathcal{B}\}}] \tag{1}$$

where the subscript p denotes sampling from the given input density p and $1_{\{x \in \mathcal{R}_\mathcal{B}\}}$ is the indicator of the rare set $\mathcal{R}_\mathcal{B}$, i.e., $1_{\{x \in \mathcal{R}_\mathcal{B}\}} = 1$ if $x \in \mathcal{R}_\mathcal{B}$ and $1_{\{x \in \mathcal{R}_\mathcal{B}\}} = 0$ if $x \notin \mathcal{R}_\mathcal{B}$. The standard simulation estimate of γ_B is $\hat{\gamma}_B = (1/N)\sum_{n=1}^{N} I_n$ where $X_1, \ldots, X_N$ are sampled from the density p and $I_n = 1_{\{X_n \in \mathcal{R}_\mathcal{B}\}}$. Then $E_p[\hat{\gamma}_B] = \gamma$ and the variance of $\hat{\gamma}_B$ is $\gamma_B(1-\gamma_B)/N$. The relative error, defined to be the standard deviation divided by the expected value, is proportional to $1/\sqrt{\gamma_B N}$ which, for any fixed sample size N, approaches ∞ as $B \to \infty$. This means that very large sample sizes are required to achieve accurate estimates of γ_B as $B \to \infty$.

Importance sampling, when applied properly, can avoid this problem. Consider the integral representation for γ_B in Equation 1. Multiplying and dividing the integrand by another density function $p'(x)$ yields

$$\gamma_B = \int 1_{\{x \in \mathcal{R}_\mathcal{B}\}} \frac{p(x)}{p'(x)} p'(x) dx = E_{p'} \left[1_{\{X \in \mathcal{R}_\mathcal{B}\}} \frac{p(X)}{p'(X)} \right] = E_{p'}[1_{\{X \in \mathcal{R}_\mathcal{B}\}} L(X)] \tag{2}$$

where $L(x) = p(x)/p'(x)$ is called the likelihood ratio and the subscript p' denotes sampling from the density p'. Equation 2 is valid so long as $p'(x) > 0$ for all $x \in \mathcal{R}_\mathcal{B}$ such that $p(x) > 0$. Importance sampling is based on Equation 2 and can be described as follows. Draw N samples $X_1, \ldots, X_N$ using density p' and define $Z_n = L_n I_n$ where $L_n = L(X_n)$. Then, by Equation 2, $E_{p'}[Z_n] = \gamma_B$. Thus an unbiased (and strongly consistent as $N \to \infty$) estimate of γ_B is given by

$$\hat{\gamma}_B(p') = \frac{1}{N} \sum_{n=1}^{N} Z_n = \frac{1}{N} \sum_{n=1}^{N} I_n L_n, \tag{3}$$

i.e., γ_B can be estimated by simulating a random variable with a different density and unbiasing the output by multiplying by the likelihood ratio. Most papers on importance sampling deal with how to choose a change of measure (in this case the density $p'(x)$) so as to obtain variance reduction. In general, importance sampling must be applied with great care, since there are examples when it increases the variance, and, in some cases, it can even make the variance infinite. However, variance reduction is obtained by making the likelihood ratio small on $\mathcal{R}_\mathcal{B}$, which, roughly speaking, is accomplished by making $p'(x)$ large for $x \in \mathcal{R}_\mathcal{B}$, i.e., by picking p' so as to make the rare event $\mathcal{R}_\mathcal{B}$ likely to happen. Suppose there are constants d_1 and d_2 and a function $f(B)$ such that $f(B) \to 0$ as $B \to \infty$ and

$$d_1 f(B) \leq L(X) \leq d_2 f(B) \tag{4}$$

for all $X \in \mathcal{R}_\mathcal{B}$, and $P'(\mathcal{R}_\mathcal{B}) = E_{p'}[1_{\{X \in \mathcal{R}_\mathcal{B}\}}]$ stays bounded away from zero as $B \to \infty$. Then $\gamma_B = E_{p'}[L_n I_n] \geq d_1 f(B) P'(\mathcal{R}_\mathcal{B})$ and $E_{p'}[(L_n I_n)^2] \leq (d_2 f(B))^2$. Combining these two facts implies that the relative error of $\hat{\gamma}_B(p')$ remains bounded as $B \to \infty$, i.e.,

$$\limsup_{B \to \infty} \frac{\text{Standard Deviation}[\hat{\gamma}_B(p')]}{\gamma_B} < \infty \tag{5}$$

and we say the importance sampling change of measure p' has "bounded relative error." This implies that only a fixed sample size is required to obtain estimates of γ_B of a given precision, no matter how rare $\mathcal{R}_{\mathcal{B}}$ is. Such an estimate is also asymptotically optimal, in the sense that its second moment approaches zero at the fastest possible rate, $[f(B)]^2$. Note that bounded relative error is also obtained if $d_1 f(B) \leq L(X) \leq D f(B)$ for all $X \in \mathcal{R}_{\mathcal{B}}$ where D is a random variable such that $E_{p'}[D^2] < \infty$.

To estimate the steady state packet loss rate, r, we exploit a ratio formula [5, 11] that is an extension of the ratio formula that holds in regenerative systems [12, 38]. (While, with multiple Markovian sources the model may theoretically regenerative, the regenerations may occur so infrequently as to make the regenerative method useless for practical purposes.) For some subset, A, of the state space define A-cycles to begin whenever the process enters A. (We will choose A to correspond to an empty buffer.) Then

$$r = \frac{E[Y]}{E[\alpha]} \tag{6}$$

where Y is the number of packets lost during an A-cycle and α is the number of packets that arrive during an A-cycle. In Equation 6, the initial distribution is the stationary distribution conditioned on the process just entering A. Successive $A-$cycles are identically distributed, but are not independent as they would be in the regenerative method. For large buffer sizes, the event $\mathcal{R}_{\mathcal{B}} = \{Y > 0\}$ is a rare event; $\mathcal{R}_{\mathcal{B}}$ is the event that, starting with an empty buffer, the buffer reaches size B before emptying. Importance sampling can be used to estimate the numerator, while standard simulation can be used to estimate the denominator. In particular, $E[Y] = E[Y|\mathcal{R}_{\mathcal{B}}]P(\mathcal{R}_{\mathcal{B}})$. Estimation of $E[Y]$ is difficult because $P(\mathcal{R}_{\mathcal{B}})$ is small. Therefore we will concentrate on asymptotically optimal procedures for estimating $P(\mathcal{R}_{\mathcal{B}})$.

We next review how to estimate $P(\mathcal{R}_{\mathcal{B}})$ for the case of a switch of size B without priority classes. The model is a discrete time queueing model having K independent (potentially) heterogeneous Markovian sources. (Somewhat more general arrival processes can be handled; see [9, 10].) The switch can service up to c ($c \geq 1$) packets in each time slot. Let $a_k(t)$ denote the number of packets from source k that arrive during time slot t. Let $a(t) = a_1(t) + \ldots + a_K(t)$ denote the total number of arrivals during time slot t, $A_k(t) = a_k(1) + \ldots + a_k(t)$ denote the total number of source k arrivals in the interval $[1, t]$, and $A(t) = A_1(t) + \ldots + A_K(t)$ denote the total number of arrivals during the interval $[1, t]$. If $Q(t)$ denotes the queue length at time t, then $Q(t)$ is given by Lindley type recursion $Q(t+1) = [\max(Q(t) + a(t+1), B) - c]^+$; this recursion takes into account the finite buffer of capacity B. For source k, there is an environment variable $X_k(t)$ describing the state of the source. The distribution of arrivals has a Markovian structure given by $P_k(a, j|i) = P(a_k(t) = a, X_k(t) = j|X_k(t-1) = i)$ ($0 \leq a \leq b$, $1 \leq i, j \leq M$) where b and M are finite constants. To apply importance sampling, we draw on large deviations results for Markov additive processes. Define the θ-conjugate process (the exponentially twisted, or tilted, distribution) with parameter θ, as follows. Let $\lambda_k(\theta)$ be the spectral radius of the matrix

$$A_{k,\theta}(i,j) = \sum_{a=0}^{b} e^{\theta a} P_k(a,j|i) \tag{7}$$

and let $h_k(i,\theta)$ be the corresponding eigenvector. (We assume that $A_{k,\theta}(i,j)$ is irreducible so that, by the Perron-Frobenius theorem, $\lambda_k(\theta)$ is real valued and $h_k(i,\theta) > 0$.) Thus

$$\sum_{j=1}^{M} A_{k,\theta}(i,j) h_k(j,\theta) = \lambda_k(\theta) h_k(i,\theta). \tag{8}$$

The twisted distribution for source k is defined by

$$P_{k,\theta}(a,j|i) = P_\theta(a_k(t) = a, X_k(t) = j | X_k(t-1) = i) = \frac{e^{\theta a} P_k(a,j|i) h_k(j,\theta)}{\lambda_k(\theta) h_k(i,\theta)}, \tag{9}$$

which is seen to be a probability distribution by Equations 7 and 8. Each arrival process will be simulated (independently) using the same twisting parameter θ for some as yet unspecified value of θ. Because of the form of $P_{k,\theta}(a,j|i)$, the likelihood ratio after T transitions has a simple form:

$$\begin{aligned} L(\theta,T) &= \textstyle\prod_{k=1}^{K} \prod_{t=1}^{T} \frac{P_k(a_k(t),X_k(t)|X_k(t-1))}{P_{k,\theta}(a_k(t),X_k(t)|X_k(t-1))} \qquad (10)\\ &= \textstyle\prod_{k=1}^{K} \lambda_k(\theta)^T e^{-\theta A_k(T)} \frac{h_k(X_k(0),\theta)}{h_k(X_k(T),\theta)} \\ &= \exp\left\{-\theta A(T) + T \textstyle\sum_{k=1}^{K} \log(\lambda_k(\theta))\right\} H(\theta,T) \end{aligned}$$

where $H(\theta,T) = \prod_{k=1}^{K} h_k(X_k(0),\theta)/h_k(X_k(T),\theta)$.

Suppose the queue is empty at time 0 and the event $\mathcal{R}_\mathcal{B}$ occurs at time T_B, i.e., the queue length first reaches or exceeds level B at time T_B and is non-empty in the interval $[1, T_B]$. Then there are exactly $c \times T_B$ departures in $[1, T_B]$ and, since $Q(T_B) \geq B$, we must have $A(T_B) \geq B + cT_B$. Let $O(T_B)$ be the (nonnegative) "overshoot" defined by $A(T_B) = O(T_B) + B + cT_B$. Thus, by Equation 10

$$L(\theta, T_B) = \exp\left\{-\theta B - \theta O(T_B) + \left[\sum_{k=1}^{K} \log(\lambda_k(\theta)) - \theta c\right] T_B\right\} H(\theta, T_B). \tag{11}$$

By setting $\theta = \theta^*$ where θ^* satisfies

$$\sum_{k=1}^{K} \frac{\log(\lambda_k(\theta^*))}{\theta^*} = c \tag{12}$$

(assuming it exists) results in the further simplification that

$$L(\theta^*, T_B) = e^{-\theta^* B - \theta^* O(T_B)} H(\theta^*, T_B). \tag{13}$$

Because the state space of each source is finite, both $H(\theta^*, T_B)$ and $O(T_B)$ are bounded. Thus, on $\mathcal{R}_{\mathcal{B}}$, there exist constants d_1 and d_2 such that

$$d_1 e^{-\theta^* B} \leq L(\theta^*, T_B) \leq d_2 e^{-\theta^* B}. \tag{14}$$

Furthermore, it can be shown that the queue is unstable when simulated with this value of θ^* and thus $P_{\theta^*}(\mathcal{R}_{\mathcal{B}})$ stays bounded away from zero. Thus, importance sampling with θ^* is asymptotically optimal for estimating $P(\mathcal{R}_{\mathcal{B}})$ and $\lim_{B\to\infty} \log(P(\mathcal{R}_{\mathcal{B}}))/B = -\theta^*$. Note that while the decay rate of $P(\mathcal{R}_{\mathcal{B}})$, $-\theta^*$, is known, the constant in front of it is unknown; in essence, importance sampling is estimating this constant.

In this example, the effective bandwidth of source k is known to be $a_k^*(\theta) = \log(\lambda_k(\theta))/\theta$ and the effective bandwidth of all the sources is $a^*(\theta) = a_1^*(\theta) + \ldots + a_K^*(\theta)$. Note that to find the optimal importance sampling change of measure requires solving Equation 12, $a^*(\theta^*) = c$, a nonlinear equation involving the spectral radiuses of the sources.

A "splitting" technique [9, 31] can be used to efficiently estimate both $E[Y]$ and $E[\alpha]$. The process is simulated (without importance sampling) until it is approximately in steady state. Then whenever the process enters A, two A-cycles are simulated using the same starting conditions; one using importance sampling and one without using importance sampling. These provide samples for the ratio estimate. Also, the A-cycle simulated without importance sampling provides a starting point (with approximately the steady state distribution on A) for the next pair of samples. In particular, let Y_i, α_i, and L_i be the samples of Y, α and the likelihood ratio, respectively, obtained from the i-th pair of $A-$cycles. Then, the estimate of r is given by

$$\hat{r}_n(\theta^*) = \frac{\sum_{i=1}^n L_i Y_i}{\sum_{i=1}^n \alpha_i}. \tag{15}$$

Because the $A-$cycles are not independent and identically distributed, the method of batch means [4] can be used for variance estimation.

3 Fast Simulation of the Shared Buffer Switch

Let us now turn to the simulation of the shared buffer switch. Let $\mathcal{H}$ denote the set of indices of high priority arrival streams and let $\mathcal{L}$ denote the set of indices of low priority arrival streams. Let $a_{\mathcal{H}}^*(\theta) = \sum_{\{k\in\mathcal{H}\}} a_k^*(\theta)$ be the effective bandwidth of the high priority streams and let $a^*_{\mathcal{L}}(\theta) = \sum_{\{k \in \mathcal{L}\}} a^*_k(\theta)$ be the effective bandwidth of the low priority streams. Note that $a^*(\theta) = a^*_{\mathcal{H}}(\theta) + a^*_{\mathcal{L}}(\theta)$.

We again consider efficient estimation of $\mathcal{R}_{\mathcal{B}}$, which is the event that, starting empty, the buffer reaches size $B = B_1 + B_2$ before becoming empty. This event entails first reaching level B_1 and then reaching B. On an $A-$cycle, let J be the (random) number of upcrossings of level B_1; on $\mathcal{R}_{\mathcal{B}}$, $J \geq 1$. Let F be the upcrossing index on which the buffer first reaches B, i.e., if the queue first reaches B after the j-th upcrossing of B_1 but before the $(j+1)$-st upcrossing (assuming it exists), then $F = j$. (If the buffer level B is not hit, define $F = \infty$.) Note that

$$1_{\{\mathcal{R}_{\mathcal{B}}\}} = \sum_{j=1}^{J} 1_{\{F=j\}} = \sum_{j=1}^{\infty} 1_{\{F=j, j \leq J\}} = \sum_{j=1}^{\infty} 1_{\{F=j\}} \tag{16}$$

since if $F = j$ then $j \leq J$ is automatically satisfied. Thus, assuming the $A-$cycle is started in the stationary distribution,

$$P(\mathcal{R}_{\mathcal{B}}) = E[\sum_{j=1}^{\infty} 1_{\{F=j\}}] = \sum_{j=1}^{\infty} E[1_{\{F=j\}}] = \sum_{j=1}^{\infty} P(F=j), \tag{17}$$

i.e., $\sum_{j=1}^{J} 1_{\{F=j\}}$ is an unbiased estimate of $P(\mathcal{R}_{\mathcal{B}})$. Note that, given $J \geq 1$, the event that $J > 1$ may not be rare, i.e., given at least one upcrossing of B_1, there may be many upcrossings of B_1. This complicates the application of importance sampling, because once the process crosses B_1, just pushing the process immediately towards B (using an appropriate change of measure) will not work well (since doing so will not produce accurate estimates of $P(F = j)$ for $j > 1$).

One may also look at this in another way. For each j, the optimal importance sampling distribution for estimating $P(F = j)$ is different. Also there are infinitely many $P(F = j)$'s, so one cannot run a different simulation for estimating each of the $P(F = j)$'s. The technique described below provides a way out of this problem.

We proceed as follows. We first use importance sampling to move the system up to level B_1. When B_1 is crossed we turn off importance sampling until the end of the $A-$cycle is reached. We call this the base path. This base path is used to generate samples of J, the number of upcrossings of B_1. Whenever the base path crosses B_1 (including the first upcrossing), we "split" the base path. On upcrossing j ($j = 1, \ldots, J$) importance sampling is applied on the "j-th split path" to try to bring the system up to level B during this j-th upcrossing of B_1. The initial conditions of the j-th split path are the same as that of the base path just after it crosses B_1 for the j-th time. The j-th split path is simulated until either B is hit or the queue length drops below B_1, whichever comes first. This splitting of the base path is illustrated in Figure **??**.

Let L_0 denote the likelihood ratio during the buildup from 0 to B_1 and L_j denote the likelihood ratio during the j-th split path. Define

$$Z = \sum_{j=1}^{J} L_0 L_j 1_{\{\mathcal{R}_j \cap \mathcal{C}_j\}} = \sum_{j=1}^{\infty} L_0 L_j 1_{\{\mathcal{R}_j \cap \mathcal{C}_j\}} 1_{\{j \leq J\}} = \sum_{j=1}^{\infty} L_0 L_j 1_{\{\mathcal{R}_j \cap \mathcal{C}_j\}} \tag{18}$$

where $\mathcal{R}_j$ is the event that the j-th split path hits B before downcrossing B_1 and $\mathcal{C}_j$ is the event that the base cycle has not hit B before the start of the j-th split path, i.e., $\mathcal{C}_j = \{F \geq j\} \cap \{j \leq J\}$. With this notation, we associate the random variables F and J only with the base path, and not the split paths. Then $L_0 L_j 1_{\{\mathcal{R}_j \cap \mathcal{C}_j\}}$ is an unbiased estimate of $P(F = j)$. This is because, the part of base path from the beginning of the $A-$ cycle until the j-th split, combined with the j-th split path, may be viewed as a new base path (that also has $J \geq j$) that has been generated using the following change of measure: at the beginning of the $A-$cycle, push the the process towards B_1 and then wait until B_1 is upcrossed $j-1$ times more before pushing it towards B. Denote this change of measure by p'_j. The random variable $1_{\{\mathcal{R}_j \cap \mathcal{C}_j\}}$ defined on the original $A-$cycle with split paths, is equivalent to the random variable $1_{\{F=j\}}$ defined on this new base path. Then we get unbiasedness from the fact that

$$E_{p'_j}[L_0 L_j 1_{\{F=j\}}] = E[1_{\{F=j\}}] = P(F = j). \tag{19}$$

Thus, Z is an unbiased estimate of $P(\mathcal{R}_{\mathcal{B}})$. ¿From Equation 18 we see that we need not generate split paths off the base path if the base path has already hit level B.

We now turn to the specifics of how importance sampling is applied. On the build-up from 0 to B_1, the queue behaves identically to the queue analyzed in Section 2. Thus an asymptotically optimal estimate of $P(J \geq 1)$ (hitting B_1) is obtained by exponential twisting of both the high and low priority arrival streams with parameter θ_1^* defined to be the solution of

$$a^*(\theta_1^*) = a^*_{\mathcal{H}}(\theta_1^*) + a^*_{\mathcal{L}}(\theta_1^*) = c, \tag{20}$$

assuming such a θ_1^* exists. (Assuming it exists, $\theta_1^* > 0$ since the queue is stable.) In particular, there exist constants d_1 and d_2 such that

$$d_1 e^{-\theta_1^* B_1} \leq L_0 \leq d_2 e^{-\theta_1^* B_1}. \tag{21}$$

Similarly, during the j-th upcrossing of B_1, the system behaves like a queue with only high priority arrivals building up from some level O_j to B_2 (without going below 0) where O_j is the overshoot on the j-th upcrossing of B_1. Since the sources are finite, there exists some constant d such that $O_j \leq d$. Thus exponential twisting of the high priority streams with parameter θ_2^* defined to be the solution (assuming it exists) of

$$a_{\mathcal{H}}^*(\theta_2^*) = c \tag{22}$$

should be effective for large B_2. (Again, assuming it exists, $\theta_2^* > 0$ since the queue is stable.) During this phase, low priority streams are simulated from their given distributions (and therefore do not contribute to the likelihood ratio). To make this more precise, if the time axis is renumbered so that the j-th upcrossing of B_1 occurs at time 0 and level $B_1 + B_2$ is first hit on this up-crossing at time T_B, then the number of high priority arrivals in $[1, T_B]$ is between $B_2 - d + T_B c$ and $B_2 + T_B c + d$, corresponding to the bounding cases that (i) $O_j = d$ and the overshoot upon hitting B is zero, and (ii) $O_j = 0$ and the overshoot upon hitting B is d (its maximum), respectively. Thus when $\theta = \theta_2^*$, there exist constants d_3 and d_4 such that on $\mathcal{R}_j \cap \mathcal{C}_j$

$$d_3 e^{-\theta_2^* B_2} \leq L_j \leq d_4 e^{-\theta_2^* B_2}. \tag{23}$$

In addition, the queue is unstable when simulated with at θ_2^*. Therefore, by Equation 18, the importance sampling estimate (from a single $A-$cycle) satisfies

$$1_{\{\mathcal{R}_1 \cap \mathcal{C}_1\}} d_1 d_3 e^{-\theta_1^* B_1 - \theta_2^* B_2} \leq Z \leq d_2 d_4 J e^{-\theta_1^* B_1 - \theta_2^* B_2}. \tag{24}$$

The event $\mathcal{R}_1 \cap \mathcal{C}_1$ is simply the event that the base path hits B_1 and the first split path hits B. Thus

$$P(\mathcal{R}_{\mathcal{B}}) \geq d_1 d_3 P_{\theta_1^*, \theta_2^*}(\mathcal{R}_1 \cap \mathcal{C}_1) e^{-\theta_1^* B_1 - \theta_2^* B_2}. \tag{25}$$

Note that $\mathcal{C}_1 = \{J \geq 1\}$, so $P_{\theta_1^*, \theta_2^*}(\mathcal{R}_1 \cap \mathcal{C}_1) = P_{\theta_2^*}(\mathcal{R}_1 | J \geq 1) P_{\theta_1^*}(J \geq 1)$. Under importance sampling $P_{\theta_2^*}(\mathcal{R}_1 | J \geq 1)$ and $P_{\theta_1^*}(J \geq 1)$ are both bounded away from zero, since, as described above, they both represent the probability that an appropriate unstable queue reaches some queue length before returning to zero. (Note that the distribution of J depends only on θ_1^*, since the base path turns importance sampling off after B_1 is crossed for the first time.) Similarly, by Equation 24,

$$E_{\theta_1^*, \theta_2^*}[Z^2] \leq [d_2 d_4]^2 E_{\theta_1^*}[J^2] e^{-2\theta_1^* B_1 - 2\theta_2^* B_2}. \tag{26}$$

Thus if, $E_{\theta_1^*}[J^2] < \infty$, the importance sampling estimate of $P(\mathcal{R}_{\mathcal{B}})$ has bounded relative error (and is thus asymptotically optimal). It is intuitively clear that this must be so, since $(J-1)$ is the number of upcrossings of level B_1 of a queue with negative drift started just above level B_1. A proof that $E_{\theta_1^*}[J^2] < \infty$ is given in Appendix A. To formalize this result, we state the following theorem.

Theorem 1 *Let $B_1 = fB$ and $B_2 = (1-f)B$ for some constant f, $0 < f < 1$. If there exist constants $\theta_1^* > 0$ and $\theta_2^* > 0$ such that $a_{\mathcal{H}}^*(\theta_1^*) + a_{\mathcal{L}}^*(\theta_1^*) = c$ and $a_{\mathcal{H}}^*(\theta_2^*) = c$, then*

$$\lim_{B\to\infty} \frac{\log(P(\mathcal{R}_{\mathcal{B}}))}{B} = -\theta_1^* f - \theta_2^*(1-f) \tag{27}$$

$$\lim_{B\to\infty} \frac{\log(E_{\theta_1^*,\theta_2^*}[Z^2])}{B} = -2\theta_1^* f - 2\theta_2^*(1-f) \tag{28}$$

i.e., the importance sampling estimate is asymptotically optimal and has bounded relative error.

4 Experimental Results

In this section we will present experimental results using the importance sampling algorithm described in the previous section. As is Section 2 define an A-cycle to begin whenever the process enters the set of states where the buffer is empty. Also, if α denotes the number of high priority arrivals in a typical A-cycle and Y denotes the number of high priority packets lost in an A-cycle, then the long run fraction of high priority packets lost is given by Equation 6.

As mentioned in Section 2, asymptotically optimal changes of measure for estimating $P(\mathcal{R}_{\mathcal{B}})$ (where now $B = B_1 + B_2$) also work well for estimating $E[Y]$. However the simulation procedure differs slightly. Similar to Equation 17, define

$$W = \sum_{j=1}^{J} L_0 L_j N_j 1_{\{\mathcal{R}_j \cap \mathcal{C}_j\}} \tag{29}$$

where N_j is the number of high priority packets lost in the j-th split path until it hits the buffer empty state. Then W is an unbiased estimate (from a single $A-$cycle) of $E[Y]$. Unlike the simulation for $P(\mathcal{R}_{\mathcal{B}})$ (where we stop simulating the j-th split path once it hits level B or downcrosses level B_1) if the j-th split path hits B before it downcrosses level B_1, then we have to simulate it until the buffer empty state (in order to get a sample of N_j).

We adopt the following simulation procedure. We first run enough A-cycles so that the process (approximately) reaches steady state. Then we start collecting samples of W and α using the splitting technique referred to at the end of Section 2, i.e., whenever the process enters A we generate two parallel cycles, one with importance sampling and the other without importance sampling. The one with importance sampling is used to generate samples of W and the one without importance sampling is used to generate samples of α as well as to get a starting point for the next pair of A-cycles. ¿From n such sets of parallel cycles, samples $W_1, \ldots, W_n$ and $\alpha_1, \ldots, \alpha_n$ of W and α, respectively, are obtained. Then r is estimated by $\hat{r}_n(\theta_1^*, \theta_2^*) = \sum_{i=1}^n W_i / \sum_{i=1}^n \alpha_i$. However, since the successive W_i's (and α_i's) are not independent, the method of batch means is used to construct confidence intervals. We will now briefly review this procedure.

Divide $W_1, \ldots, W_n$ and $\alpha_1, \ldots, \alpha_n$ into b batches of size $k \equiv n/b$. (Assume that n/b is an integer.) Let $\delta_i = \sum_{j=(i-1)k+1}^{ik} W_j/k$ and $\gamma_i = \sum_{j=(i-1)k+1}^{ik} \alpha_j/k$, i.e., these are the sample means corresponding to the ith batch. The method of batch means works on the assumption that for k sufficiently large, the successive δ_j's and γ_j's are (approximately) normally distributed and independent. Hence if we estimate r by $\hat{r} = \sum_{i=1}^b \delta_i / \sum_{i=1}^b \gamma_i$ (this is the same estimator as when we did not use batch means), then, if b is also large, $\sqrt{b}(\hat{r} - r) \approx N(0, \sigma^2)$ where

$$\sigma^2 = \frac{\mathrm{Var}[\delta_j] + 2r\mathrm{Cov}[\delta_j, \gamma_j] + r^2 \mathrm{Var}[\gamma_j]}{E[\gamma_j]}. \tag{30}$$

Then an approximate 99% confidence interval can be constructed as $(\hat{r} - 2.56\sqrt{\sigma^2/b}, \hat{r} + 2.56\sqrt{\sigma^2/b})$. The relative error RE is thus defined to be $2.56\sqrt{(\sigma^2/b)}/r$. In practice, since σ^2 is unknown, it must be estimated, which can be done using the standard estimators of all the quantities in Equation 30.

We now consider two examples. In the first example there are four input sources, three of which are high priority and one of which is low priority. Each source is a two state Markov modulated process, with a deterministic number of arrivals in each state; in state 1 there is no arrival and in state 2 there is exactly 1 arrival. Let $p_{ij}^{(k)}$ denote the the transition matrix of the state of the kth source, i.e., $p_{ij}^{(k)} = P(X_k(t) = j | X_k(t-1) = i)$. For $1 \le k \le 3$, $p_{11}^{(k)} = 0.7$ and $p_{22}^{(k)} = 0.5$; for $k = 4$, $p_{11}^{(k)} = 0.9$ and $p_{22}^{(k)} = 0.5$. (This notation can be easily mapped to the notation of Section 2.) Note that the individual input streams are positively correlated, i.e., $p_{11}^{(k)} + p_{22}^{(k)} > 1$ for all k. The total high priority and low priority arrival rate can be computed to be 1.125 and 0.167 respectively. We choose the switch speed $c = 2$. For such two state Markov processes, the appropriate Eigenvalues are known in closed form (see, e.g., [9, 10]). Using Equation 20 and Equation 22, the θ_1^* and θ_2^* were computed to be 0.75 and 2.00, respectively.

We ran experiments for several values of B_1 and B_2; we fixed $f = 0.5$ and varied B. Note that the Markov chain model for this example will have $8(B+1)$ states and it is not very difficult to solve this particular model numerically (for moderate values of B). We use this model mainly as a means of comparing the effectiveness of our importance sampling algorithm to that of standard simulation; for moderate values of B, the loss rates can be accurately estimated within a reasonable amount of time using standard simulation. In each experiment (using importance sampling) we first simulated 300 consecutive A-cycles so that the process (approximately) reaches steady state. Then we simulated for 60,000 A-cycle pairs. We used $b=$ 2,000 batches with 30 A-cycle pairs in each batch. (The variance estimate was fairly insensitive to the batch size for the same total number of A-cycles). For the sake of completeness, we also estimated the loss probabilities (r_1) of the low priority packets using the same simulation runs; we used the A-cycles for the estimation of the denominator (in Equation 6) to get samples of the total number of low priority arrivals in an A-cycle and we used the base sample path of the cycle used for the estimation of the numerator, to get samples of the number of low priority arrivals lost during an A-cycle.

Results for these experiments are presented in Table 1. Note how the RE using importance sampling remains almost constant as B becomes larger. For comparison purposes we also ran standard simulations for the same CPU time that was used for the importance sampling case. By standard simulation we mean that (after discarding the first 300 $A-$cycles) at each entrance to the set A, instead of simulating a pair of A-cycles, we only simulate the A-cycle with no importance sampling. In addition to getting samples of the total number of arrivals of each priority level from this A-cycle, we also obtain samples of the number lost. For a given number of samples, i.e, pair of numerator/denominator $A-$cycle samples, importance sampling requires twice the number of $A-$cycle simulations as compared to standard simulation. Also, an $A-$cycle with importance sampling is more costly to simulate due to the split paths and the computation of the likelihood ratios. However, as we can see in Table 1, these inefficiencies in the importance sampling method are insignificant as compared to the speedup gained due to the variance reduction. Note that with standard simulation, in many cases, we do not even get any samples of buffer overflow in an $A-$cycle. (In the tables, such cases are indicated by a $\pm$?. The superscript $*$ in some of the table entries means that the estimate had not yet stabilized.)

To compute speedup, we need to compute the ratio of the CPU time required by standard simulation to achieve a given level of accuracy to the CPU time required by importance sampling to achieve the same level of accuracy. Due to the large amount of CPU time it took for the standard simulation estimates to stabilize, we were not able to compute the speedup for all cases, although this was possible for $B=6$ and $B=8$. For $B=6$, for the same amount of CPU time, the RE of the standard simulation estimate (for r) is about 6.7 (=20.2/3) times larger than that of the importance sampling estimate. Since the RE decreases at rate $1/\sqrt{\text{sample size}}$, the standard simulation would have to be run about 45

($\approx 6.7^2$) times longer in order to achieve the same RE as using importance sampling. Thus the $B = 6$ speedup is about 45. For $B = 8$, the variance estimate using standard simulation had not yet stabilized. For this case, we continued the standard simulation run until we obtained the same RE ($\pm 3.2\%$) as that obtained in Table 1 using importance sampling. The CPU time required to achieve this same accuracy was about 450 times greater than that measured for importance sampling, i.e., the $B = 8$ speedup is about 450. (The standard simulation estimate of r was $3.13 \times 10^{-5} \pm 3.2\%$ compared to $3.11 \times 10^{-5} \pm 3.2\%$ using importance sampling.) As r decreases, the speedup using importance sampling increases, although the speedup becomes more difficult to compute.

Next we consider a larger example with 16 two-state Markov modulated input sources (of the same type as in the previous example). The p_{11}'s and p_{22}'s of the 16 sources were

$$(0.7,0.6,\ 0.6,0.5,0.8,0.8,0.8,0.8,0.8,0.7,0.7,0.7,0.9,0.9,0.9,0.9)$$

and

$$(0.5,0.8,0.6,0.9,0.3,0.6,0.6,0.4,0.8,0.5,0.9,0.6,0.7,0.8,0.6,0.5),$$

respectively. The first twelve sources are considered to be of high priority with a total arrival rate of 5.57 and the last four sources are considered to be of low priority with a total arrival rate of 0.95. The switch speed c was chosen to be 8. Using Equation 20 and Equations 22 the θ_1^* and θ_2^* were computed as 0.34375 and 1.0625, respectively.

We again estimated loss probabilities for several values of B and a fixed f. Note that the Markov chain model for this example will have approximately $2^{16}(B+1)$ states and hence it is very difficult to compute the loss probabilities numerically. Therefore, simulation using importance sampling is very useful. The initial transient deletion period was again 300 A-cycles. The simulation was run for 150,000 A-cycle pairs, i.e., 5,000 batches with 30 A-cycle pairs in each batch. Notice again how the RE using importance sampling remains bounded as $B \to \infty$ whereas with standard simulation, in all cases, we did not observe any high priority packet losses. With importance sampling, loss rates as low as 10^{-23} were estimated in a reasonable amount of time with a high degree of accuracy.

5 Summary

This paper has considered the application of importance sampling to estimating the packet loss rate in an ATM communications switch having a shared buffer and two priority classes. When the buffer contents reach a threshold level, B_1, low priority packets are rejected. When the buffer contents reaches its capacity, $B_1 + B_2$, high priority packets are rejected. The object of the analysis is to estimate the loss rate of high priority packets. An asymptotically optimal importance sampling approach was described. Experiments using this approach showed that many orders of magnitude speedup, as compared to standard simulation, can be obtained. The speedup increases as the packet loss rate decreases.

This importance sampling approach is unusual in several respects:

1. The rare event of interest occurs because of the occurrence of two different events, each of which is rare. Importance sampling has to be done so as to make each of these rare events happen, one after the other. Since the best change of measure is different for each of these rare events, the simulation program has to shift from one importance sampling strategy to another at appropriate instances in time.

2. The situation is further complicated by the fact that the second rare event may not happen immediately after the first rare event. As a result, a more complicated "splitting" procedure needs to be introduced so as to get an efficient overall simulation procedure.

Based on the analysis, it is clear that this approach can be extended to the case when there are more than two priority levels, e.g., three priority levels with an admission policy determined by three threshold levels, $B_1, B_1 + B_2$, and $B_1 + B_2 + B_3$. There would be three different values of θ, θ_1^*, θ_2^* and θ_3^*, with which to do importance sampling. Exponential twisting of the high, medium and low priority streams with parameter θ_1^* satisfying $a_{\mathcal{H}}^*(\theta_1^*) + a_{\mathcal{M}}^*(\theta_1^*) + a_{\mathcal{L}}^*(\theta_1^*) = c$ is done initially at the start of an A-cycle until B_1 is reached. Then exponential twisting of the high and medium priority streams with parameter θ_2^* satisfying $a_{\mathcal{H}}^*(\theta_2^*) + a_{\mathcal{M}}^*(\theta_2^*) = c$ is done until $B_1 + B_2$ is reached. Finally, exponential twisting of the high priority streams with parameter θ_3^* satisfying $a_{\mathcal{H}}^*(\theta_3^*) = c$ is done until $B_1 + B_2 + B_3$ is reached. In addition, a multi-level splitting procedure needs to be applied; the j-th split path obtained on the j-th upcrossing of B_1 also needs to be split upon upcrossings of $B_1 + B_2$. This results in a rather complicated simulation algorithm.

Other extensions are not so clear. For example, suppose when the buffer reaches level B_1, the switch displaces low priority packets already in the buffer with high priority arriving packets. Now the change of measure required to move the buffer from level B_1 to $B_1 + B_2$ is not obvious.

This paper thus illustrates both the potential and limitations of importance sampling. In certain applications, there is enough structure so that highly effective importance sampling schemes can be devised. In such cases, importance sampling results in spectacular improvements over standard simulation. However, while some applications of genuine interest, such as the one considered here, can be handled optimally, the class of such applications is rather narrow. Furthermore, small changes in the problem structure can easily lead to models for which the optimal, or even a good, importance sampling change of measure is unknown.

A Appendix A

In this appendix we prove $E_{\theta_1^*}[J^2] < \infty$. Our approach is based on a sequence of stochastic comparisons. We note that $J-1$ is bounded above by the last time that the queue is below B_1 in an A-cycle when started from the level B_1+d (where d is the bound for the overshoot when level B_1 is upcrossed). Denote this last time by τ_1. Now we consider an auxiliary queue in which the lower priority arrival streams are allowed to be stored in the buffer even when the buffer reaches the level B_1. Let τ_2 be the last time that the auxiliary queue is below B_1 in an A-cycle when started from B_1+d. Since there are more arrivals in the auxiliary queue than the original queue, it then follows from a standard sample path argument that τ_2 is stochastically larger than or equal to τ_1, i.e., $\tau_2 \geq_{st} \tau_1$. Now we remove both boundaries at levels B and 0 of the auxiliary queue and consider the corresponding random walk started from the same level B_1+d. Let τ_3 be the last time that the random walk is below level B_1. Clearly, $\tau_3 \geq \tau_2$ since there are no losses in the random walk and there are sample paths that the random walk might reach level 0 and then bounce back to level B_1. Now we will show that the distribution of τ_3 is bounded by an exponential tail and hence all the moments exist. Via the stochastic comparisons made above, all the moments of $J-1$ exist and $E_{\theta_1^*}[J^2] < \infty$.

To show that the tail distribution of τ_3 is bounded exponentially, let $A(t)$ denote the total number of arrivals in the interval $[1,t]$. Thus, the level of the random walk at time t is $A(t)-ct+B_1+d$. From the definition for the last time, $P(\tau_3 > s) = P(\max_{t>s}[A(t)-ct+d] \geq 0)$. Note that for all $\epsilon > 0$ and $\theta > 0$

$$P(\max_{t>s}[A(t)-ct+d] \geq 0) = P(\max_{t>s}[A(t)-ct+d+\epsilon t-\epsilon t] \geq 0) \tag{31}$$

$$\leq P(\max_{t>s}[A(t)-(c-\epsilon)t+d] \geq \epsilon s) \tag{32}$$

$$\leq P(\max_{t>0}[A(t)-(c-\epsilon)t+d] \geq \epsilon s) \tag{33}$$

$$= P\left(\max_{t>0}[\exp(\theta(A(t)-(c-\epsilon)t+d))] \geq \exp(\theta\epsilon s)\right). \tag{34}$$

Applying Markov's inequality and replacing the maximum by the sum yields

$$P(\tau_3 > s) \leq e^{-\theta\epsilon s} e^{\theta d} \sum_{t>0} E e^{\theta(A(t)-(c-\epsilon)t)}. \tag{35}$$

¿From Equation 10, it follows that

$$E[e^{\theta(A(t)-(c-\epsilon)t)}] = E_\theta[L(\theta,t)e^{\theta(A(t)-(c-\epsilon)t)}] \tag{36}$$

$$= E_\theta[H(\theta,t)e^{(a^*(\theta)-(c-\epsilon))t}] \tag{37}$$

Since the state space of each source is finite, $H(\theta,t)$ is bounded. Thus, there exists a constant $d_2(\theta) < \infty$ for $\theta > 0$ such that $H(\theta,t) \leq d_2(\theta)$. Also, since $\theta_1^* > 0$ is the solution of $a^*(\theta) = c$, for any ϵ sufficiently small, there exists a $0 < \theta' < \theta_1^*$ such that $a^*(\theta') \leq c - 2\epsilon$. Thus,

$$E[e^{\theta'(A(t)-(c-\epsilon)t)}] \leq d_2(\theta')e^{-\theta'\epsilon t}. \tag{38}$$

In conjunction with (35), one has

$$P(\tau_3 > s) \leq \phi(\theta')e^{-\theta'\epsilon s}, \tag{39}$$

where

$$\phi(\theta') = d_2(\theta')e^{\theta'(d-\epsilon)}(1 - e^{-\theta'\epsilon})^{-1}.\square \tag{40}$$

B	f	Quantity Estimated	Importance Sampling Estimate	Standard Simulation Estimate
6	0.5	r	$4.34 \times 10^{-4} \pm 3.0\%$	$4.72 \times 10^{-4} \pm 20.2\%$
		r_1	$3.72 \times 10^{-2} \pm 4.7\%$	$3.92 \times 10^{-2} \pm 6.1\%$
8	0.5	r	$3.11 \times 10^{-5} \pm 3.2\%$	$4.19 \times 10^{-5} \pm 68.4\%^*$
		r_1	$1.59 \times 10^{-2} \pm 4.7\%$	$1.73 \times 10^{-2} \pm 8.9\%$
20	0.5	r	$5.47 \times 10^{-12} \pm 3.8\%$	$0.00\pm?$
		r_1	$9.53 \times 10^{-5} \pm 5.1\%$	$12.98 \times 10^{-5} \pm 70.8\%^*$
30	0.5	r	$1.24 \times 10^{-17} \pm 4.4\%$	$0.00\pm?$
		r_1	$1.29 \times 10^{-6} \pm 5.0\%$	$0.00\pm?$
40	0.5	r	$3.02 \times 10^{-23} \pm 4.8\%$	$0.00\pm?$
		r_1	$1.79 \times 10^{-8} \pm 5.2\%$	$0.00\pm?$

Table 1: Estimates of steady-state loss probabilities in the small example. Importance sampling and standard simulation were run for the same amount of CPU time.

B	f	Quantity Estimated	Importance Sampling Estimate	Standard Simulation Estimate
24	1/3	r	$1.13 \times 10^{-7} \pm 4.4\%$	$0.00 \pm ?$
		r_1	$1.15 \times 10^{-2} \pm 2.2\%$	$1.17 \times 10^{-2} \pm 3.9\%$
30	1/3	r	$8.63 \times 10^{-10} \pm 4.8\%$	$0.00 \pm ?$
		r_1	$5.74 \times 10^{-3} \pm 2.2\%$	$5.96 \times 10^{-3} \pm 5.1\%$
45	1/3	r	$4.32 \times 10^{-15} \pm 7.6\%$	$0.00 \pm ?$
		r_1	$1.04 \times 10^{-3} \pm 2.2\%$	$1.06 \times 10^{-3} \pm 10.0\%$
60	1/3	r	$2.15 \times 10^{-20} \pm 4.9\%$	$0.00 \pm ?$
		r_1	$1.94 \times 10^{-4} \pm 2.2\%$	$2.01 \times 10^{-4} \pm 19.8\%$

Table 2: Estimates of steady-state loss probabilities in the large example. Importance sampling and standard simulation were run for the same amount of CPU time.

References

[1] Asmussen, S. 1985. Conjugate processes and the simulation of ruin problems. *Stochastic Processes and their Applications* 20: 213-229.

[2] Asmussen, S. 1987 . *Applied Probability and Queues.* New York, NY: J. Wiley & Sons, Inc.

[3] Asmussen, S. 1989. Risk theory in a Markovian environment. *Scand. Actuarial J.*, 69-100.

[4] Bratley, P., B.L. Fox, and L.E. Schrage. 1987 *A Guide to Simulation*, Second Edition. New York, NY: Springer Verlag.

[5] Breiman, L. 1968. *Probability.* Reading, MA: Addison-Wesley.

[6] Bucklew, J. 1990. *Large Deviation Techniques in Decision, Simulation and Estimation.* New York, NY: J. Wiley & Sons, Inc.

[7] Bucklew, J.A., P. Ney, and J.S. Sadowsky. 1990. Monte Carlo simulation and large deviations theory for uniformly recurrent Markov chains. *J. Appl. Prob.* 27: 44-99.

[8] Chang, C.S. 1992. Stability, queue length and delay, Part II: Stochastic queueing networks. IBM Research Report RC 17709, Yorktown Heights, New York. Part of the report is published in *Proceedings of the IEEE CDC'92 Conference*, 1005-1010, IEEE Computer Society Press, Tucson, Arizona, 1992.

[9] Chang, C.S., P. Heidelberger, S. Juneja, and P. Shahabuddin. 1992. Effective bandwidth and fast simulation of ATM intree networks. IBM Research Report RC 18586, Yorktown Heights, New York. To appear in *Proceedings of the Performance '93 Conference.*

[10] Chang, C.S., P. Heidelberger, S. Juneja, and P. Shahabuddin. 1993. The application of effective bandwidth to fast simulation of communication networks. IBM Research Report RC 18877, Yorktown Heights, New York.

[11] Cogburn, R. 1975. A uniform theory for sums of Markov chain transition probabilities. *The Annals of Probability* 3: 191-214.

[12] Crane, M.A., and D.L. Iglehart. 1975. Simulating stable stochastic systems III: regenerative processes and discrete event simulation. *Operations Research* 23: 33-45.

[13] Ellis, R. 1984. Large deviations for a general class of random vectors. *Annals of Probability* 12: 1-12.

[14] Elwalid, A.I. and D. Mitra. 1993. Effective bandwidth of general Markovian traffic sources and admission control of high speed networks. *IEEE/ACM Transactions on Networking* 1: 329 - 343.

[15] Elwalid, A.I. and D. Mitra. 1991. Fluid models for the analysis and design of statistical multiplexing with loss priorities on multiple classes of bursty traffic. *Proceedings of the IEEE Global Telecommunications Conference*, 415-425, IEEE Press.

[16] Elwalid, A.I. and D. Mitra. 1993. Statistical multiplexing with loss priorities in rate-based congestion control of high speed networks. To appear in *IEEE Transaction on Communications.*

[17] Frater, M.R., and B.D.O. Anderson. 1989. Fast estimation of the statistics of excessive backlogs in tandem networks of queues. *Australian Telecommunications Research* 23: 49-55.

[18] Frater, M.R., R.R. Bitmead, R.A. Kennedy, and B.D.O. Anderson. 1989. Rare events and reverse time models. In *Proceedings of the 28th Conference on Decision and Control*, 1180-1183, IEEE Press.

[19] Frater, M.R., T.M. Lenon, and B.D.O. Anderson. 1991. Optimally efficient estimation of the statistics of rare events in queueing networks. *IEEE Transactions on Automatic Control* 36: 1395-1405.

[20] Gärtner, J. 1977. On large deviations from invariant measure. *Theory Probab. Appl.* 22: 24-39.

[21] Gibbens, R.J., and P.J. Hunt. 1991. Effective bandwidths for the multi-type UAS channel. *Queueing Systems* 9: 17-28.

[22] Glasserman, P. and S.G. Kou. 1993. Overflow probabilities in Jackson networks. To appear in *Proceedings of the 32nd IEEE Conference on Decision and Control.*

[23] Glynn, P.W., and D.L. Iglehart. 1989. Importance sampling for stochastic simulations. *Management Science* 35: 1367-1392.

[24] Guérin, R., H. Ahmadi and M. Naghshineh. 1991. Equivalent capacity and its application to bandwidth allocation in high-speed networks. *IEEE J. Select. Areas Commun.* 9: 968-981.

[25] Hammersley, J.M., and D.C. Handscomb, 1964. *Monte Carlo Methods.* London: Methuen and Co., Ltd.

[26] Heidelberger, P. 1993. Fast simulation of rare events in queueing and reliability models. IBM Research Report RC 19028, Yorktown Heights, New York. Preliminary version appears in *Performance Evaluation of Computer and Communications Systems*, Springer Verlag, Lecture Notes in Computer Science, Volume 729, 165-202.

[27] Kelly, F.P. 1991. Effective bandwidths at multi-class queues. *Queueing Systems* 9: 5-16.

[28] Kesidis, G., and J. Walrand. 1993. Quick simulation of ATM buffers with on-off multiclass Markov fluid sources. To appear in *ACM Transactions on Modeling and Computer Simulation.*

[29] Lehtonen, T., and H. Nyrhinen. 1992. Simulating level-crossing probabilities by importance sampling. *Advances in Applied Probability* 24: 858-874.

[30] Lehtonen, T., and H. Nyrhinen. 1992. On asymptotically efficient simulation of ruin probabilities in a Markovian environment. *Scand. Actuarial. J.* 60-75.

[31] Nicola, V.F., P. Shahabuddin, P. Heidelberger and P.W. Glynn. 1993. Fast simulation of steady-state availability in non-Markovian highly dependable systems. In *Proceedings of the Twenty-Third International Symposium on Fault-Tolerant Computing*, 38-47, IEEE Computer Society Press.

[32] Nicola, V.F., P. Shahabuddin, and P. Heidelberger. 1993. Techniques for fast simulation of highly dependable systems. IBM Research Report RC 18956, Yorktown Heights, New York.

[33] Parekh, S., and J. Walrand. 1989. A quick simulation method for excessive backlogs in networks of queues. *IEEE Transactions on Automatic Control* 34: 54-56.

[34] Sadowsky, J.S. 1991. Large deviations and efficient simulation of excessive backlogs in a GI/G/m queue. *IEEE Transactions on Automatic Control* 36: 1383-1394.

[35] Sadowsky, J.S. 1993. On the optimality and stability of exponential twisting in Monte Carlo estimation. *IEEE Transactions on Information Theory* 39: 119-128.

[36] Sadowsky, J.S., and J.A. Bucklew. 1990. On large deviations theory and asymptotically efficient Monte Carlo estimation. *IEEE Transactions on Information Theory* 36: 579-588.

[37] Siegmund, D. 1976. Importance sampling in the Monte Carlo study of sequential tests. *The Annals of Statistics* 4: 673-684.

[38] Smith, W.L. 1955. Regenerative stochastic processes. *Proc. Roy. Soc. Ser. A.* 232: 6-31.

[39] Whitt, W. 1993. Tail probability with statistical multiplexing and effective bandwidths in multi-class queues. To appear in *Telecommunication Systems.*