Importance Sampling for the Estimation of Buffer Overflow Probabilities via Trace-Driven Simulations

Ioannis Ch. Paschalidis and Spyridon Vassilaras

***Abstract*—We develop an importance sampling technique that can be used to speed up the simulation of a model of a buffered communication multiplexer fed by a large number of independent sources. The sources generate traffic according to a periodic function with a random phase. This traffic model accommodates a wide range of situations of practical interest, including ON–OFF periodic traffic models and sequences of bit rates generated by actual Variable Bit Rate sources, such as MPEG video compressors. The simulation seeks to obtain estimates for the buffer overflow probability that in most cases of interest is very small. We use a large deviations result to devise the change of measure used in the importance sampling technique and demonstrate through numerical results that this change of measure leads to a dramatic reduction in the required simulation time over direct Monte Carlo simulation. Possible practical applications include short-term network resource planning and even real-time Call Admission Control.**

***Index Terms*—Importance sampling, large deviations, MPEG traces, simulation, statistical multiplexing, variance reduction.**

I. Introduction

REAL-TIME digital telecommunication services (such as Internet telephony, teleconferencing, and video-on-demand) are very sensitive to packet losses and delays and require from the network stringent *Quality of Service (QoS)* guarantees. End users, however, can tolerate the loss of a very small fraction of the transmitted packets. This observation led researchers to propose the provisioning of statistical QoS guarantees (see [3] and references therein), which allow for a more efficient utilization of the network resources than worst case-based QoS provisioning.

More specifically, the objective becomes to operate the network in a regime where congestion phenomena, that lead to packet losses or delays, occur with very small probabilities (e.g., on the order of 10^{-6} to 10^{-8}). Packet drops due to erroneous transmission occur with an even smaller probability in modern high-speed telecommunication networks. Estimating the probabilities of congestion phenomena, such as packet losses due to buffer overflows, is a very hard problem, especially in view of complicated traffic models. Consequently, researchers have relied on analytical asymptotic results—using mainly *Large Deviations (LD)* theory (see [4] for a survey, [3], [5], and references therein)—and simulation.

One of the limitations of analytical (e.g., LD) techniques is that they require the knowledge of a traffic model and its detailed statistics. In practice, such a model may be unavailable or hard to obtain (e.g., by fitting actual traffic realizations to a model and estimating its parameters). Hence, simulation emerges as the only alternative. Even if a traffic model is available and its analysis feasible, simulation is commonly used to validate the model and the analytical results.

Obtaining accurate estimates of very small congestion probabilities via Monte Carlo simulation can require huge sample sizes, and thus, prohibitively long execution times. *Importance sampling* (see [6] and [7]) is a powerful technique that can be used to drastically reduce the sample size (and execution time) required to achieve a given level of estimation accuracy. It has been effectively applied to speed up rare event simulations in a variety of queueing models (see [8] for a survey, [9]–[15] and references therein). Simulation using importance sampling is often referred to as *quick simulation*. Essentially, importance sampling is based on the notion of *changing* (or *biasing* or *twisting*, as it is commonly called) the underlying probability measure in such a way that the rare events occur much more frequently. To correct for this biasing, the results are weighted in a way that yields a statistically unbiased estimator. Identifying the appropriate change of measure, though, is a complex problem-specific task which can depend heavily on the stochastic processes involved and requires solving an LD problem. It is its problem-specific nature and lack of generality that make importance sampling hard to use effectively.

In this paper, we develop an important sampling approach for significantly speeding up a very popular simulation scheme. Under this scheme, a buffered multiplexer is fed by a large number of independent traffic sources and drained at a constant rate. Each arrival process is simulated by repeating an existing finite sequence of bits per unit of time (e.g., bits per frame of an MPEG coded movie) over and over again. The sources are randomly synchronized by starting transmission to the buffer from a random point in their associated sequences. When some source reaches the end of its sequence before the end of the simulation, it continues from the start of the sequence. The simulation seeks to obtain an estimate of the overflow probability at the multiplexer's buffer by averaging the fraction of time that the buffer

Manuscript received June 18, 2001; revised October 4, 2002; approved by IEEE/ACM TRANSACTIONS ON NETWORKING Editor J. Liebeherr. This work was supported in part by the National Science Foundation (NSF) under CAREER Award ANI-9983221, Grants DMI-0330171, DMI-0300359, CNS-0435312, and ECS-0426453, and in part by the Army Research Office under the ODDR&E MURI 2001 Program Grant DAAD19-01-1-0465 to the Center for Networked Communicating Control Systems. A preliminary conference version has appeared in the Proceedings of 39th Annual Allerton Conference on Communication, Control, and Computing 2001.

I. Ch. Paschalidis is with the Center for Information and Systems Engineering (CISE) and the Department of Manufacturing Engineering, Boston University, Brookline, MA 02446 USA (e-mail: yannisp@bu.edu).

S. Vassilaras is with Athens Information Technology, Peania 19002, Greece (e-mail: svas@ait.gr).

Digital Object Identifier 10.1109/TNET.2004.836139

is above a given threshold over the random synchronization phases.[1] This scheme has been used by many researchers and practitioners for *off-line* performance analysis purposes (e.g., dimensioning the multiplexer's buffer, estimating QoS, tuning admission control algorithms). It has also been used to check the accuracy of analytical results and validate traffic models (see, e.g., [16]–[18]). Its use for real-time Call Admission Control (CAC) or load balancing was prohibited by the long simulation times required by the direct Monte Carlo approach. Our quick simulation method reduces computational times enough to make such applications possible.

To develop an appropriate change of measure for the scheme described above we start by considering a simpler setup where each source generates traffic according to a periodic ON–OFF traffic model (this case was considered in [2] and the change of measure we propose was developed there). There is only one ON interval in each period and the only random variable is the starting time ϕ_i of the ON interval (same in each period but different for each source). Obviously, buffer overflows occur when the ON intervals of many sources overlap. We employ an LD result from [19] to infer an appropriate change of measure for the probability distribution of the ϕ_is and demonstrate through numerical results that a drastic improvement in simulation time can be achieved. This model is of independent interest since such periodic ON–OFF processes are suggested in [20] to have worst-case characteristics amongst all leaky-bucket regulated stationary processes, in the sense of maximizing the steady-state buffer overflow probability. Thus, it can be used to obtain a conservative estimate of this probability under more realistic traffic conditions.

We generalize the change of measure derived for this simple ON–OFF model to handle the general case where arrivals are derived from *actual traffic traces*. Our importance sampling approach is general enough to accommodate both continuous and discrete time models, as well as, multiplexing of both homogeneous and heterogeneous sources. We report extensive numerical results that show dramatic improvements in simulation time when compared with direct Monte Carlo simulation. Among the contributions of our work we consider the fact that arrival processes can be based on arbitrary traffic traces. This is in contrast to importance sampling work in the literature where specific stochastic traffic models are assumed (as, for example, in [14]).

After submitting this paper for publication, [21] was brought to our attention. The authors in [21] address the same problem and suggest an equivalent change of measure to the one we introduce in (10). They arrive to this change of measure by following a reasoning different than ours. However, as we explain in Section III-B, this change of measure failed to reduce the variance in many of our experiments. This led us to develop a modified change of measure [in (14)] which proved rather effective and is the one we use in all of our experiments. In Section III-B we attempt to intuitively explain the apparent inefficiency of the change of measure in (10).

The remainder of this paper is organized as follows. We begin in Section II with a primer on Large Deviations and importance sampling. In Section III we consider the simpler case of homogeneous periodic ON–OFF sources. We derive a change of measure that we use to speed up the simulation and compare the performance of the quick simulation to that of the direct Monte Carlo simulation. In Section IV we extend our analysis to the more general case of homogeneous periodic sources and demonstrate the validity of the results through simulation based on actual MPEG traces. In Section V we examine the most general case of multiplexing heterogeneous periodic sources. We discuss the applicability of our method to network resource planning and real-time CAC in Section VI. Conclusions are in Section VII.

II. Large Deviations and Importance Sampling Primer

In this section we briefly introduce some concepts from Large Deviations theory and discuss the main idea behind the importance sampling technique. Standard texts on LD are [9], [22], and [23]. A detailed discussion of quick simulation techniques can be found in [9].

The theory of large deviations is concerned with the estimation of rare event probabilities. Consider, for instance, a sequence of i.i.d. random variables $X_i, i \geq 1$, with mean $\mathbf{E}[X_1] = m$. Let $S_n = \sum_{i=1}^{n} X_i$. The strong law of large numbers asserts that S_n/n converges to m, as $n \to \infty$, with probability one (w.p. 1). Thus, for large n the event $S_n > na$, where $a > m$, (or $S_n < na$, for $a < m$) is a rare event. In particular, its probability behaves as $e^{-nr(a)}$, as $n \to \infty$, where the function $r(\cdot)$ determines the rate at which the probability of this event diminishes. Cramér's theorem [24] determines $r(\cdot)$, and is considered the first LD statement.

To state Cramér's theorem let us define $\Lambda(\theta)$ to be the logarithm of the moment generating function of X_1, i.e.,

$$\Lambda(\theta) \triangleq \log \mathbf{E}[e^{\theta X_1}].$$

We also denote by $\Lambda^*(\cdot)$ the Legendre transform of $\Lambda(\cdot)$, i.e.,

$$\Lambda^*(a) \triangleq \sup_{\theta}[\theta a - \Lambda(\theta)].$$

Cramér [24] established the following theorem:

Theorem II.1:

a) ***Upper Bound***: For any closed set $F \subset \mathbb{R}$

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbf{P}\left[\frac{S_n}{n} \in F\right] \leq - \inf_{x\in F} \Lambda^*(x).$$

b) ***Lower bound***: For any open set $G \subset \mathbb{R}$

$$\liminf_{n\to\infty} \frac{1}{n} \log \mathbf{P}\left[\frac{S_n}{n} \in G\right] \geq - \inf_{x\in G} \Lambda^*(x).$$

We say that $\{S_n\}$ satisfies a Large Deviations Principle (LDP) with *good rate function* $\Lambda^*(\cdot)$. The term "good" refers to the fact that the level sets $\{a \mid \Lambda^*(a) \leq k\}$ are compact for

[1] A variation of this scheme is also used in practice: sources are periodically "restarted" from some random point in their associated sequence. Typically, the restarting period is less than or equal to the length of the sequence and the sequence is wrapped-around if the end is reached before the restarting time. Both variants exhibit approximately equal buffer overflow probabilities when the length of the sequence is large. Thus, we will use the former variant as a good approximation of the latter.

all $k < \infty$. Cramér's Theorem intuitively asserts that for large enough n and for small $\epsilon > 0$

$$\mathbf{P}[S_n \in (na - n\epsilon, na + n\epsilon)] \sim e^{-n(\Lambda^*(a)+O(\epsilon))}.$$

Although Cramér's theorem applies to i.i.d. random variables it has been extended by Gärtner and Ellis [22] to cover the case where X_i are dependent random variables, e.g., state functionals of a Markov chain. This latter theorem holds under a certain technical assumption [22] which is satisfied by processes that are typically used to model traffic in communication networks, such as renewal processes, Markov-modulated processes, and stationary processes with mild mixing conditions.

We next turn to importance sampling. The importance sampling idea is rather simple. Consider a random variable X, and assume that we want to estimate $\mathcal{L} = \mathbf{E}_P[1_B(X)]$, where $1_B(X)$ denotes the indicator function of the event B, and the expectation is with respect to the distribution P of X. Assume also that X has a density $p(\cdot)$.

If we were to calculate the expectation through direct Monte Carlo simulation we would draw a sequence of D i.i.d. samples $x_1, \ldots, x_D$ from P and obtain the estimate

$$\hat{\mathcal{L}}_P = \frac{1}{D}\sum_{i=1}^{D} 1_B(x_i). \tag{1}$$

However, when the event B is very rare we need a huge sample size D to obtain a good estimate. Let now Q be some arbitrary distribution with density $q(\cdot)$ and consider a sequence of i.i.d. samples $y_1, \ldots, y_D$ drawn from Q. We can now form the estimate

$$\hat{\mathcal{L}}_Q = \frac{1}{D}\sum_{i=1}^{D} \frac{p(y_i)}{q(y_i)} 1_B(y_i). \tag{2}$$

Notice that the expected value of $\hat{\mathcal{L}}_Q$ with respect to Q is $\mathcal{L}$, since

$$\mathbf{E}_Q\left[\hat{\mathcal{L}}_Q\right] = \frac{1}{D}\sum_{i=1}^{D}\int \frac{p(y_i)}{q(y_i)} 1_B(y_i) q(y_i) dy_i = \mathcal{L}.$$

We will hereafter call the distribution Q *change of measure* and the ratio $(p(x)/q(x))$ *likelihood ratio of P versus Q*. The problem is to find a *change of measure* which reduces the variance of the estimator which is given by

$$\begin{aligned}\mathbf{Var}\left(\hat{\mathcal{L}}_Q\right) &= \frac{1}{D}\int \left(\frac{p(x)}{q(x)}1_B(x) - \mathcal{L}\right)^2 q(x)dx \\ &= \frac{1}{D}\left[\int \frac{(p(x)1_B(x))^2}{q(x)}dx - \mathcal{L}^2\right].\end{aligned}$$

The expression above is minimized when $q(x)$ is proportional to $p(x)1_B(x)$. But the normalizing constant is $1/\mathcal{L}$, precisely what we are trying to estimate. In general it is hard to obtain a good change of measure since it essentially requires solving an LD problem. The intuition behind the change of measure idea, is that we find a Q under which the event B is typical (as opposed to rare) and we use it to obtain an estimate that has the desired mean. In this paper, we do this heuristically for a special case of interest, and show numerical and theoretical evidence that indeed the simulation speeds up dramatically.

A particular case where an efficient change of measure is known, and will be useful later on in our analysis, concerns deviations of a sum of i.i.d. random variables. More specifically, consider a random variable X with distribution F_X and let $X_1, X_2, \ldots, X_N$ be N i.i.d. copies of X. Suppose we want to estimate the probability $\mathbf{P}\left[(1/N)\sum_{i=1}^{N} X_i > a\right]$ using importance sampling. Let $\tilde{X}$ be a random variable with the twisted measure

$$dF_{\tilde{X}}(x) = \frac{e^{\theta^* x} dF_X(x)}{\mathbf{E}[e^{\theta^* X}]} \tag{3}$$

where θ^* is the optimal solution of the optimization problem

$$\lambda^*(a) = \sup_{\theta}[\theta a - \lambda(\theta)]$$

and

$$\lambda(\theta) = \log \mathbf{E}[e^{\theta X}].$$

It is a direct consequence of the results presented in [9, Ch. VIII] and [10] that the variance of the direct Monte Carlo estimator for $\mathbf{P}\left[(1/N)\sum_{i=1}^{N} X_i > a\right]$ [cf. (1)] goes to zero like $e^{-N\lambda^*(a)}/\sqrt{N}$, while the variance of the estimator in the quick simulation (i.e., (2) using the measure in (3)) goes to zero like $e^{-2N\lambda^*(a)}/\sqrt{N}$. Assuming that $\lambda^*(a) > 0$ (e.g., if $a > \mathbf{E}[X]$ and $\lambda^*(a)$ is strictly convex), the twisted estimator can give orders of magnitude better performance for large N. Furthermore, [9, Ch. VIII] and [10] show that the change of measure in (3) is *uniquely asymptotically optimal* among all i.i.d. distributions. In particular, letting $\hat{\mathcal{L}}_Q$ be the twisted estimator for $\mathbf{P}\left[(1/N)\sum_{i=1}^{N} X_i > a\right]$ with twisted distribution Q, and D the sample size used in the quick simulation, [9, Ch. VIII] and [10] show that $dF_{\tilde{X}}$ as defined in (3) is the unique Q (among all i.i.d. distributions) that minimizes the speed factor

$$SF(\hat{\mathcal{L}}_Q) \triangleq \lim_{N\to\infty} \frac{1}{N}\log(D\mathbf{Var}[\hat{\mathcal{L}}_Q]). \tag{4}$$

III. Simple ON–OFF Traffic Case

As discussed in the Introduction we start in this Section with the simple ON–OFF traffic model. The change of measure we will devise will serve as the basis for the change of measure we will use for the more complicated cases we address.

A. Traffic Model and Problem Definition

Consider N independent traffic sources multiplexed into an infinite buffer which is served at a constant rate c. The arrival rate from source i at time t is denoted by $A_i(t)$. Arriving traffic is serviced according to a single-class, work-conserving, FCFS serving discipline. Both time and amount of traffic (arriving, in queue or departing) are considered continuous variables (continuous-time fluid model).

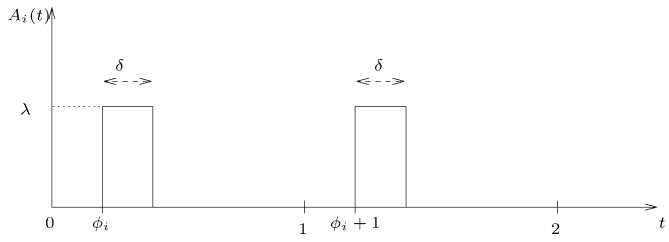

Fig. 1. Periodic ON–OFF with random phase source model.

The traffic sources conform to the model of Fig. 1. They generate traffic in a periodic fashion. Assume, without loss of generality, that the period is $T = 1$. Each source can be in one of two states: ON or OFF. In the ON state, it generates traffic at a constant rate λ, whereas in the OFF state it is silent. There is only one ON interval per period with a duration of $\delta < 0.5$. The phases of the sources, ϕ_i, are independent random variables uniformly distributed in $(0, 1]$. Thus, the probability density function for the vector of phases $\boldsymbol{\phi} = (\phi_1, \phi_2, \ldots, \phi_N)$ is $p(\boldsymbol{\phi}) = 1$ in $(0, 1]^N$ and zero elsewhere. We assume that the system is stable, that is, it holds $N\delta\lambda < c$. We want to estimate the probability that the queue length in the above system exceeds a given threshold U. When this level crossing probability is very small, it can be used to approximate the loss probability for an identical system with a finite buffer of size U. Assume that the sources start feeding the buffer at time $-\infty$. Let us now define two random variables: $S^{A,i}_{t_1,t_2}$ which represents the amount of work that source i generates in the interval $(t_1, t_2]$ (where $t_1 \leq t_2$), and $S^{A,\text{tot}}_{t_1,t_2}$ which denotes the total arriving traffic in that interval. In other words, $S^{A,i}_{t_1,t_2} = \int_{t_1}^{t_2} A_i(t)dt$ and $S^{A,\text{tot}}_{t_1,t_2} = \sum_{i=1}^{N} S^{A,i}_{t_1,t_2}$.

Let $L(t)$ be the queue length, in the infinite buffer system, at some arbitrary time t. Since the source model is periodic with period 1 and the system is stable, it is not hard to verify that the queue length $L(t)$ is also periodic with the same period. Given the above definitions, the probability we seek to estimate is

$$P_{\text{loss}} = \int \left(\frac{1}{T} \int_0^T 1_{\{L(t)>U\}}(\boldsymbol{\phi})dt \right) p(\boldsymbol{\phi})d\boldsymbol{\phi} = \int_0^1 \left(\int 1_{\{L(t)>U\}}(\boldsymbol{\phi})p(\boldsymbol{\phi})d\boldsymbol{\phi} \right) dt$$

where we write $1_{\{L(t)>U\}}(\boldsymbol{\phi})$ to explicitly denote the dependence of the indicator function on the vector of phases. Observe now that for two arbitrary times t and $t' \in (0, 1]$ the bijection $g : (0, 1]^N \to (0, 1]^N$ with $\boldsymbol{\phi}' = g(\boldsymbol{\phi}) = (\phi_1 + t' - t \bmod 1, \phi_2 + t' - t \bmod 1, \ldots, \phi_N + t' - t \bmod 1)$ has the property $1_{\{L(t)>U\}}(\boldsymbol{\phi}) = 1_{\{L(t')>U\}}(\boldsymbol{\phi}')$. And since for the uniformly distributed and independent ϕ_is, $p(\boldsymbol{\phi}) = p(\boldsymbol{\phi}')\ \forall\, \boldsymbol{\phi} \in (0, 1]^N$, we conclude that

$$\int 1_{\{L(t)>U\}}(\boldsymbol{\phi})p(\boldsymbol{\phi})d\boldsymbol{\phi} = \int 1_{\{L(t')>U\}}(\boldsymbol{\phi}')p(\boldsymbol{\phi}')d\boldsymbol{\phi}'$$

which implies

$$P_{\text{loss}} = \int 1_{\{L(t)>U\}}(\boldsymbol{\phi})p(\boldsymbol{\phi})d\boldsymbol{\phi} \tag{5}$$

for any arbitrary time $t \in (0, 1]$.

B. Loss Probability and Change of Measure

In order to devise an appropriate change of measure for the loss probability we will rely on LD asymptotics. In particular, we consider the many sources regime in which both the buffer size and the service capacity are scaled by N. That is, $U = Nb$ and $c = Ns$, where b and s are fixed quantities that have the interpretation of buffer size per source and capacity per source, respectively. We will next discuss an LD result for the loss probability in the many sources regime which was obtained in [19] and use it to infer an appropriate change of measure for importance sampling.

From the Lindley equation we obtain

$$L(0) = \max_{t \geq 0} \left[S^{A,tot}_{-t,0} - sNt \right]. \tag{6}$$

Due to the periodicity of the queue length, the maximum in (6) can be equivalently taken only over $0 \leq t \leq 1$. Let us now define

$$\lambda_t(\theta) \triangleq \frac{1}{t} \log \mathbf{E}\left[e^{\theta(S^{A,i}_{-t,0} - st)} \right]$$

which is the same for all sources i, and denoted by $\lambda^*_t(\cdot)$ the convex dual of $\lambda_t(\cdot)$, i.e.,

$$\lambda^*_t(a) = \sup_{\theta} [\theta a - \lambda_t(\theta)].$$

Under a technical assumption on $\lambda_t(\cdot)$, very similar to the technical assumption required for the Gärtner–Ellis theorem (see [22]), and a local regularity condition on the sample paths of the workload process $\left\{ S^{A,tot}_{-t,0} - sNt; t \geq 0 \right\}$, the following theorem is proved in [19].

Theorem III.1: ([19]) For each $b > 0$

$$\lim_{N\to\infty} \frac{1}{N} \log \mathbf{P}[L(0) > Nb] = -I(b)$$

where

$$I(b) = \inf_{t \geq 0} t\lambda^*_t\left(\frac{b}{t}\right).$$

Proof: We present the proof of the lower bound since this is informative on the change of measure that we use in order to estimate the probability above through simulation. The upper bound proof can be found in [19].

We have that

$$\begin{aligned} &\liminf_{N\to\infty} \frac{1}{N} \log \mathbf{P}\left[\max_{t\geq 0} \left(S^{A,\text{tot}}_{-t,0} - sNt \right) > Nb \right] \\ &\quad \geq \liminf_{N\to\infty} \frac{1}{N} \sup_{t\geq 0} \log \mathbf{P}\left[\left(S^{A,\text{tot}}_{-t,0} - sNt \right) > Nb \right] \\ &\quad \geq \sup_{t\geq 0} \liminf_{N\to\infty} \frac{1}{N} \log \mathbf{P}\left[\left(S^{A,\text{tot}}_{-t,0} - sNt \right) > Nb \right]. \end{aligned} \tag{7}$$

Now note that the moment generating function of $\left(S^{A,\text{tot}}_{-t,0} - sNt \right)$ is

$$\mathbf{E}\left[e^{\theta\left(S^{A,\text{tot}}_{-t,0} - sNt\right)} \right] = \left(\mathbf{E}\left[e^{\theta\left(S^{A,i}_{-t,0} - st\right)} \right] \right)^N$$

which by using the definition of $\lambda_t(\cdot)$ implies that

$$t\lambda_t(\theta) = \log \mathbf{E}\left[e^{\theta\left(S^{A,i}_{-t,0} - st\right)}\right] = \frac{1}{N}\log \mathbf{E}\left[e^{\theta\left(S^{A,\text{tot}}_{-t,0} - sNt\right)}\right].$$

Thus, applying the lower bound of Cramér's theorem to the right-hand side of (7), we obtain

$$\liminf_{N\to\infty} \frac{1}{N}\log \mathbf{P}\left[\left(S^{A,\text{tot}}_{-t,0} - sNt\right) > Nb\right] \geq -(t\lambda_t)^*(b) = -t\lambda_t^*\left(\frac{b}{t}\right)$$

where the superscript $*$ denotes the convex dual and the last equality above is obtained by using convex duality properties (see [25]). Combining the above with (7) we finally obtain the desired result. ∎

The above theorem intuitively asserts that for large N the overflow probability behaves as

$$\mathbf{P}[L(0) > Nb] \sim e^{-NI(b)}.$$

Let τ the optimal solution of the optimization problem associated with the LD rate function $I(b)$. We can interpret τ as the most likely duration of the busy period that leads to the overflow. Due to the periodicity of the queue length process we have $\tau \in [0,1]$.

We next present a heuristic change of measure that we use to speed up the simulation. Consider the lower bound on $\mathbf{P}[L(0) > Nb]$ developed in the proof of Theorem III.1. As it is typical with LD results, it identifies the most likely way in which the buffer exceeds Nb, hence it also suggests a change of measure. In particular, the proof of Theorem III.1 asserts that $\mathbf{P}[L(0) > Nb]$ has the same LD exponent as $\mathbf{P}\left[\left(S^{A,\text{tot}}_{-\tau,0} - sN\tau\right) > Nb\right]$ [see (7)], where, as defined above, τ is the optimal solution of the optimization problem in the definition of $I(b)$. Since the latter two probabilities are asymptotically equal, changing the measure in a way that $S^{A,\text{tot}}_{-\tau,0} - sN\tau > Nb$ occurs more often will also lead to more frequent buffer overflows. Next note that $\left(S^{A,\text{tot}}_{-\tau,0} - sN\tau\right)$ is a sum of N i.i.d. random variables $W^i_\tau \triangleq S^{A,i}_{-\tau,0} - s\tau$, $i = 1,\ldots,N$. Let F_{W_τ} denote their common distribution and let W_τ be a random variable with the same distribution. It follows that $\mathbf{P}\left[\left(S^{A,\text{tot}}_{-\tau,0} - sN\tau\right) > Nb\right] = \mathbf{P}\left[(1/N)\sum_{i=1}^N W^i_\tau > b\right]$. To estimate the latter we can use the exponential change of measure for sums of i.i.d. random variables presented in Section II. In particular, let $\tilde{W}_\tau$ be a random variable with the asymptotically optimal changed measure with [cf. (3)]

$$dF_{\tilde{W}_\tau}(w) = \frac{e^{\theta^*_\tau w} dF_{W_\tau}(w)}{\mathbf{E}[e^{\theta^*_\tau W_\tau}]} \tag{8}$$

where θ^*_τ is the optimal solution of the optimization problem in

$$\lambda^*_\tau(b/\tau) = \sup_\theta \left[\theta b/\tau - \lambda_\tau(\theta)\right].$$

Recall now that only the phase ϕ_i of the source is random, thus, W^i_τ is a deterministic function of ϕ_i, which is uniformly distributed in $(0,1]$. To explicitly denote this we will write $W^i_\tau(\phi_i) = S^{A,i}_{-\tau,0}(\phi_i) - s\tau$, where $S^{A,i}_{t_1,t_2}(\cdot)$ is a deterministic function that maps the phase ϕ_i of a source i to the amount of traffic transmitted from this source in the interval $(t_1,t_2]$. Thus, from (8) we obtain the change of measure for the phase

$$q_i(\phi_i) = p_i(\phi_i)\frac{e^{\theta^*_\tau S^{A,i}_{-\tau,0}(\phi_i)}}{\mathbf{E}\left[e^{\theta^*_\tau S^{A,i}_{-\tau,0}(\phi_i)}\right]} \tag{9}$$

where $p_i(\phi_i)$ is the original uniform density and $q(\boldsymbol{\phi}) = \prod_{i=1}^N q_i(\phi_i)$.

Let us now return to the particular ON–OFF source model we are considering and explicitly calculate the density $q(\cdot)$ resulting from the change of measure in (9). In the following calculations we have shifted time by τ, hence focusing on the overflow probability $\mathbf{P}[L(\tau) > Nb]$. The change of measure then becomes

$$q_i(\phi_i) = p_i(\phi_i)\frac{e^{\theta^*_\tau S^{A,i}_{0,\tau}(\phi_i)}}{\mathbf{E}\left[e^{\theta^*_\tau S^{A,i}_{0,\tau}(\phi_i)}\right]}. \tag{10}$$

We divide $[0,1]$ in the three subintervals[2] $[0,\delta]$, $[\delta,1-\delta]$, and $[1-\delta,1]$. After a fair amount of routine calculations we obtain that for $\tau \in [0,\delta]$

$$S^{A,i}_{0,\tau}(\phi_i) = \begin{cases} (\tau-\phi_i)\lambda, & \text{if } \phi_i \in [0,\tau] \\ 0, & \text{if } \phi_i \in [\tau, 1-\delta] \\ (\phi_i+\delta-1)\lambda, & \text{if } \phi_i \in [1-\delta, \tau+1-\delta] \\ \tau\lambda, & \text{if } \phi_i \in [\tau+1-\delta, 1] \end{cases} \tag{11}$$

and

$$\mathbf{E}\left[e^{\theta S^{A,i}_{0,\tau}}\right] = 2\frac{e^{\theta\tau\lambda}-1}{\theta\lambda} + (1-\delta-\tau) + e^{\theta\tau\lambda}(\delta-\tau). \tag{12}$$

For $\tau \in [\delta, 1-\delta]$ we obtain

$$S^{A,i}_{0,\tau}(\phi_i) = \begin{cases} \lambda\delta, & \text{if } \phi_i \in [0,\tau-\delta] \\ (\tau-\phi_i)\lambda, & \text{if } \phi_i \in [\tau-\delta,\tau] \\ 0, & \text{if } \phi_i \in [\tau, 1-\delta] \\ (\phi_i+\delta-1)\lambda, & \text{if } \phi_i \in [1-\delta, 1] \end{cases}$$

and

$$\mathbf{E}\left[e^{\theta S^{A,i}_{0,\tau}}\right] = 2\frac{e^{\theta\delta\lambda}-1}{\theta\lambda} + (1-\delta-\tau) + e^{\theta\delta\lambda}(\tau-\delta).$$

Finally, for $\tau \in [1-\delta, 1]$ we obtain

$$S^{A,i}_{0,\tau}(\phi_i) = \begin{cases} \lambda\delta, & \text{if } \phi_i \in [0,\tau-\delta] \\ (\tau-\phi_i)\lambda, & \text{if } \phi_i \in [\tau-\delta, 1-\delta] \\ (\delta-1+\tau)\lambda, & \text{if } \phi_i \in [1-\delta,\tau] \\ (\phi_i+\delta-1)\lambda, & \text{if } \phi_i \in [\tau, 1] \end{cases}$$

and

$$\mathbf{E}\left[e^{\theta S^{A,i}_{0,\tau}}\right] = e^{\theta\delta\lambda}(\tau-\delta) + 2\frac{e^{\theta\delta\lambda} - e^{\theta(\tau-1+\delta)\lambda}}{\theta\lambda} + (\tau-1+\delta)e^{\theta(\delta-1+\tau)\lambda}.$$

Fig. 2 shows a plot of the density $q_i(\phi_i)$ when the arrival rate from one source during an ON interval is $\lambda = 2500$, the buffer size per source is $b = 6$, the ON interval duration is $\delta = 0.1$ and the service capacity per source is $s = 500$. For these parameters it turns out that $\tau = 0.0799$, so (11) and (12) are applicable.

[2]Recall that we have assumed $\delta < 0.5$.

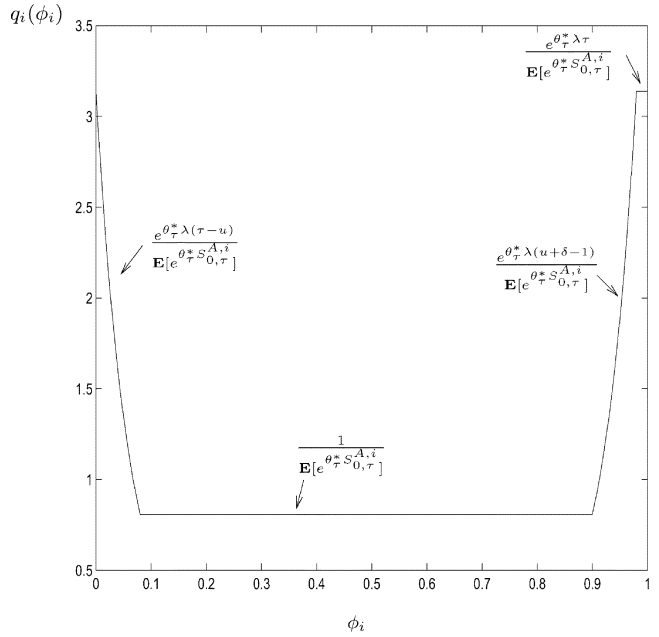

Fig. 2. Plot of the change of measure described by (10)–(12), ($\tau \leq \delta$).

Note that as long as we scale the buffer size and capacity with the number of sources, the change of measure is independent of the number of sources. As it is shown in the plot, in order to increase the probability of an overflow at τ, the sources need to get to their ON state with much higher probability in the intervals $[1-\delta, 1]$ and $[0, \tau]$.

Let now $\hat{\mathcal{L}}_P$ be the estimate of P_{loss} obtained from the direct simulation and $\hat{\mathcal{L}}_Q$ the one obtained from the simulation that uses importance sampling. As discussed above, the change of measure $q_i(\phi_i)$ in (10) increases the probability that an overflow occurs at $t = \tau$. In reality, an overflow is equally likely to occur at any time, due to the uniform distribution of the phases. Therefore, if we want to use $q_i(\phi_i)$ in order to estimate P_{loss} via quick simulation, we can take one of the following two approaches.

- Use the change of measure $q_i(\phi_i)$ in (10) to obtain an estimate of $\mathbf{P}[L(\tau) > U]$ which is equal to $P_{\text{loss}} = \int 1_{\{L(\tau)>U\}}(\boldsymbol{\phi}) p(\boldsymbol{\phi}) d\boldsymbol{\phi}$ by (5). In this case, the estimators of P_{loss} take the form

$$\hat{\mathcal{L}}_P = \frac{1}{D} \sum_{i=1}^{D} 1_{\{L(\tau)>U\}}(\boldsymbol{\phi}^{(i)})$$

and

$$\hat{\mathcal{L}}_Q = \frac{1}{D} \sum_{i=1}^{D} \frac{p(\boldsymbol{\phi}^{(i)})}{q(\boldsymbol{\phi}^{(i)})} 1_{\{L(\tau)>U\}}(\boldsymbol{\phi}^{(i)}) \tag{13}$$

  where $\boldsymbol{\phi}^{(i)}$ is the ith random vector of phases in a sample $\boldsymbol{\phi}^{(1)}, \boldsymbol{\phi}^{(2)}, \ldots, \boldsymbol{\phi}^{(D)}$.
- Use (10) to derive a different change of measure that increases the probability of overflow uniformly at all times. To achieve this, we employ the following idea. We first generate a sequence of random phases from $q(\boldsymbol{\phi})$. Let $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_N)$ be the resulting random sequence. We then let u be a uniform random variable in $(0, 1]$ and shift the vector $\boldsymbol{\phi}$ by u to obtain

$$\boldsymbol{\xi} = (\xi_1, \ldots, \xi_N) = (\phi_1 + u, \ldots, \phi_N + u) \, (\text{mod} 1).$$

The intuitive idea is that with the shift the overflow time will be uniformly distributed in $(0, 1]$. We use samples of $\boldsymbol{\xi}$ to generate the phases of the sources in the simulation. Note that the likelihood ratio for a sample $\boldsymbol{\xi}$, is $1/r(\boldsymbol{\xi})$, where

$$\begin{aligned} r(\boldsymbol{\xi}) &= \int_0^1 q(\xi_1 - a, \ldots, \xi_N - a \mid u = a) da \\ &= \int_0^1 q(\xi_1 - a, \ldots, \xi_N - a) da \\ &= \int_0^1 \prod_{i=1}^{N} q_i(\xi_i - a) da. \end{aligned} \tag{14}$$

The second equality above holds since u is independent of the phase vector $\boldsymbol{\phi}$ drawn from $q(\cdot)$. In this case the estimators of P_{loss} take the form

$$\hat{\mathcal{L}}_P = \frac{1}{D} \sum_{i=1}^{D} \rho(\boldsymbol{\phi}^{(i)})$$

and

$$\hat{\mathcal{L}}_Q = \frac{1}{D} \sum_{i=1}^{D} \frac{p(\boldsymbol{\phi}^{(i)})}{r(\boldsymbol{\phi}^{(i)})} \rho(\boldsymbol{\phi}^{(i)})$$

where $\boldsymbol{\phi}^{(i)}$ is the ith random vector of phases in a sample $\boldsymbol{\phi}^{(1)}, \boldsymbol{\phi}^{(2)}, \ldots, \boldsymbol{\phi}^{(D)}$ generated in the direct Monte Carlo and the quick simulations, respectively, and where

$$\rho(\boldsymbol{\phi}^{(i)}) = \int_0^1 1_{\{L(t)>U\}}(\boldsymbol{\phi}^{(i)}) dt$$

is the fraction of the $[0, 1]$ interval for which $L(t) > U$.

Now, as explained earlier, the change of measure $q(\boldsymbol{\phi})$, is an optimal change of measure for $\mathbf{P}\left[\left(S^{A,\text{tot}}_{0,\tau} - sN\tau\right) > Nb\right]$, and only a candidate heuristic change of measure for the estimation of $\mathbf{P}[L(\tau) > Nb]$ via importance sampling. It turns out that while the first approach fails to demonstrate any speed up of the simulation, the variance reduction obtained by the second one is very satisfactory, as it will be demonstrated through numerical results in Section III-C. An intuitive explanation of this discrepancy is that the change of measure $q(\boldsymbol{\phi})$ assigns a very small probability (much smaller than their original probability) to $\boldsymbol{\phi}$s for which $L(\tau) > Nb$ with $\left(S^{A,\text{tot}}_{0,\tau} - sN\tau\right) < Nb$. When such a $\boldsymbol{\phi}$ is encountered, during the course of quick simulation, it immediately blows up the variance of the estimator. To see this, note from (13) that when $q(\boldsymbol{\phi}^{(i)})$ is small and $1_{\{L(\tau)>U\}}(\boldsymbol{\phi}^{(i)}) = 1$ the sample $\boldsymbol{\phi}^{(i)}$ contributes a very large value to the estimator and increases its variance. Apparently though, for most of these $\boldsymbol{\phi}$s there exist some moment $\sigma \neq 0$ for which $\left(S^{A,\text{tot}}_{\sigma,\sigma+\tau} - sN\tau\right) > Nb$. Therefore, they are assigned a higher probability under the change of measure $r(\boldsymbol{\phi})$. The remaining $\boldsymbol{\phi}$s occur with a very small probability

TABLE I
COMPARISON OF RESULTS FROM DIRECT MONTE CARLO SIMULATION AND QUICK SIMULATION FOR THE ON–OFF MODEL

	Direct simulation		Quick Simulation		
N	P_{loss}	D	P_{loss}	D	SU
20	$(3.40\pm0.17)\times10^{-2}$	4.0×10^{3}	$(3.58\pm0.14)\times10^{-2}$	2.0×10^{3}	2.0
40	$(3.47\pm0.17)\times10^{-3}$	3.5×10^{4}	$(3.56\pm0.18)\times10^{-3}$	3.0×10^{3}	11.7
60	$(4.05\pm0.20)\times10^{-4}$	2.5×10^{5}	$(4.39\pm0.20)\times10^{-4}$	5.0×10^{3}	50
80	$(5.00\pm0.25)\times10^{-5}$	1.8×10^{6}	$(5.11\pm0.25)\times10^{-5}$	6.0×10^{3}	300
100	$(6.39\pm0.32)\times10^{-6}$	1.3×10^{7}	$(6.55\pm0.32)\times10^{-6}$	8.0×10^{3}	1625
120	$(8.35\pm0.42)\times10^{-7}$	8.9×10^{7}	$(8.32\pm0.40)\times10^{-7}$	1.2×10^{4}	7417

under the original measure and being assigned an even smaller probability by $r(\boldsymbol{\phi})$ does not affect the efficiency of this change of measure.

C. Numerical Results

Next we report numerical results that indicate the speed up in the simulation with the change of measure $r(\cdot)$ derived above.

We simulated the system we have considered in this paper for different values of N with the following parameters: $\lambda = 2500$, $\delta = 0.1$, $b = 6$, and $s = 500$. In Table I we compare the estimates obtained from direct Monte Carlo simulation and quick simulation with the change of measure $r(\cdot)$ described above. D denotes the sample size required in order to estimate P_{loss} within $\pm5\%$ with a 95% confidence level. We calculated D with an accuracy of ±1000, that is, we gradually increased the sample size in the simulation by 1000 until half the size of the 95% confidence interval was less than 5% of the estimated P_{loss}. In the table we report the resulting value of D rounded off to show only its first two significant digits. The last column of the table presents the achieved speed-up (SU) as the ratio of the sample size for the direct simulation over the sample size needed for the quick simulation to obtain the same confidence interval. To provide a flavor of the required simulation running times we note that for the last row of the table quick simulation took 18 s while direct Monte Carlo took 19 h 16 min.[3]

IV. A MORE GENERAL TRAFFIC MODEL

A. Problem Generalization and Change of Measure

Let us now consider the case of multiplexing N generic periodic sources that are randomly synchronized. More specifically, the arrival rate from source i at time t is given by $A_i(t) = f(t-\phi_i)$, where $f(\cdot)$ is a deterministic periodic function with period T and ϕ_i is a random phase, uniformly distributed in $(0,T]$. Thus, the probability density function for the vector of phases $\boldsymbol{\phi} = (\phi_1, \phi_2, \ldots, \phi_N)$ is $p(\boldsymbol{\phi}) = \prod_{i=1}^{N}(1/T) = (1/T^N)$ in $(0,T]^N$ and zero elsewhere. The stability condition $S_{0,T}^{A,\text{tot}} < cT$ for this system can be written as

$$N\int_0^T f(t)dt < cT.$$

Let $L(t)$ be the queue length, in the infinite buffer system, at some arbitrary time t. Since the source model is periodic with period T and the system is stable, it is not hard to verify that the queue length $L(t)$ is also periodic with the same period. The probability we seek to estimate is [cf. (5)]

$$P_{\text{loss}} = \int\left(\frac{1}{T}\int_0^T 1_{\{L(t)>U\}}(\boldsymbol{\phi})dt\right)p(\boldsymbol{\phi})d\boldsymbol{\phi} = \int 1_{\{L(\tau)>U\}}(\boldsymbol{\phi})p(\boldsymbol{\phi})d\boldsymbol{\phi}$$

for any arbitrary τ.

The development of the change of measure $r(\cdot)$ in (14) of the previous section was in fact independent of the special ON–OFF shape of $f(\cdot)$. This change of measure can therefore be used for any $f(\cdot)$. Note that the expression in (14) holds for the special case where $T = 1$ and that in the general case

$$r(\boldsymbol{\xi}) = \int_0^T \prod_{i=1}^{N} q_i(\xi_i - a)da. \tag{15}$$

We have started our analysis from the simple case in which the analytical calculation of $S_{0,\tau}^{A,i}(\phi_i)$ and consequently $q_i(\phi_i)$ was possible. This allowed for plotting $q_i(\phi_i)$ and gaining some insight into the way the proposed change of measure works. Calculating $S_{0,\tau}^{A,i}(\phi_i)$ analytically, is practically nontractable for all but some trivial functions $f(\cdot)$. For example, we have calculated the change of measure $r(\cdot)$ for several ON–OFF periodic sources with two ON intervals (with different durations and transmission rates) in each period and experimentally verified that they achieve speed-ups similar to those reported in Section III-C.

When $S_{0,\tau}^{A,i}(\phi_i)$ cannot be calculated analytically, numerical calculation through time discretization can be used. In practice, it is most common that $f(\cdot)$ is given in the form of some discrete time sequence $f_d : \{1,2,\ldots,K\} \to \mathbb{R}_+$ where $f_d(k)$ represents the amount of traffic arriving in the interval $(k-1,k]$. This discretization typically arises naturally from the system we are interested in simulating. For example, the sequence f_d may have been obtained by actual observations of traffic transmitted through a communication link or may be the sequence of bits per frame at the output of a video compressor as explained in the Introduction. But even in the case that the arrival rate is modeled by a continuous function $f(\cdot) : (0,T] \to \mathbb{R}_+$ it is quite straightforward to discretize time by setting $f_d(k) = \int_{(k-1)\delta}^{k\delta} f(t)dt$ where $\delta = T/K$.

The appropriate change of measure for the discrete time case is almost identical to the one developed for the continuous time case although it applies to a probability mass function instead of a probability density function. For the discrete time model ($t \in \mathbb{Z}$), the period is denoted by K and $\phi_i \in \{1,2,\ldots,K\}$. $A_i(t)$ denotes the amount of traffic generated by source i at

[3]All runs reported in the paper were performed on a Pentium III, 600 MHz, with 384 MB RAM. Occasionally during the runs the PC was used for other typical office tasks, so the reported times should not be taken as being very accurate. We report them to provide an indication of how the SU factor translates into running times.

time slot t and the capacity of the link c indicates the maximum amount of traffic that can depart from the buffer at any time slot t. The phases ϕ_i are independent, uniformly distributed discrete random variables in $\{1,\dots,K\}$. Thus, the probability mass function for the vector of phases $\boldsymbol{\phi} = (\phi_1, \phi_2, \dots, \phi_N)$ is $p(\boldsymbol{\phi}) = \prod_{i=1}^{N} (1/K) = (1/K^N)$ in $\{1,\dots,K\}^N$ and zero elsewhere.

The random variables $S_{t_1,t_2}^{A,i}$ and $S_{t_1,t_2}^{A,\text{tot}}$ can be seen as $S_{t_1,t_2}^{A,i} = \sum_{t=t_1+1}^{t_2} A_i(t)$ and $S_{t_1,t_2}^{A,\text{tot}} = \sum_{i=1}^{N} S_{t_1,t_2}^{A,i}$. The stability condition $S_{1,K}^{A,\text{tot}} < cK$ can be written as

$$N \sum_{t=1}^{K} f_d(t) < cK.$$

Similarly to the continuous time case, the probability we seek to estimate is

$$\begin{aligned} P_{\text{loss}} &= \frac{1}{K^N} \sum_{\boldsymbol{\phi}} \left(\frac{1}{K} \sum_{t=1}^{K} 1_{\{L(t)>U\}}(\boldsymbol{\phi}) \right) \\ &= \frac{1}{K^N} \sum_{\boldsymbol{\phi}} 1_{\{L(\tau)>U\}}(\boldsymbol{\phi}) \end{aligned}$$

for any given τ.

Given the above definitions, the analysis of Section III-B holds for the discrete time case as well, up to the calculation of the change of measure in (10). The change of measure in (14) needs to be adapted to discrete time as follows

$$\begin{aligned} r(\boldsymbol{\xi}) &= \frac{1}{K} \sum_{a=1}^{K} q(\xi_1 - a, \dots, \xi_N - a) \\ &= \frac{1}{K} \sum_{a=1}^{K} \prod_{i=1}^{N} q_i(\xi_i - a). \end{aligned} \tag{16}$$

B. Quick Simulation Overhead

Before we proceed with numerical results, a couple of comments are in order regarding the computational overhead imposed by the quick simulation versus the direct Monte Carlo simulation.

We distinguish between the initialization phase (during which we first calculate the parameters τ and θ_τ^* and then the probability mass function $q_i(\phi_i)$ for all K values of ϕ_i) and the simulation phase which requires a much smaller sample size D than the direct simulation but a more complex calculation to generate each sample.

Calculating the parameters τ and θ_τ^* involves solving the following nested optimization problem

$$\inf_{1\le t\le T} \left[t \sup_{\theta} \left(\frac{\theta b}{t} - \frac{1}{t} \log \mathbf{E}\left[e^{\theta\left(S_{-t,0}^{A,i}(\phi_i) - st\right)} \right] \right) \right].$$

Notice that in this expression θ is continuous while t is discrete and that the function to be maximized is concave with respect to θ while the function to be minimized is convex with respect to t. The computational time to obtain an optimal solution (τ, θ_τ^*) depends on the optimization algorithms used and the desirable precision in determining θ_τ^*. Note that it is not necessary to determine θ_τ^* with high precision since the efficiency of the change of measure is robust to small discrepancies from the optimal θ (see [11] for a detailed discussion on the robustness of exponentially twisted changes of measure). For every evaluation of the objective function of the optimization problem, $\mathbf{E}\left[e^{\theta\left(S_{-t,0}^{A,i}(\phi_i) - st\right)} \right]$ needs to be computed, which can be done in $O(tK) = O(K^2)$ elementary operations.

Once we have the values of τ and θ_τ^*, calculating $q_i(\phi_i)$ for all K values of ϕ_i using (10) takes $O(K)$ steps if we have kept the values of $e^{\theta_\tau^* S_{-\tau,0}^{A,i}(\phi_i)}$ and $\mathbf{E}\left[e^{\theta_\tau^* S_{-\tau,0}^{A,i}(\phi_i)} \right]$ from the previous calculations. We next compute the cumulative distribution $Q_i(\phi_i)$ for all K values of ϕ_i (time complexity $O(K^2)$) which will be used to generate samples of ϕ_is. We conclude that the time complexity of the initialization phase is polynomial in K.[4]

During the simulation phase, the time overhead in one iteration for the quick simulation lies in the generation of the vector of phases $\boldsymbol{\phi}$ with the twisted measure and the calculation of the likelihood ratio. To that end, we first generate N random variables $h_i, i = 1,\dots,N$, with uniform distribution in $[0,1]$. Then, for each i we "invert $Q(\cdot)$," that is, we use some array searching algorithm to find the index j of the smallest Q_j for which $Q_j \ge h_i$. This index is a random variable ϕ_i with probability mass $q_i(\phi_i)$. This array search can be performed by a simple binary search with time complexity $O(\log K)$ although more sophisticated algorithms involving hash tables are also possible. Then, as outlined in Section III-B we shift (modulo K) all ϕ_is by the same random amount u, which is uniformly distributed in $\{1,\dots,K\}$, to generate a random vector $\boldsymbol{\xi}$ distributed according to $r(\boldsymbol{\xi})$. Calculating $r(\boldsymbol{\xi})$ for a particular $\boldsymbol{\xi}$ using (16) has time complexity $O(NK)$. Therefore, the time overhead for generating $\boldsymbol{\xi}$ in the quick simulation is $O(N \log K) + O(NK) = O(NK)$. On the other hand, the number of samples D needed to obtain the same level of estimation accuracy grows exponentially in N for the direct simulation versus linearly in N for the quick simulation. Hence for N large enough (equivalently, P_{loss} small enough) the performance of quick simulation is better by orders of magnitude.

C. Numerical Results

We next report numerical results that demonstrate the effectiveness of the above change of measure for generic sequences f_d. We draw the data for our experiments from sequences of bits per group of pictures (GOP) in MPEG video traces. In the first experiment the sequence f_d depicts the numbers of Mbits/GOP for the first 1000 GOPs of the "Star Wars" MPEG trace. Fig. 3 shows a graph of f_d versus time. We plot this sequence to demonstrate the generality of our numerical results in the sense that it shows a pretty random sequence of Mbits/GOP with no particular pattern that could suggest limited applicability of our method. The parameters used in this simulation are K =1000 GOPs, s = 0.212 Mbits/GOP and b = 0.4 Mbits which result to τ = 58 GOPs and θ_τ^* = 0.1517. In Table II we compare the estimates obtained from the direct Monte Carlo

[4]To provide an indication on how this translates into running time, let us note that the initialization phase for the experiments we report in the sequel took on the order of 20 s for the "Star Wars" and "Super Bowl" traces, 45 s for the "Asterix" trace, and 90 s for the experiment with heterogeneous sources. The computations were performed using Matlab and no particular attention was paid to optimize the code. We will not include those initialization times in the quick simulation running times we report later on.

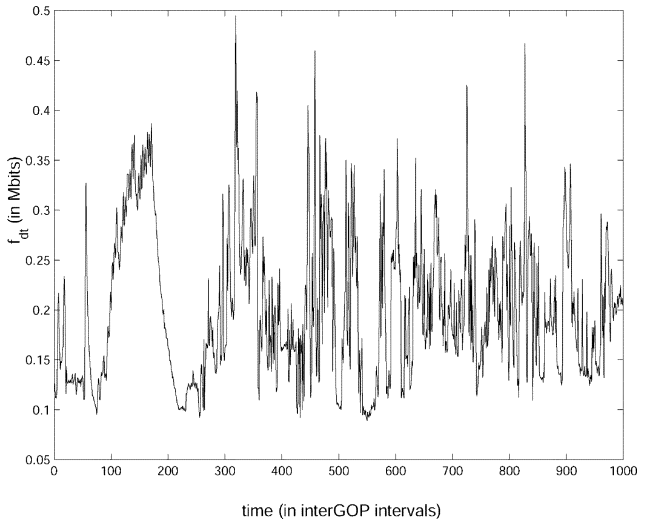

Fig. 3. Megabits per GOP for the first 1000 GOPs of the "Star Wars" trace.

simulation and the quick simulation with the change of measure $r(\cdot)$ in (16). The quantities D and SU have the same meaning as in Section III-C. The running times for the last row of the table were 1 min 45 s for the quick simulation versus 1day 9 h 56 min for the direct Monte Carlo simulation.

In the second experiment the sequence f_d is formed from the number of Mbits/GOP for the first 1000 GOPs of an MPEG trace taken from a "Super Bowl" game of American football. The parameters used in this simulation are $K = 1000$ frames, $s = 0.275$ Mbits/frame and $b = 0.5$ Mbits which result to $\tau = 107$ frames and $\theta^*_\tau = 0.0993$. In Table III we compare the estimates obtained from the direct Monte Carlo simulation and the quick simulation with the change of measure $r(\cdot)$ in (16). The running times for the last row of the table were 1 min 22 s for the quick simulation versus 11 days 6 h 29 min for the direct Monte Carlo simulation.

In the third experiment the sequence f_d is formed from the number of Mbits/GOP for the first 3000 GOPs of an "Asterix" cartoon MPEG trace. The parameters used in this simulation are $K = 3000$ GOPs, $s = 0.29$ Mbits/GOP, and $b = 0.5$ Mbits which result to $\tau = 80$ GOPs and $\theta^*_\tau = 0.0692$. In Table IV we compare the estimates obtained from the direct Monte Carlo simulation and the quick simulation with the change of measure $r(\cdot)$ in (16). The running times for the last row of the table were 11 min 35 s for the quick simulation versus 14 days 20 h 39 min for the direct Monte Carlo simulation.

The time complexity of these simulation experiments is proportional to $K \times D$. In all the experiments of this section and the one in Section V we are limiting the length K of the generating sequences so that the direct simulation could be completed in a reasonable time (no more than two weeks). When using quick simulation we can afford a much larger K.

V. Multiplexing Heterogeneous Sources

A. Derivation of the Change of Measure

So far, for the sake of simplicity, we have assumed that all multiplexed sources are homogeneous. Let us now consider the case where there are J types of sources with N_j of them being of type j and $\sum_{j=1}^{J} N_j = N$. For a continuous time model $(t \in \mathbb{R}, T \in \mathbb{R}_+, \phi_i \in (0,T])$, the arrival rate from source i of type j at time t is given by $A_i(t) = f_j(t - \phi_i)$, where $f_j(\cdot)$ is a deterministic periodic function with period T and $\phi_i \in (0,T]$ is a random phase. For a discrete time model $(t \in \mathbb{Z}, T \in \mathbb{Z}_+, \phi_i \in \{1,2,\ldots,T\})$, the same definition applies where $A_i(t)$ denotes the number of arrivals at time t. For uniformity of notation we will stretch the notion of an interval to the set of integers $\mathbb{Z}$ to denote the subset of $\mathbb{Z}$ contained in that interval. For instance: $(t_1,t_2] \triangleq \{t_1+1, t_1+2, \ldots, t_2\}$. The phases ϕ_i are independent, uniformly distributed random variables. Thus, the probability distribution for the vector of phases $\boldsymbol{\phi} = (\phi_1, \phi_2, \ldots, \phi_N)$ is $p(\boldsymbol{\phi}) = \prod_{i=1}^{N} (1/T) = (1/T^N)$ in $(0,T]^N$ and zero elsewhere.

For the case of multiplexing heterogeneous sources in addition to the already known random variables $S^{A,i}_{t_1,t_2}$ and $S^{A,\mathrm{tot}}_{t_1,t_2}$ we define $S^{A,\mathrm{type\_}j}_{t_1,t_2}$ expressing the amount of work generated by a source of type j in the interval $(t_1,t_2]$. The stability condition $S^{A,\mathrm{tot}}_{0,T} < cT$ now becomes

$$\sum_{j=1}^{J} \left(N_j \int_0^T f_j(t)dt \right) < cT$$

when time is continuous and

$$\sum_{j=1}^{J} \left(N_j \sum_{t=1}^{T} f_j(t) \right) < cT$$

when time is discrete.

To apply the change of measure developed for the homogeneous case to the heterogeneous case the definition of $\lambda_t(\theta)$ has to be extended to account for the different source types. Theorem III.1 is proven in [19] to hold in this case (for both discrete and continuous time) with

$$\lambda_t(\theta) \triangleq \frac{1}{t} \sum_{j=1}^{J} \zeta_j \log \mathbf{E}\left[e^{\theta\left(S^{A,\mathrm{type\_}j}_{-t,0} - st\right)} \right]$$

where we assume that $N_j = \zeta_j N$, $j = 1, \ldots, J$, for some constants $\zeta_j \in [0,1]$ that satisfy $\sum_{j=1}^{J} \zeta_j = 1$. As in the homogeneous case, we solve the optimization problem

$$\inf_{0 \le t \le T} \left[t \sup_{\theta} \left(\frac{\theta b}{t} - \lambda_t(\theta) \right) \right].$$

Let (τ, θ^*_τ) be the optimal solution. We can again interpret τ as the most likely duration of the busy period that leads to an overflow. The difference is that in the heterogeneous case $S^{A,\mathrm{tot}}_{0,\tau} - sN\tau$ is a sum of N independent but not identically distributed random variables, so the analysis in Section III-B does not go directly through. Nevertheless, it is quite straightforward to extend the asymptotically optimal change of measure for the estimation of the probability $\mathbf{P}\left[(1/N)\sum_{i=1}^{N} X_i > a\right]$, which we used in Section III-B, to the case where the X_is are independent but not identically distributed. We do this in the following Lemma.

Lemma V.1: Consider the sum of N independent random variables X_i, $i = 1,2,\ldots,N$ that belong to J different classes. Let $N_j = \zeta_j N$ the number of random variables from

TABLE II
COMPARISON OF RESULTS FROM DIRECT MONTE CARLO SIMULATION AND QUICK SIMULATION FOR THE "STAR WARS" TRACE

	Direct simulation		Quick Simulation		
N	P_{loss}	D	P_{loss}	D	SU
20	$(5.42\pm0.24)\times10^{-2}$	2.0×10^{3}	$(5.52\pm0.20)\times10^{-2}$	1.0×10^{3}	2.0
40	$(7.82\pm0.39)\times10^{-3}$	1.0×10^{4}	$(7.79\pm0.36)\times10^{-3}$	2.0×10^{3}	5.0
60	$(1.21\pm0.06)\times10^{-3}$	5.0×10^{4}	$(1.21\pm0.06)\times10^{-3}$	3.0×10^{3}	16.7
80	$(1.99\pm0.10)\times10^{-4}$	2.6×10^{5}	$(1.99\pm0.09)\times10^{-4}$	4.0×10^{3}	65
100	$(3.32\pm0.17)\times10^{-5}$	1.4×10^{6}	$(3.20\pm0.15)\times10^{-5}$	5.0×10^{3}	280
120	$(5.66\pm0.28)\times10^{-6}$	7.4×10^{6}	$(5.59\pm0.27)\times10^{-6}$	5.0×10^{3}	1480

TABLE III
COMPARISON OF RESULTS FROM DIRECT MONTE CARLO SIMULATION AND QUICK SIMULATION FOR THE "SUPER BOWL" TRACE

	Direct simulation		Quick Simulation		
N	P_{loss}	D	P_{loss}	D	SU
20	$(6.72\pm0.31)\times10^{-2}$	3.0×10^{3}	$(6.85\pm0.26)\times10^{-2}$	1.0×10^{3}	3.0
40	$(7.14\pm0.35)\times10^{-3}$	2.4×10^{4}	$(7.32\pm0.31)\times10^{-3}$	2.0×10^{3}	12.0
60	$(8.95\pm0.45)\times10^{-4}$	1.6×10^{5}	$(8.53\pm0.36)\times10^{-4}$	3.0×10^{3}	53.3
80	$(1.06\pm0.05)\times10^{-4}$	1.1×10^{6}	$(1.05\pm0.05)\times10^{-4}$	4.0×10^{3}	275
100	$(1.32\pm0.07)\times10^{-5}$	7.8×10^{6}	$(1.29\pm0.06)\times10^{-5}$	4.0×10^{3}	1950
120	$(1.70\pm0.08)\times10^{-6}$	5.5×10^{7}	$(1.73\pm0.09)\times10^{-6}$	4.0×10^{3}	13750

TABLE IV
COMPARISON OF RESULTS FROM DIRECT MONTE CARLO SIMULATION AND QUICK SIMULATION FOR THE "ASTERIX" TRACE

	Direct simulation		Quick Simulation		
N	P_{loss}	D	P_{loss}	D	SU
40	$(1.20\pm0.06)\times10^{-2}$	6.0×10^{3}	$(1.21\pm0.05)\times10^{-2}$	2.0×10^{3}	3.0
60	$(1.96\pm0.10)\times10^{-3}$	2.9×10^{4}	$(1.96\pm0.09)\times10^{-3}$	3.0×10^{3}	9.7
80	$(3.39\pm0.17)\times10^{-4}$	1.4×10^{5}	$(3.20\pm0.15)\times10^{-4}$	4.0×10^{3}	33.8
100	$(5.77\pm0.29)\times10^{-5}$	6.8×10^{5}	$(5.49\pm0.26)\times10^{-5}$	6.0×10^{3}	113.3
120	$(9.53\pm0.47)\times10^{-6}$	3.4×10^{6}	$(9.13\pm0.45)\times10^{-6}$	7.0×10^{3}	485.7
140	$(1.60\pm0.08)\times10^{-6}$	1.9×10^{7}	$(1.67\pm0.08)\times10^{-6}$	8.0×10^{3}	2375

this collection that belong to class j, where $\zeta_j \in [0,1]$ and $\sum_{j=1}^{J}\zeta_j = 1$. Denote by $X_{\text{type\_}j}$ a random variable with distribution $F_{X_{\text{type\_}j}}$, the cumulative distribution function for every random variable in class j. Let Q the following twisted distribution of $X_1,\ldots,X_N$ where if X_i, $i=1,\ldots,N$, is of class j it is distributed according to

$$dF_{\tilde{X}_{\text{type\_}j}}(x) = \frac{e^{\theta^* x} dF_{X_{\text{type\_}j}}(x)}{\mathbf{E}\left[e^{\theta^* X_{\text{type\_}j}}\right]}$$

where θ^* is the optimal solution of the optimization problem

$$\lambda^*(a) = \sup_{\theta}[\theta a - \lambda(\theta)] \tag{17}$$

and

$$\lambda(\theta) = \sum_{j=1}^{J}\zeta_j \log \mathbf{E}\left[e^{\theta X_{\text{type\_}j}}\right]. \tag{18}$$

Let $\mathcal{H}$ the class of distributions of $X_1,\ldots,X_N$, where X_i, $i=1,\ldots,N$, are independent and all X_i that belong to class j are identically distributed. Then Q is the asymptotically optimal twisted distribution in $\mathcal{H}$ for estimating $\mathbf{P}\left[(1/N)\sum_{i=1}^{N}X_i > a\right]$, in the sense that it minimizes the speed factor in (4).

Proof: We will show that the exponentially twisted measure Q is a minimizer of the speed factor. Uniqueness can be shown along the lines of [9, Ch. VIII] and [10]. Without loss of generality assume that densities exist. Let $p_1(x_1)\cdots p_N(x_N)$ be the density of $X_1,\ldots,X_N$, that is, $p_i(x_i)$, $i=1,\ldots,N$, is the density corresponding to $F_{X_{\text{type\_}j}}$ if x_i is of class j. Let also R be a distribution in $\mathcal{H}$ with density $r_1(x_1)\cdots r_N(x_N)$. To compute the twisted estimator $\hat{\mathcal{L}}_R$ of $\mathbf{P}\left[(1/N)\sum_{i=1}^{N}X_i > a\right]$ we generate D i.i.d. "blocks" of data $\boldsymbol{\Xi}^{(1)},\ldots,\boldsymbol{\Xi}^{(D)}$. Each data block $\boldsymbol{\Xi}^{(i)}$, $i=1,\ldots,D$, consists of N random samples $x_1^{(i)},\ldots,x_N^{(i)}$, drawn from $r_1(x_1)\cdots r_N(x_N)$. For each data block i we form the sum

$$Y^{(i)} = \frac{1}{N}\sum_{k=1}^{N}x_k^{(i)}.$$

The twisted estimator $\hat{\mathcal{L}}_R$ of $\mathbf{P}\left[(1/N)\sum_{i=1}^{N}X_i > a\right]$ is given by

$$\hat{\mathcal{L}}_R = \frac{1}{D}\sum_{i=1}^{D}1_{(a,\infty)}(Y^{(i)})\frac{p_1\left(x_1^{(i)}\right)\cdots p_N\left(x_N^{(i)}\right)}{r_1\left(x_1^{(i)}\right)\cdots r_N\left(x_N^{(i)}\right)}.$$

The variance of $\hat{\mathcal{L}}_R$ is given by

$$\mathbf{Var}[\hat{\mathcal{L}}_R] = \frac{1}{D}\int 1_{(a,\infty)}(Y^{(i)})\frac{\left(p_1\left(x_1^{(i)}\right)\cdots p_N\left(x_N^{(i)}\right)\right)^2}{r_1\left(x_1^{(i)}\right)\cdots r_N\left(x_N^{(i)}\right)} \times dx_1^{(i)}\cdots dx_N^{(i)} - \frac{1}{D}\left(\mathbf{P}\left[\sum_{i=1}^{N}X_i > Na\right]\right)^2.$$

Letting $\Gamma(R)$ denote the first term in the right-hand side of the above multiplied by D and since the variance is nonnegative we have

$$\lim_{N\to\infty}\frac{1}{N}\log(\Gamma(R)) \geq \lim_{N\to\infty}\frac{1}{N}\log\left(\mathbf{P}\left[\sum_{i=1}^{N}X_i > Na\right]\right)^2 = -2\lambda^*(a) \tag{19}$$

where $\lambda^*(a)$ is defined in (17). The last equality above is obtained by noticing that $\lambda(\theta)$ in (18) is the limiting log-moment generating function of $\sum_{i=1}^{N}X_i$ and by applying the Gärtner–Ellis theorem [22].

We will next consider the twisted distribution Q and show that it achieves the bound in (19). To that end, we will reduce the heterogeneous case to the homogeneous one. To compute the twisted estimator $\hat{\mathcal{L}}_Q$ of $\mathbf{P}\left[(1/N)\sum_{i=1}^{N}X_i > a\right]$ we generate D i.i.d. "blocks" of data as before. Each data block $\Xi^{(i)}$, $i = 1,\ldots,D$, consists of N random samples $x_1^{(i)},\ldots,x_N^{(i)}$, N_j of which belong to class j for $j = 1,\ldots,J$ and are now drawn from $F_{\tilde{X}_{\text{type\_}j}}$. Let $x_{1,\text{type\_}j}^{(i)},\ldots,x_{N_j,\text{type\_}j}^{(i)}$ denote the class j samples in the ith block. For each data block i we form the sum

$$Y^{(i)} = \frac{1}{N}\sum_{k=1}^{N}x_k^{(i)} = \frac{1}{N}\sum_{j=1}^{J}\sum_{k=1}^{N_j}x_{k,\text{type\_}j}^{(i)}.$$

The twisted estimator $\hat{\mathcal{L}}_Q$ of $\mathbf{P}\left[(1/N)\sum_{i=1}^{N}X_i > a\right]$ is given by

$$\hat{\mathcal{L}}_Q = \frac{1}{D}\sum_{i=1}^{D}\left(1_{(a,\infty)}(Y^{(i)}) \mathbf{E}\left[e^{\theta^*\sum_{j=1}^{J}\sum_{k=1}^{N_j}X_{k,\text{type\_}j}^{(i)}}\right]e^{-\theta^*\sum_{j=1}^{J}\sum_{k=1}^{N_j}x_{k,\text{type\_}j}^{(i)}}\right)$$

where the random variables $X_{k,\text{type\_}j}^{(i)}$ are independent copies of $X_{\text{type\_}j}$.

Next note that

$$\mathbf{E}\left[e^{\theta^*\sum_{j=1}^{J}\sum_{k=1}^{N_j}X_{k,\text{type\_}j}^{(i)}}\right] = \prod_{j=1}^{J}\left(\mathbf{E}[e^{\theta^*X_{\text{type\_}j}}]\right)^{N_j} = e^{N\lambda(\theta^*)}$$

where $\lambda(\theta^*)$ was defined in (18). Consequently

$$\hat{\mathcal{L}}_Q = \frac{1}{D}\sum_{i=1}^{D}1_{(a,\infty)}(Y^{(i)})e^{N\lambda(\theta^*)}e^{-\theta^*\sum_{k=1}^{N}x_k^{(i)}}. \tag{20}$$

Consider now a random variable Z with log-moment generating function equal to $\lambda(\theta)$ and let F_Z be its distribution. Let also $\tilde{Z}$ be a random variable with the following exponentially twisted distribution

$$dF_{\tilde{Z}}(z) = \frac{e^{\theta^*z}dF_Z(z)}{e^{\lambda(\theta^*)}}.$$

Suppose we are interested in estimating $\mathbf{P}\left[(1/N)\sum_{i=1}^{N}X_i > a\right]$ where $X_1,\ldots,X_N$ are i.i.d. and drawn from F_Z. The key observation is that this latter probability has a twisted estimator equal to $\hat{\mathcal{L}}_Q$ in (20).

TABLE V
MPEG SEQUENCES USED IN THE SIMULATION EXPERIMENT WITH HETEROGENEOUS SOURCES

MPEG sequence	mean bit rate (in Mbits/GOP)	max bit rate (in Mbits/GOP)	N_j/N
Star Wars	0.201	0.495	1/20
Asterix	0.257	0.765	1/10
James Bond: Goldfinger	0.248	0.727	3/20
Super Bowl	0.258	0.628	3/20
Jurassic Park	0.157	0.421	1/4
German Talk Show	0.177	0.471	3/10

Namely, $\hat{\mathcal{L}}_Q$ in (20) is also the twisted estimator obtained from D i.i.d. "blocks" of data $\Xi^{(1)},\ldots,\Xi^{(D)}$ where each block i consists of N i.i.d. samples $x_1^{(i)},\ldots,x_N^{(i)}$ drawn from $F_{\tilde{Z}}$. Note also that by construction $\sum_{i=1}^{N}X_i$ has the same distribution with $\sum_{j=1}^{J}\sum_{k=1}^{N_j}X_{k,\text{type\_}j}$ since they have the same moment generating function. Thus, the results established in [9, Ch. VIII] and [10] for the i.i.d. case apply to $\hat{\mathcal{L}}_Q$. In particular, the second moment $\Gamma(Q)$ of $D\hat{\mathcal{L}}_Q$ satisfies

$$\lim_{N\to\infty}\frac{1}{N}\log(\Gamma(Q)) = -2\lambda^*(a).$$

Considering (19) it follows that Q minimizes the speed factor in (4). ■

Using the same argument as in the case of homogeneous sources we can apply the previous lemma to derive the change of measure for the probability distribution of $\boldsymbol{\phi}$. If source i is a source of type j, then

$$q_i(\phi_i) = p_i(\phi_i)\frac{e^{\theta_\tau^*S_{-\tau,0}^{A,\text{type\_}j}(\phi_i)}}{\mathbf{E}\left[e^{\theta_\tau^*S_{-\tau,0}^{A,\text{type\_}j}(\phi_i)}\right]}. \tag{21}$$

Finally, the change of measure $r(\boldsymbol{\xi})$ can be obtained as in (15) or (16) depending on whether the time is continuous or discrete. Note that while the twisted distribution is different for each type of source, the parameters τ and θ_τ^* are identical for all types.

B. Numerical Results

We next report numerical results for the case of heterogeneous sources. In the following experiment sources of six different types are being multiplexed. The sequence f_j is the number of Mbits/GOP for the first 1000 GOPs of a different MPEG trace for each source type as summarized in Table V.

The parameters used in this simulation are $K = 1000$ GOPs, $s = 0.22$ Mbits/GOP, and $b = 0.6$ Mbits which results to $\tau = 99$ GOPs and $\theta_\tau^* = 0.0784$. In Table VI we compare the estimates obtained from the direct Monte Carlo simulation and the quick simulation with the change of measure $r(\cdot)$ in (16). The quantities D and SU have the same meaning as in the previous sections. The running times for the last raw of the table were 2 min 6 s for the quick simulation versus 4 days 4 h 4 min for the direct Monte Carlo simulation.

VI. APPLICATIONS IN QOS PROVISIONING

In this section we discuss how our quick simulation technique can be used to facilitate network resource planning and real-time *Connection Admission Control (CAC)*.

TABLE VI
COMPARISON OF RESULTS FROM THE DIRECT MONTE CARLO SIMULATION AND THE QUICK SIMULATION WHEN THE TRAFFIC MODEL IS BASED ON RANDOMLY STARTED HETEROGENEOUS MPEG SEQUENCES

	Direct simulation		Quick Simulation		
N	P_{loss}	D	P_{loss}	D	SU
20	$(6.60\pm0.27)\times10^{-2}$	3.0×10^{3}	$(6.28\pm0.29)\times10^{-2}$	1.0×10^{3}	3.0
40	$(8.42\pm0.41)\times10^{-3}$	1.8×10^{4}	$(8.36\pm0.35)\times10^{-3}$	3.0×10^{3}	6.0
60	$(1.26\pm0.06)\times10^{-3}$	1.0×10^{5}	$(1.22\pm0.06)\times10^{-3}$	3.0×10^{3}	33.3
80	$(1.85\pm0.09)\times10^{-4}$	6.1×10^{5}	$(1.84\pm0.09)\times10^{-4}$	4.0×10^{3}	152.5
100	$(2.75\pm0.14)\times10^{-5}$	3.5×10^{6}	$(2.91\pm0.14)\times10^{-5}$	5.0×10^{3}	700
120	$(4.29\pm0.21)\times10^{-6}$	2.0×10^{7}	$(4.32\pm0.20)\times10^{-6}$	6.0×10^{3}	3333

We first discuss applications in network resource planning. Consider, for example, the provider of a video-on-demand service. The service is provided by a video server accessed through a fixed bandwidth line leased from a backbone provider. The amount of leased bandwidth is renegotiated on a weekly basis. The server operator uses statistical multiplexing to make maximum use of the leased bandwidth and needs to predict its next week needs for bandwidth. The server contains a certain number of MPEG-coded video segments (e.g., recently released movies) and the set of available movies is constant for the week in question and known in advance. Users can start viewing a movie of their choice at any time. They can also pause, replay a scene, or fast forward the movie. The quick simulation scheme presented in this paper can be used by the server operator to calculate the buffer overflow probability at the access link interface in a number of scenarios with different mix of movies (based on forecasts of expected demand) and bandwidth values. Computational efficiency (i.e., fast simulation times) is of essence since it allows the operator to quickly explore a multitude of scenarios and perform a cost/utility analysis in order to come up with the optimal amount of bandwidth to be leased for the following week. The provider of the video-on-demand service can actually be seen as a special case of an *Application Service Provider (ASP)*. It follows, that our quick simulation can be used by ASPs to assess loss probabilities at their access link and optimize its capacity.

We next turn our attention to applications in CAC. To use quick simulation as a tool to support real-time CAC providing QoS guarantees, the estimation time for buffer overflow probabilities has to be within the limits of acceptable session setup times (let us say 10 s). To get an idea of how close we are to this objective, we considered the video server scenario of the previous paragraph with the following parameters. The server supports 16 movies with a duration of 27.5 min or 3300 GOPs (our entire data set!). We use quick simulation to estimate the buffer overflow probability when 100 video streams are multiplexed through a fixed bandwidth ($c = 50.4$ Mbits/s) buffered link ($U = 15$ Mbits). The 100 streams are drawn from the 16 movies according to the following vector of N_js: $[14, 11, 10, 8, 7, 6, 6, 5, 5, 5, 5, 4, 4, 4, 3, 3]$ (we use the notation of Section V). In order to further expedite the calculation of the buffer overflow probabilities we encoded the initialization phase routines in C and relaxed the simulation accuracy requiring that half the size of the 95% confidence interval be less than 10% of the estimated P_{loss}. With these modifications, the total calculation time for $P_{\text{loss}} = 2.57 \times 10^{-6}$, on the same PC used for all the experiments in this paper, was about 50 s. Further speed improvements could be achieved with code optimization and the use of a faster PC. This suggests that, depending on the problem size (number of sources and length of movies), running times on general purpose hardware are on the same order of magnitude with desired session initialization times. As a result, the use of quick simulation in real-time CAC might be feasible. If one assumes that the movie mix and available bandwidth evolve in a slower time-scale than connection requests (e.g., on the order of several minutes), it is certainly feasible to use the quick simulation scheme to precompute loss probabilities in a number of scenarios and extrapolate from those to make real-time CAC decisions. The speed of quick simulation allows for very frequent updates of those precomputed scenarios. Finally, note that running the simulations on dedicated hardware could reduce execution times by at least a factor of 100 and make simulation-based real-time CAC feasible for a wide range of practical situations. This could never be possible without the use of quick simulation.

VII. CONCLUSIONS

The importance sampling technique presented in this paper can be used to speed up the estimation of very small buffer overflow probabilities via simulation. It pertains to a model of a buffered communication multiplexer fed by a large number of independent sources that generate traffic according to a periodic function with a random phase. The sources can be homogeneous or heterogeneous. This traffic model accommodates a wide range of situations of practical interest, including ON–OFF traffic models and sequences of bit rates generated by actual Variable Bit Rate sources, such as MPEG video compressors. Extensive numerical results demonstrate dramatic savings in computational time compared with direct Monte Carlo simulation. This makes quick simulation applicable in network resource planning and real time CAC.

Trace-based simulation can also be applied to investigate buffered multiplexing of arrival processes with known stochastic models. In particular, one can generate sufficiently long sample paths of arrival processes and use our quick simulation method to simulate the system based on these paths. This approach will work even in the case that sources do not generate traffic according to the same stochastic model (heterogeneous case). The alternative, is to develop an appropriate change of measure for each arrival traffic model of interest which might be hard to do. It should be noted that the results obtained by simulation based on given traces of the stochastic processes involved are not necessarily identical to the results that we would

obtain if we could use the "real" arrival stochastic processes to drive the simulation. In the case of homogeneous sources, we run the risk of generating dependent streams when the starting points are not sufficiently separated (as discussed in [16]). But even when the sources are heterogeneous, a finite realization might or might not be a good representative of the process itself. However, if we have no way of knowing the "true" process this is the best we can do and a common practice in the related literature. The approximation of the "true" buffer overflow probability improves as the length of the finite realization and the number of multiplexed sources become larger.

ACKNOWLEDGMENT

The authors would like to thank A. Weiss and D. Mitra for suggesting this problem and for many helpful discussions, as well as, for their hospitality at Bell Labs. This work started when the first author was visiting the Mathematics of Networks and Systems Department at Bell Labs, Lucent Technologies. The MPEG traces were taken from O. Rose's data set. They can be found at http://www-info3.informatik.uni-wuerzburg.de/MPEG.

REFERENCES

[1] I. Paschalidis and S. Vassilaras, "Quick simulation of a queue fed by arbitrary traffic traces," in *Proc. 39th Annu. Allerton Conf. Communication, Control, and Computing*, Monticello, IL, Oct. 2001, pp. 172–181.

[2] I. Ch. Paschalidis, "Large deviations in high speed communication networks," Ph.D. dissertation, Massachusetts Institute of Technology, Cambridge, MA, 1996.

[3] ——, "Class-specific quality of service guarantees in multimedia communication networks," *Automatica (Special Issue on Control Methods for Communication Networks)*, vol. 35, pp. 1951–1968, 1999.

[4] A. Weiss, "An introduction to large deviations for communication networks," *IEEE J. Select. Areas Commun.*, vol. 13, pp. 938–952, Aug. 1995.

[5] D. Bertsimas, I. Ch. Paschalidis, and J. N. Tsitsiklis, "Asymptotic buffer overflow probabilities in multiclass multiplexers: An optimal control approach," *IEEE Trans. Automat. Contr.*, vol. 43, pp. 315–335, Mar. 1998.

[6] J. Hammersley and D. Handscomb, *Monte Carlo Methods*. London, U.K.: Methuen, 1964.

[7] P. Glynn and D. Iglehart, "Importance sampling for stochastic simulations," *Manage. Sci.*, vol. 35, pp. 1367–1392, 1989.

[8] P. Heidelberger, "Fast simulation of rare events in queueing and reliability models," *ACM Trans. Modeling Comp. Simulation*, vol. 5, no. 1, pp. 43–85, 1995.

[9] J. A. Bucklew, *Large Deviation Techniques in Decision, Simulation, and Estimation*. New York: Wiley, 1990.

[10] J. Bucklew, P. Ney, and J. Sadowsky, "Monte Carlo simulation and large deviations theory for uniformly recurrent Markov chains," *J. Appl. Prob.*, vol. 27, pp. 44–59, 1990.

[11] J. Sadowsky, "On the optimality and stability of exponential twisting in Monte Carlo estimation," *IEEE Trans. Inform. Theory*, vol. 39, pp. 119–128, Jan. 1993.

[12] J. Townsend, Z. Harastzi, J. Freebersyser, and M. Devetsikiotis, "Simulation of rare events in communication networks," *IEEE Commun. Mag.*, vol. 36, pp. 36–41, Aug. 1998.

[13] S. Parekh and J. Walrand, "A quick simulation method for excessive backlogs in networks of queues," *IEEE Trans. Automat. Contr.*, vol. 34, pp. 54–66, Jan. 1989.

[14] C. Chang, P. Heidelberger, S. Juneja, and P. Shahabuddin, "Effective bandwidth and fast simulation of ATM in tree networks," *Perform. Eval.*, vol. 20, pp. 45–65, 1994.

[15] C. Huang, M. Devetsikiotis, I. Lambadaris, and A. Kaye, "Modeling and simulation of self-similar variable bit rate compressed video: A unified approach," presented at the ACM SIGCOMM Conf., Cambridge, MA, Aug. 1995.

[16] M. Krunz and A. Makowski, "Modeling video traffic using M/G/∞ input processes: A compromise between Markovian and LRD models," *IEEE J. Select. Areas Commun.*, vol. 16, pp. 733–748, June 1998.

[17] B. Melamed and D. Pendarakis, "Modeling full-length VBR video using Markov-renewal-modulated TES models," *IEEE J. Select. Areas Commun.*, vol. 16, pp. 600–611, June 1998.

[18] E. Knightly and N. Shroff, "Admission control for statistical QoS: Theory and practice," *IEEE Network*, vol. 13, pp. 20–29, Mar. 1999.

[19] D. Botvich and N. Duffield, "Large deviations, economies of scale, and the shape of the loss curve in large multiplexers," *Queueing Syst.*, vol. 20, pp. 293–320, 1995.

[20] D. Mitra and J. Morrison, "Multiple time scale regulation and worst case processes for ATM network control," in *Proc. IEEE 34th Conf. Decision and Control*, New Orleans, LA, Dec. 1995, pp. 353–358.

[21] C.-S. Chang, Y.-M. Chiu, and W. Song, "On the performance of multiplexing independent regulated inputs," *Proc. ACM Sigmetrics 2001/Performance 2001*, vol. 29, no. 1, pp. 184–193, 2001.

[22] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications*, 2nd ed. New York: Springer-Verlag, 1998.

[23] A. Shwartz and A. Weiss, *Large Deviations for Performance Analysis*. New York: Chapman & Hall, 1995.

[24] H. Cramér, "Sûr un nouveau théorème-limite de la théorie des probabilités," in *Actualités Scientifiques et Industrielles*. Paris, France: Hermann, 1938, vol. 736, Colloque consacré à la théorie des probabilités, pp. 5–23.

[25] R. Rockafellar, *Convex Analysis*. Princeton, NJ: Princeton Univ. Press, 1970.

Ioannis Ch. Paschalidis (M'96) was born in Athens, Greece, in 1968. He received the Diploma degree in electrical and computer engineering from the National Technical University of Athens, Athens, Greece, in 1991, and the S.M. and Ph.D. degrees in electrical engineering and computer science from the Massachusetts Institute of Technology (MIT), Cambridge, MA, in 1993 and 1996, respectively.

After a brief period as a Postdoctoral Associate at the Laboratory for Information and Decision Systems (LIDS), MIT, he joined Boston University, Boston, MA, in September 1996 where he currently is Associate Professor of Manufacturing Engineering. During his sabbatical in 2003, he held visiting appointments with LIDS, MIT, and the Columbia University Business School. His current research interests include the analysis and control of stochastic systems, large deviations theory, queueing theory, optimization, pricing, and revenue management. The main application areas he is targeting include communication and sensor networks, manufacturing systems, and supply chains.

Dr. Paschalidis has received a National Science Foundation CAREER Award (2000), and the second prize in the 1997 George E. Nicholson paper competition by INFORMS, and was an invited participant at the 2002 Frontiers of Engineering Symposium, organized by the National Academy of Engineering. He has served in the program committees of several conferences, including the INFORMS Applied Probability Conference, the IEEE Conference on Decision and Control, and the INFOCOM. He is an Associate Editor of *Operations Research Letters* and of *Automatica*.

Spyridon Vassilaras was born in Athens, Greece, in 1971. He received the Diploma degree in electrical and computer engineering from the National Technical University of Athens, Athens, in 1995, and the M.S. degree in computer systems engineering and the Ph.D. degree in computer engineering from Boston University, Boston, MA, in 1997 and 2002, respectively.

He is currently a researcher with the Athens Information Technology, Athens, a recently founded academic institution dedicated to research and graduate education. His research interests include performance analysis of telecommunication networks, stochastic modeling and simulation, large deviations theory, nonlinear optimization, and information security.