Loss Probability Calculations and Asymptotic Analysis for Finite Buffer Multiplexers

Han S. Kim and Ness B. Shroff, *Senior Member, IEEE*

***Abstract*—In this paper, we propose an approximation for the loss probability, $P_L(x)$, in a finite buffer system with buffer size x. Our study is motivated by the case of a high-speed network where a large number of sources are expected to be multiplexed. Hence, by appealing to *Central Limit Theorem* type of arguments, we model the input process as a general Gaussian process. Our result is obtained by making a simple mapping from the tail probability in an infinite buffer system to the loss probability in a finite buffer system. We also provide a strong asymptotic relationship between our approximation and the actual loss probability for a fairly large class of Gaussian input processes. We derive some interesting asymptotic properties of our approximation and illustrate its effectiveness via a detailed numerical investigation.**

***Index Terms*—Asymptotic relationship, loss probability, queue length distribution, maximum variance asymptotic.**

I. Introduction

LOSS PROBABILITY is an important quality of service (QoS) measure in communication networks. While the overflow probability, or the tail of the queue length distribution, in an infinite buffer system has been extensively studied [1]–[7], there have been relatively few studies on the loss probability in finite buffer systems [8]–[11].

In this paper, we propose a simple method to estimate the loss probability $P_L(x)$ in a finite buffer system from the tail of the queue length distribution (or *tail probability*) $\mathbb{P}\{Q > x\}$ of an infinite buffer system. We estimate $P_L(x)$ by making a simple mapping from $\mathbb{P}\{Q > x\}$. Hence, we consider both a finite buffer queueing system and an infinite buffer queueing system. We model both systems by a discrete-time fluid queue consisting of a server with constant rate c and a fluid input λ_n. Both queues are fed with the same input. Let $\hat{Q}_n$ and Q_n denote the queue length in the finite queue and in the infinite queue at time n, respectively. We assume that λ_n is stationary and ergodic and that the system is stable, i.e., $\mathbb{E}\{\lambda_n\} < c$. Under this assumption, it has been shown that Q_n converges to a stationary and ergodic process [12]. It has also been shown that $\hat{Q}_n$ converges to a stationary process when the system is a GI/GI/m/x type of queue [13], [14], and when the system is a G/M/m/x type of queue [15]. Since proving the convergence of $\hat{Q}_n$ is not the focus of this paper, and moreover, practical measurements of $P_L(x)$ and $\mathbb{P}\{Q > x\}$ are based on "time averaging" assuming ergodicity [see (1) and (2)], we assume that both $\hat{Q}_n$ and Q_n started at $n = -\infty$, and that they are ergodic and stationary.[1] The time index n is often omitted to represent the stationary distribution, i.e., $\mathbb{P}\{Q > x\} = \mathbb{P}\{Q_n > x\}$ and $\mathbb{E}\{\lambda\} = \mathbb{E}\{\lambda_n\}$.

The loss probability, $P_L(x)$, for a buffer size x is defined as the long-term ratio of the amount of fluid lost to the amount of fluid fed. It is expressed as

$$P_L(x) = \lim_{N\to\infty} \frac{\sum_{k=1}^{N} \left(\hat{Q}_{k-1} + \lambda_k - c - x\right)^+}{\sum_{k=1}^{N} \lambda_k} = \frac{\mathbb{E}\left\{\left(\hat{Q}_{n-1} + \lambda_n - c - x\right)^+\right\}}{\mathbb{E}\{\lambda_n\}} \tag{1}$$

where $(x)^+$ denotes $\max\{x, 0\}$, and where the second equality is due to the ergodicity assumption. The tail probability (or tail of the queue length distribution, also sometimes called the *overflow probability*) $\mathbb{P}\{Q > x\}$ is defined as the amount of time the fluid in the infinite buffer system spends above level x divided by the total time. It is expressed as

$$\mathbb{P}\{Q > x\} = \lim_{N\to\infty} \frac{1}{N} \sum_{k=1}^{N} I(Q_k > x) \tag{2}$$

where $I(A) = 1$ if A is true; $I(A) = 0$ otherwise. From now on, when we write "loss probability" it will only be in the context of a finite buffer system, and when we write "tail probability" it will only be in the context of an infinite buffer system. Note that since $\mathbb{P}\{Q > x\}$ is averaged by time, and $P_L(x)$ is averaged by the input, in general there is no relationship between these two quantities. However, $P_L(x)$ is often approximated as

$$P_L(x) \approx \mathbb{P}\{Q > x\}. \tag{3}$$

This approximation usually provides an upper bound (sometimes a very poor bound) to the loss probability, although in general this cannot be proven, and in fact counterexamples can easily be constructed. What we have learned from simulation studies is that the curves $P_L(x)$ versus x and $\mathbb{P}\{Q > x\}$ versus x exhibit a similar shape (e.g., see Fig. 1), which motivates this

Manuscript received March 23, 2000; revised December 12, 2000; recommended by IEEE/ACM TRANSACTIONS ON NETWORKING Editor J. Liebeherr.

The authors are with the School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN 47907 USA (e-mail: shroff@purdue.edu).

Publisher Item Identifier S 1063-6692(01)10551-0.

[1]We refer the interested reader to our technical report [17], where we have studied the relationship between finite and infinite buffer queues without assuming ergodicity of $\hat{Q}_n$ and derived similar asymptotic results to (22) in this paper. However, this involves mathematical technicalities that take away from the main message in this paper, i.e., developing a simple approximation for the loss probability.

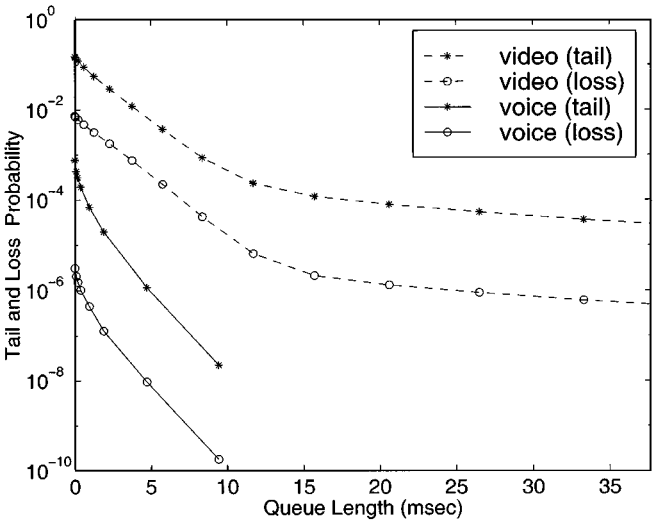


Fig. 1. Comparison of loss and tail curves in a 45-Mb/s link where 2900 voice sources or 69 MPEG video sources are multiplexed: Loss and tail curves seem to have the same shape.

work. Further, it has been shown in [16] that for M/Subexponential/1 and GI/Regularly-varying/1 with independent identically distributed (i.i.d.) interarrival times and i.i.d. service times, $\mathbb{P}\{Q > x\}/P_L(x)$ converges to a constant, as $x \to \infty$.

Hence, it seems reasonable that if we have a good estimate of the tail probability $\mathbb{P}\{Q > x\}$ and a way to calculate $P_L(a)$, the loss probability for some buffer size a, then we can calculate the loss probability $P_L(x)$ as

$$P_L(x) = \frac{P_L(a)}{\mathbb{P}\{Q > a\}} \mathbb{P}\{Q > x\}. \tag{4}$$

In particular, we will choose $a = 0$ because this allows us to compute the loss probability $[P_L(0)]$ quite easily. *This is the basic idea that drives this paper*. In addition to developing a methodology to calculate the loss probability, we will also show that asymptotically the loss probability and the tail probability curves are quite similar, and if they diverge, they do so slowly, which is an interesting result by itself.

For our study in this paper, we focus on the case when the aggregate traffic can be characterized by a stationary Gaussian process. Recently, Gaussian processes have received significant attention as good models for the arrival process to a high-speed multiplexer [3], [18]–[23]. There are many reasons for this. Due to the huge link capacity of high-speed networks, hundreds or even thousands of network applications are likely to be served by a network multiplexer. Also, when a large number of sources are multiplexed, characterizing the input process with traditional Markovian models results in computational infeasibility problems [24] that are not encountered for Gaussian processes. Finally, recent network traffic studies suggest that certain types of network traffic may exhibit self-similar or more generally asymptotic self-similar type of long-range dependence [25], [26], and various Gaussian processes can be used to model this type of behavior. Hence, our motivation to study the case when the input process λ_n can be characterized by a Gaussian process.

This paper is organized as follows. In Section II, we review the maximum variance asymptotic (MVA) results for the infinite buffer queue, and then demonstrate how to obtain similar results for the loss probability. Then, we compare our approach to an approach based on the *many-sources asymptotics*. In Section III, we validate our result with several numerical examples, including those for self-similar/long-range dependent traffic. In Section IV, we find the asymptotic relationship between the loss probability and our approximation. In Section V, we describe the applicability of our approximation for on-line traffic measurements. We finally state the conclusions in Section VI.

II. Maximum Variance Asymptotic (MVA) Approximation for Loss

Remember that the first component in our development of an approximation for $P_L(x)$ is to find a good estimate of $\mathbb{P}\{Q > x\}$. Fortunately, this part of the problem has already been solved in [20], [21], [27]. By developing results based on Extreme Value Theory, it has been found that the MVA approach (first named in [20]) provides an accurate estimate of the tail probability. We briefly review it here. As mentioned before, we focus on the case when the aggregate traffic can be characterized by a Gaussian process, hence λ_n, the input process to the queue is Gaussian. Let $\overline{\lambda} := \mathbb{E}\{\lambda_n\}$ and $\sigma^2 := \mathrm{Var}\{\lambda_n\}$.

The queue length Q_n (or workload) at time n in the infinite buffer system is expressed by Lindley's equation:

$$Q_n = (Q_{n-1} + \lambda_n - c)^+. \tag{5}$$

We define a stochastic process X_n as

$$X_n := \sum_{k=1}^{n} \lambda_k - cn. \tag{6}$$

We assume that λ_n is stationary and ergodic and that the system is stable, i.e., $\mathbb{E}\{\lambda_n\} < c$. Then, it has been shown that the distribution of Q_n converges to the steady state distribution as $n \to \infty$ and that the supremum distribution of X_n is the steady state queue distribution [12]:

$$\mathbb{P}\{Q > x\} = \mathbb{P}\left\{\sup_{n \geq 1} X_n > x\right\}. \tag{7}$$

Let $C_\lambda(l)$ be the autocovariance function of λ_n. Then, the variance of X_n can be expressed in terms of $C_\lambda(l)$. For each $x > 0$, define the normalized variance $\sigma^2_{x,n}$ of X_n as

$$\sigma^2_{x,n} := \frac{\mathrm{Var}\{X_n\}}{(x - \mathbb{E}\{X_n\})^2} = \frac{nC_\lambda(0) + 2\sum_{l=1}^{n-1}(n-l)C_\lambda(l)}{(x + \kappa n)^2} \tag{8}$$

where $\kappa := c - \overline{\lambda}$. Let m_x be the reciprocal of the maximum of $\sigma^2_{x,n}$ for given x, i.e.,

$$m_x := \frac{1}{\max_{n \geq 1} \sigma^2_{x,n}} = \min_{n \geq 1} \frac{(x + \kappa n)^2}{\mathrm{Var}\{X_n\}} \tag{9}$$

and we define n_x to be the time n at which the normalized variance $(\mathrm{Var}\{X_n\}/(x + \kappa n)^2)$ is maximized. Although the esti-

mate $e^{-(m_x/2)}$ called the MVA approximation has been theoretically shown to be only an *asymptotic* upper bound, simulation studies in different papers have shown that it is an accurate approximation even for small values of x [27], [18], [20], [28].

Now, for some a, we need to evaluate the ratio $P_L(a)/\mathbb{P}\{Q > a\}$ given in (4). As mentioned earlier, it is easy to find $P_L(a)$ for $a = 0$, hence what we need to do is to first estimate $\mathbb{P}\{Q > 0\}$ from the MVA result. For a given x both n_x and m_x in the MVA approximation cannot generally be obtained in a simple closed form, hence, search algorithms[2] are likely to be used to evaluate them, but n_x may not be unique especially for a small value of x. However, when $x = 0$, we can obtain them right away, as demonstrated in the following proposition.

Proposition 1: Let n_x be the value of n at which $\sigma^2_{x,\,n}$ attains its maximum $\langle\sigma^2_x\rangle$. Then, $n_0 = 1$ and

$$m_0 = \frac{\kappa^2}{C_\lambda(0)}. \tag{10}$$

Proof of Proposition 1: To prove the proposition, it suffices to show that

$$\sup_{n\geq 1} \sigma^2_{0,\,n} = \sigma^2_{0,\,1} = \frac{C_\lambda(0)}{\kappa^2}. \tag{11}$$

Since $C_\lambda(0) \geq C_\lambda(n)$, for all $n \geq 1$

$$\begin{aligned}\sigma^2_{0,\,n} &= \frac{1}{(\kappa n)^2}\left[nC_\lambda(0) + 2\sum_{m=1}^{n-1}(n-m)C_\lambda(m)\right]\\ &\leq \frac{1}{(\kappa n)^2}\left[nC_\lambda(0) + 2\sum_{m=1}^{n-1}(n-m)C_\lambda(0)\right]\\ &= \frac{n^2 C_\lambda(0)}{(\kappa n)^2}\\ &= \frac{C_\lambda(0)}{\kappa^2}.\end{aligned} \tag{12}$$

Since $\sigma^2_{0,\,1} = (C_\lambda(0)/\kappa^2)$, we have (11). ∎

Now, we show how to calculate $P_L(0)$. Since λ_n is assumed Gaussian, the mean and the variance provide sufficient information to calculate $P_L(0)$, i.e.

$$\begin{aligned}P_L(0) &= \frac{\mathbb{E}\{(\lambda_n - c)^+\}}{\mathbb{E}\{\lambda_n\}}\\ &= \frac{1}{\overline{\lambda}\sqrt{2\pi}}\int_c^\infty (r-c)e^{-((r-\overline{\lambda})^2/2\sigma^2)}\,dr\end{aligned} \tag{13}$$

where $\overline{\lambda} := \mathbb{E}\{\lambda_n\}$. As long as the number of input sources is large enough for the aggregate traffic to be characterized as a Gaussian process, (13) gives an accurate estimate (exact for a Gaussian input) and is often called the Gaussian approximation [29]. Note that $C_\lambda(0) = \sigma^2$ and $\kappa = c - \overline{\lambda}$ in (10). From (4), (10), and (13), we have

$$P_L(x) \approx \frac{P_L(0)}{e^{-(m_0/2)}}\,e^{-(m_x/2)} = \alpha e^{-(m_x/2)} \tag{14}$$

[2]Simple local search algorithms starting at $(\beta x/(2-\beta)\kappa)$ are good enough to find n_x within a small number of iterations.

where

$$\alpha = \frac{1}{\overline{\lambda}\sqrt{2\pi}}\exp\left(\frac{(c-\overline{\lambda})^2}{2\sigma^2}\right)\int_c^\infty (r-c)\exp\left(-\frac{(r-\overline{\lambda})^2}{2\sigma^2}\right)dr.$$

We call this above approximation the *MVA approximation for loss*. The MVA approach is based on the *large buffer asymptotics* and it also applies in the context of the *many-sources asymptotics* [20], [28]. We next compare this approach with an approximation based on the many-sources asymptotics.

The many-sources asymptotics have been widely studied and can be found in many papers on queueing analysis using large-deviation technique [5], [30]–[32]. Most of the papers deal with the tail distribution rather than the loss probability. In [9], the authors developed the first result on the loss probability based on the many-sources asymptotics. We call this the Likhanov–Mazumdar (L–M) approximation for loss. Since the L–M result was obtained for a fairly general class of arrival processes and is much stronger than typical large-deviation types of results, we feel that it is important to compare our result with the L–M result.

Consider N i.i.d. sources, each with input rate $\lambda_n^{(i)}$, $n \in \{0, 1, 2, \ldots\}$, $i \in \{1, 2, \ldots, N\}$. It is assumed that the moment generating function of $\lambda_n^{(1)}$ exists, and that the input rate $\lambda_n^{(i)}$ is bounded. The L–M approximation has the following form:[3]

$$P_L(NB) \approx \frac{e^{-NI_{\hat{n}}(C,B)}}{\theta^2_{\hat{n}}\overline{\lambda}_1\sqrt{2\pi\sigma^2_{\hat{n}}N^3}} \tag{15}$$

and is theoretically justified by

$$P_L(NB) = \frac{e^{-NI_{\hat{n}}(C,B)}}{\theta^2_{\hat{n}}\overline{\lambda}_1\sqrt{2\pi\sigma^2_{\hat{n}}N^3}}(1+O(1/N)) \tag{16}$$

where N is the number of sources, NC is the link capacity, NB is the buffer size, $\overline{\lambda}_1 = \mathbb{E}\{\lambda_n^{(1)}\}$, $\phi_n(\theta) = \mathbb{E}\left\{e^{\theta\sum_{k=1}^n \lambda_k^{(1)}}\right\}$, θ_n is a value of θ such that

$$\begin{aligned}\frac{\phi_n'(\theta)}{\phi_n(\theta)} &= Cn + B,\\ \sigma_n^2 &= \frac{\phi_n''(\theta_n)}{\phi_n(\theta_n)} - (Cn+B)^2,\\ I_n(C,B) &= (Cn+B)\theta_n - \log\phi_n(\theta_n)\end{aligned}$$

and $\hat{n}$ is a value of n that maximizes $I_n(C, B)$, for a given C and B. This approximation (15) becomes exact as $N \to \infty$.

Consider the numerical complexity of (16). Suppose that we calculate (16) for given N, C, B, and $\lambda_n^{(1)}$. In general, since there are no closed-form solutions for θ_n and $\hat{n}$, we have to find them numerically. Two iteration loops are nested. The inner loop iterates over θ to find θ_n for given n, and the outer loop iterates over n to find $\hat{n}$. Hence, it can take a long time to find a solution of (16) by numerical iteration. However, the MVA approximation requires only a one-dimensional iteration over n to find n_x at which m_x is minimized.

[3]This expression is just a rewriting of equation (2.6) in [9].

There is another problem in applying the L–M approximation for control based on on-line measurements. When the distribution of a source is not known beforehand, in the L–M approach the moment generating function of a source should be evaluated for the two-dimensional arguments (θ, n), whereas only the first two moments are evaluated for the one argument, n, in the MVA approach (see Section V).

Note that one could avoid the above problems by making a Gaussian approximation on the aggregate source first, and then using the L–M approximation given by (16). Specifically, if we assume that the input process is Gaussian, we have a closed-form solution for θ_n, i.e., as $\phi_n(\theta) = e^{\theta m(n)+(1/2)\theta^2 v(n)}$ with $m(n) = \mathbb{E}\{\sum_{k=1}^{n} \lambda_k^{(1)}\}$ and $v(n) = \mathrm{Var}\{\sum_{k=1}^{n} \lambda_k^{(1)}\}$, we have

$$\theta_n = \frac{Cn + B - m(n)}{v(n)}. \tag{17}$$

Hence, for given C and B, both $I_n(C, B)$ and $\sigma^2_{NB,n}$ (the normalized variance of X_n) are expressed in terms of n, $m(n)$, and $v(n)$, we can avoid the two-dimensional evaluation of the moment generating function.

The only problem is that the theoretical result that says that the L–M approximation in (16) becomes exact as the number of sources N becomes large is not proven for unbounded (e.g., Gaussian) inputs. Still, since making this approximation reduces the complexity of the search space, it would be instructive to also investigate the performance of such an approximation. In Section III, we will numerically investigate our MVA approximation for loss, the L–M approximation, and some other approximations developed in the literature.

III. Numerical Validation of the MVA Approximation for Loss

In this section, we investigate the accuracy of the proposed method by comparing our technique with simulation results. In all our simulations, we have obtained 95% confidence intervals. However, to not clutter the figures, the error bars are only shown in the figures when they are larger than $\pm 20\%$ of the estimated probability. To improve the reliability of the simulation, we use *importance-sampling* [33] whenever applicable.[4] We have attempted to systematically study the MVA approximation for various representative scenarios. For example, we begin our investigation with Gaussian input processes. Here, we only check the performance of our approximation (we do not compare with other approximations in the literature), since other approximations are not developed for Gaussian inputs. We then consider non-Gaussian input sources and compare our MVA approximation for loss with other approximations in the literature. Specifically, we consider Markoff Modulated Fluid (MMF) sources which have been used as representative of voice traffic in many different papers (e.g., [34], [35]) and also consider JPEG and MPEG video sources that have been used in other papers in the literature (e.g., [20], [36]).

[4]For interested readers, the software used for the analysis and simulation will be available upon request.

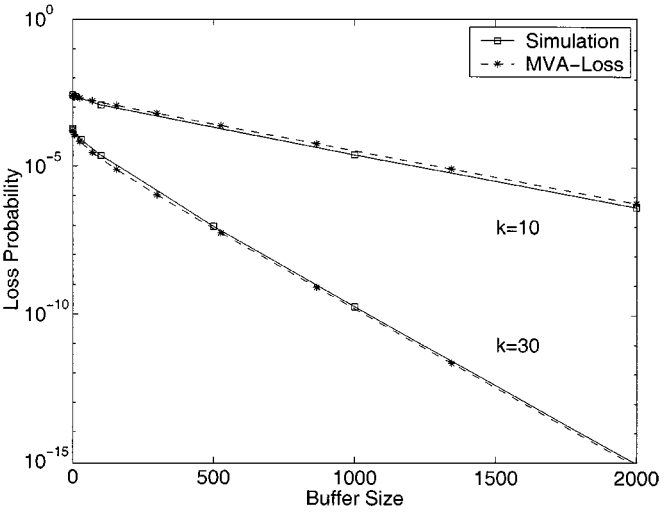


Fig. 2. Loss probability for a Gaussian input process with autocovariance function $C_\lambda(l) = 258 \times 0.9^{|l|}$ (mean rate: $\overline{\lambda} = 1000$; service rate–mean rate: $\kappa = 10, 20$).

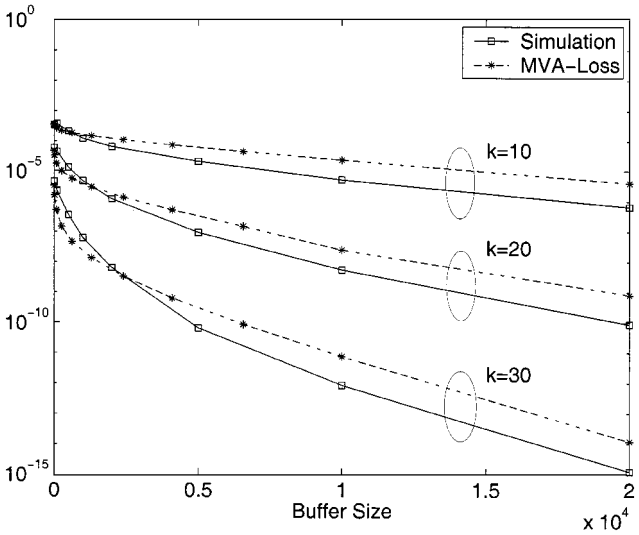


Fig. 3. Loss probability for a Gaussian input process with autocovariance function $C_\lambda(l) = 49.80 \times 0.9^{|l|} + 16.18 \times 0.99^{|l|} + 57.96 \times 0.999^{|l|}$ (mean rate: $\overline{\lambda} = 3000$; service rate–mean rate: $\kappa = 10, 20, 30$).

A. Gaussian Processes

We begin by considering the simple case when the input is a Gaussian autoregressive (AR) process with autocovariance $C_\lambda(l) = 258 \times 0.9^{|l|}$ (note that AR processes have been used to model variable bit-rate (VBR) video [22]). In Fig. 2, one can see that the simulation and MVA loss result in a close match over the entire range of buffers tested.

The next example, in Fig. 3, covers a scenario of multiple-timescale correlated traffic. Note that multiple-timescale correlated traffic is expected to be generated in high-speed networks because of the superposition of different types of sources [37]. In this case, the autocovariance function of the Gaussian input process is the weighted sum of three different powers, i.e., $C_\lambda(l) = 49.80 \times 0.9^{|l|} + 16.18 \times 0.99^{|l|} + 57.96 \times 0.999^{|l|}$. One can see from Fig. 3 that because of the multiple-timescale correlated nature of the input, the loss probability converges to its

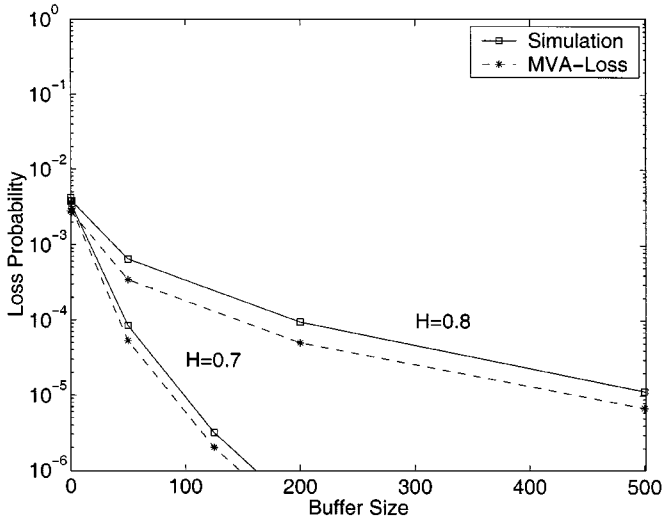


Fig. 4. Loss probability for a fractional Brownian motion process (Hurst parameter: $H = 0.7, , 0.8$; mean rate: $\overline{\lambda} = 300$; variance: $\sigma^2 = 100$; service rate–mean rate: $\kappa = 10$).

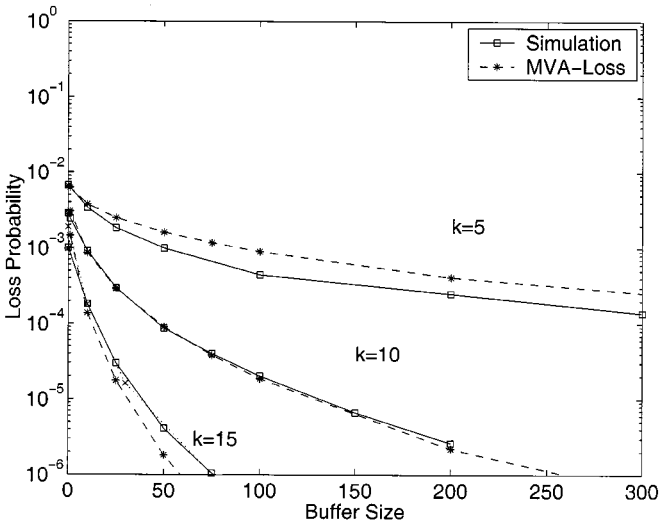


Fig. 5. Loss probability for the superpostition of two fractional Brownian motion processes. (Hurst parameters: $H_1 = 0.6, H_2 = 0.8$; mean rates: $\overline{\lambda}_1 = 150, \overline{\lambda}_2 = 150$, variances: $\sigma_1^2 = 50, \sigma_2^2 = 50$, service rate–mean rate: $\kappa = 5, 10, 15$).

asymptotic decay rate only at large buffer sizes. This observation is consistent with observations made on the tail probability when fed with multiple-timescale correlated traffic [20]. Again, it can be seen that the analytical result tracks the simulation results quite closely.

The next example deals with a well-known input process, the fractional Brownian motion process, which is the classical example of a self-similar process [23].[5] The results are shown in Figs. 4 and 5, demonstrating the accuracy of MVA loss, even for self-similar sources. Due to the difficulty in applying importance-sampling techniques to obtain loss probabilities for self-similar traffic, in Figs. 4 and 5, we show probabilities only as low as 10^{-6}. In Fig. 4, the input traffic is characterized by a single Hurst parameter. However, even if the traffic itself is long-range dependent, due to the heterogeneity of sources that high-speed networks will carry, we expect that it will be difficult to characterize the traffic by simply one parameter, such as the Hurst parameter. Hence, we also run an experiment for a more realistic scenario, i.e., the input process being the superposition of fractional Brownian motion processes with different Hurst parameters. The numerical result is shown in Fig. 5. One can see from Figs. 4 and 5 that MVA loss works well for self-similar sources.

B. Non-Gaussian Processes

In this section, we will compare the performance our MVA-loss approximation with simulations and also with other schemes in the literature. We call the Likhanov–Mazumdar technique described earlier "L–M," or "L–M:Gaussian" when further approximated by a Gaussian process, the Chernoff dominated eigenvalue technique in [38] "Chernoff-DE," the average/peak rate method in [39] "Ave/Peak," the analytical technique developed in [24] "Hybrid," and the famous effective bandwidth scheme "Effective BW" [40].

We now consider the practically important case of multiplexed voice sources. The input MMF process, which has widely been used to model voice traffic source [34], [35], has the following state transition matrix and rate vector:

$$\text{State transition matrix: } \begin{bmatrix} 0.9833 & 0.0167 \\ 0.025 & 0.975 \end{bmatrix}$$

$$\text{Input rate vector: } \begin{bmatrix} 0 \text{ cells/slot} \\ 0.85 \text{ cells/slot} \end{bmatrix}.$$

These values are chosen for a 45-Mb/s ATM link with 10-ms time slot and 53-byte ATM cell. In this example, we assume that 2900 voice sources are multiplexed on a 45-Mb/s ATM link with 10-ms time slot and 53-byte ATM cell. As shown in Fig. 6, the MVA loss obtains the loss probability calculations accurately and better than the other techniques.

We next investigate the accuracy of our approximation when the sources to the queue are generated from actual MPEG video traces. The trace used to generate this simulation result comes from an MPEG-encoded action movie (007 series) which has been found to exhibit long-range dependence [36]. In Fig. 7, 240 MPEG sources are multiplexed and served at 3667 cells/slot (OC-3 line), where we assume 25 frames/s and a 10-ms slot size. The loss probability versus buffer size result in this case is shown in Fig. 7. Again, it can be seen that the MVA-loss approximation tracks the simulation results quite closely.

C. Application to Admission Control

The final numerical result is to demonstrate the utility of MVA loss as a tool for admission control. We assume that a new flow is admitted to a multiplexer with buffer size x if the loss probability is less than the maximum tolerable loss probability ϵ.

[5]For computer simulations, since continuous-time Gaussian processes cannot be simulated, one typically uses a discrete-time version. In the case of fractional Brownian motion, the discrete-time version is called fractional Gaussian noise and has autocovariance function given by

$$C_\lambda(l) = \frac{\sigma^2}{2}(|l-1|^{2H} + |l+1|^{2H} - 2|l|^{2H})$$

where $H \in [0.5, 1)$ is the Hurst parameter.

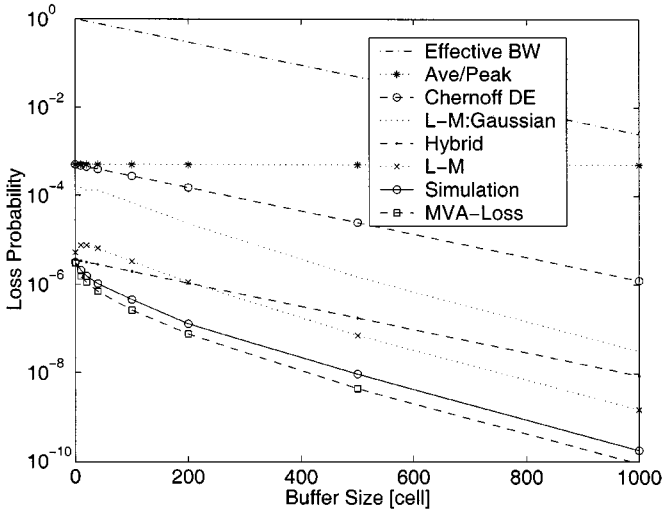


Fig. 6. Loss probability for 2900 voice sources. (Transition matrix = [0.9833, 0.0167; 0.025, 0.975]; rate vector = [0 cells/slot, 0.85 cells/slot]; mean rate (total): $\lambda = 987.18$ cells/slot; service rate–mean rate: $\kappa = 73.82$ cell/slot.)

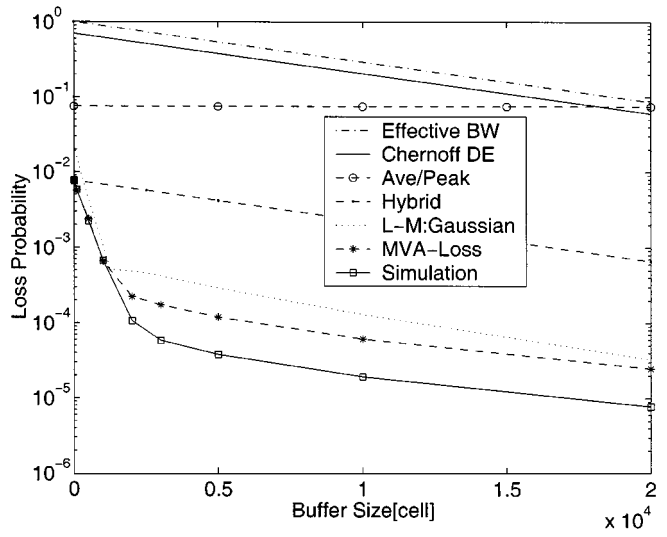
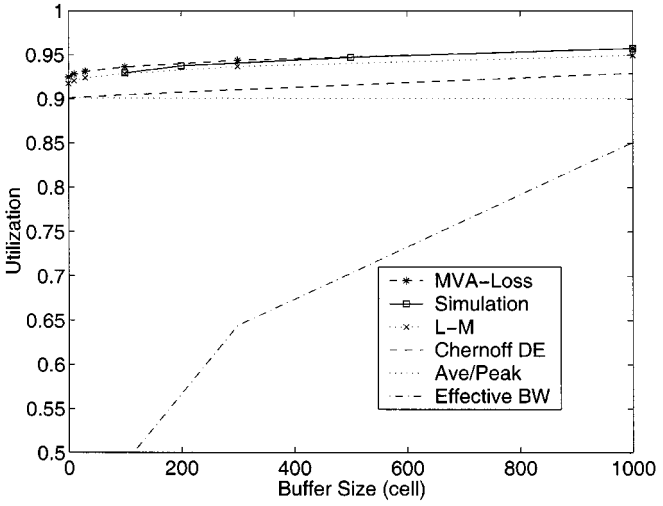
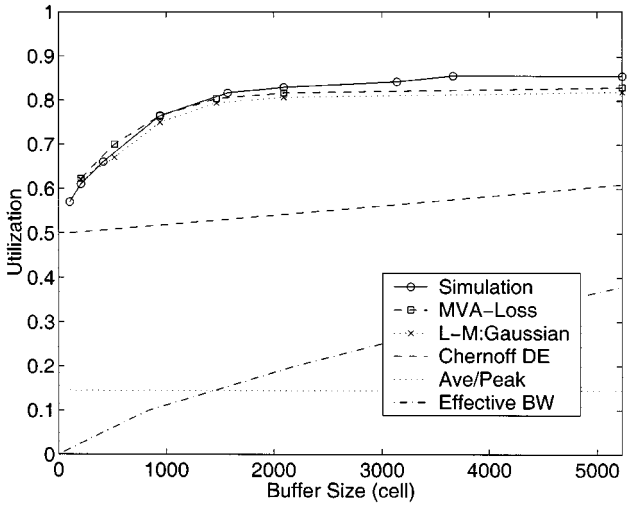


Fig. 7. Loss probability for 240 real MPEG traces from a 007 movie. [Mean rate (total): $\overline{\lambda} = 3469.45$ cells/slot; service rate–mean rate: $\kappa = 197.55$ cells/slot.]



(a)



(b)

Fig. 8. Admissible region for a 45-Mb/s link where voice/video sources are multiplexed. (Link capacity: $c = 45$ Mb/s $= 1046.7$ cells/slot; QoS parameter: $\epsilon = 10^{-6}$.)

In this example, we consider multiplexed voice sources on a 45-Mb/s link [Fig. 8(a)] or multiplexed video sources [Fig. 8(b)] for an admission-control type of application. The QoS parameter ϵ is set to 10^{-6}. For each voice source in Fig. 8(a), we use the same MMF ON–OFF process that was used for Fig. 6. For each video source, we use the same MPEG trace that was used in Fig. 7 (with start times randomly shifted). Then, the admission policy using MVA loss is the following. Let $\overline{\lambda}$ and $v(n)$ be the mean and the variance function of a single source, i.e., let $\overline{\lambda} := \mathbb{E}\{\lambda_n^{(1)}\}$ and $v(n) := \mathrm{Var}\{\sum_{k=1}^{n} \lambda_k^{(1)}\}$. When $(N-1)$ sources are currently serviced, a new source is admitted if

$$2\log\alpha - \sup_n \frac{Nv(n)}{\left((c - N\overline{\lambda})\, n + x\right)^2} < 2\log\epsilon \qquad (18)$$

where α is defined as in (14). In Fig. 8(a) and (b), we provide a comparison of admissible regions using different methods. It can be seen that MVA-loss curve most closely approximates the simulation curve in both figures. In Fig. 8(a), the L–M approximation performs as well, and the Chernoff DE approximation does only slightly worse. In Fig. 8(b), however, the Chernoff DE approximation in this case is found to be quite conservative. This is because for sources that are correlated at multiple timescales [such as the MPEG video sources in Fig. 8(b) shown here], the loss probability does not converge to its asymptotic decay rate quickly (even if there exists an asymptotic decay rate), and hence approximations such as the Chernoff DE scheme (or the hybrid scheme shown earlier) perform quite poorly.

Admission control by MVA loss can be extended to a case where heterogeneous flows are multiplexed. The link capacity is 622.02 Mb/s (OC-12 line), the buffer size x is fixed to 20 000 cells, and the QoS parameter ϵ is 10^{-6}. In this system, the input sources are of two types, JPEG video and voice. As a video source, we use a generic model that captures the multiple-timescale correlation observed in JPEG video traces.

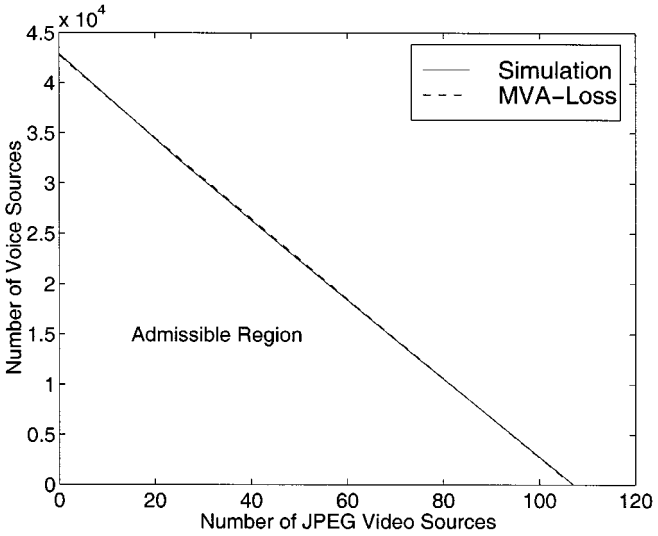


Fig. 9. Admissible region for an OC-12 line where voice and JPEG video sources are multiplexed. (Link capacity: $c = 622.02$ Mb/s $= 14\,467.7$ cells/slot; buffer size: $B = 2000$ cells; QoS parameter: $\epsilon = 10^{-6}$.)

It is a superposition of an i.i.d. Gaussian process and three two-state MMF processes:

$$\text{State transition matrices:}$$
$$\begin{bmatrix} 0.99 & 0.01 \\ 0.01 & 0.99 \end{bmatrix} \begin{bmatrix} 0.999 & 0.001 \\ 0.001 & 0.999 \end{bmatrix} \begin{bmatrix} 0.9999 & 0.0001 \\ 0.0001 & 0.9999 \end{bmatrix}$$
$$\text{Input rate vectors [cells/slot]:}$$
$$\begin{bmatrix} 0 \\ 45.516 \end{bmatrix} \begin{bmatrix} 0 \\ 31.86 \end{bmatrix} \begin{bmatrix} 0 \\ 18.204 \end{bmatrix}$$
$$\text{Mean of i.i.d. Gaussian: } 82.42$$
$$\text{Variance of i.i.d. Gaussian: } 8.6336.$$

Then, the admission policy is the following. Let $\overline{\lambda}_1$ and $v_1(n)$ be the mean and the variance function of a single voice source. Let $\overline{\lambda}_2$ and $v_2(n)$ be the mean and the variance function of a single video source. When $(N_1 - 1)$ voice and N_2 video flows are currently serviced, a new voice flow is admitted if

$$2\log\alpha - \sup_n \frac{N_1 v_1(n) + N_2 v_2(n)}{\left(\left(c - N_1\overline{\lambda}_1 - N_2\overline{\lambda}_2\right)n + x\right)^2} < 2\log\epsilon. \quad (19)$$

The boundary of the admissible region is obtained by finding maximal N_1 satisfying (19) for each N_2.

As one can see in Fig. 9, the admissible region estimated by simulations and via MVA loss is virtually indistinguishable. In fact, the difference between the two curves is less than 1% in terms of utilization.

IV. Asymptotic Properties of the MVA Approximation for Loss

We now find a strong asymptotic relationship between the loss probability and the tail probability. More specifically, under some conditions (to be defined later in Theorem 5), we find that

$$\log P_L(x) + \frac{m_x}{2} = O(\log x) \quad (20)$$

where $f = O(g)$ means that $\limsup |f/g| < \infty$. Equation (20) tells us that the divergence between the approximation $\alpha e^{-m_x/2}$, given in (14), and the loss probability is slow if at all [this may be easier to see if we rewrite (20) in the form $\log P_L(x) - \log \alpha e^{-m_x/2} = O(\log x)$].

In [27] and [28], under a set of general conditions it has been shown for the continuous-time case that

$$\log \mathbb{P}\{Q > x\} + \frac{m_x}{2} = O(\log x). \quad (21)$$

We will obtain (20) by finding a relationship between $P_L(x)$ and $\mathbb{P}\{Q > x\}$, i.e.

$$\log \mathbb{P}\{Q > x\} - \log P_L(x) = O(\log x) \quad (22)$$

under the set of conditions given in Theorem 5 [$P_L(x)$ will be bounded from above and below by some expressions in terms of $\mathbb{P}\{Q > x\}$], and then by applying (21) and some properties of m_x. Note, that finding the asymptotic relationship (22) between $\mathbb{P}\{Q > x\}$ and $P_L(x)$ is by itself a valuable and new contribution.

We first list a set of conditions for which (21) holds in the discrete-time case that are equivalent to the set of conditions in [27] defined for the continuous-time case. Let $v_n := \mathrm{Var}\{X_n\}$, $\psi(n) := \log v_n$, and $\beta := \lim_{n\to\infty} \psi(n)/\log n$ (assuming that the limit exists).

(H1) $\lim_{n\to\infty} n\left[\psi(n+1) - \psi(n)\right] = \beta.$

(H2) $v_n \overset{n\to\infty}{\sim} S n^{\beta}$ for some $S > 0$.

The notation $f(n) \overset{n\to\infty}{\sim} g(n)$ means that $\lim_{n\to\infty}(f(n)/g(n)) = 1$. The parameter β cannot be larger than 2 due to the stationarity of λ_n, and $\beta \in (0, 2)$ covers the majority of nontrivial stationary Gaussian processes. The Hurst parameter H is related to β by $\beta = 2H$. We now state the following results that are the discrete-time versions of the results in [27], [28], [41]. The proofs for these results are identical to those given in [27], [28], [41], with trivial modifications accounting for the discrete-time version, and, hence, we omit them here. These results are stated as Lemmas here, since we will be using them to prove our main theorem.

Lemma 2: Under hypotheses (H1) and (H2)

$$m_x \overset{x\to\infty}{\sim} \frac{4\kappa^{\beta}}{S\beta^{\beta}(2-\beta)^{2-\beta}} x^{2-\beta}.$$

Lemma 3: Under hypotheses (H1) and (H2)

$$\log \mathbb{P}\{Q > x\} + \frac{m_x}{2} = O(\log x).$$

It is easier for us to work with conditions on the autocovariance function of the input process rather than conditions (H1) and (H2). Hence, we first define a condition on the autocovariance function $C_\lambda(l)$ which guarantees (H1) and (H2):

$$\text{(C1)} \quad \sum_{l=-n}^{n} C_\lambda(l) \overset{n\to\infty}{\sim} S\beta n^{\beta-1}.$$

Note that condition (C1) is quite general and is satisfied not only by short-range dependent processes but also by a large class of long-range dependent processes including second-order self-similar and asymptotic self-similar processes [42].

Lemma 4: If the autocovariance function $C_\lambda(l)$ of λ_n satisfies (C1), then (H1) and (H2) hold.

Proof of Lemma 4: Let $h(n) := \sum_{l=-n}^{n} C_\lambda(l)$. Note that

$$v_n := \mathrm{Var}\{X_n\} = \sum_{k=1}^{n}\sum_{l=1}^{n} C_\lambda(l-k)$$
$$= nC_\lambda(0) + 2\sum_{l=1}^{n-1}(n-l)C_\lambda(l)$$

and that $v_{n+1} - v_n = h(n)$. First, we show condition (H2). Since both v_n and n^β approach ∞, $\lim_{n\to\infty}(v_n/n^\beta)$ should be equal to

$$\lim_{n\to\infty}\frac{v_{n+1}-v_n}{(n+1)^\beta - n^\beta}$$

if it exists (this is the discrete version of *L'Hospital's rule*). Hence

$$\lim_{n\to\infty}\frac{v_{n+1}-v_n}{(n+1)^\beta - n^\beta} = \frac{v_{n+1}-v_n}{\beta n^{\beta-1}}\frac{\beta n^{\beta-1}}{(n+1)^\beta - n^\beta}$$
$$= \frac{h(n)}{\beta n^{\beta-1}}\frac{\beta n^{\beta-1}}{(n+1)^\beta - n^\beta}$$
$$\overset{n\to\infty}{\to} S\cdot 1 \tag{23}$$

where $\lim_{n\to\infty}(v_n/n^\beta) = S$. Now, we show that (H1) also follows from (C1). Since $h(n)/v_n \sim \beta/n$, $f(x) = o(h(n)/v_n)$ implies that $f(x) = o(n^{-1})$. Note that a function $g(x)$ is $o(x)$ if $\lim_{x\to\infty} g(x)/x \to 0$. Now

$$n[\psi(n+1)-\psi(n)] = n\log\left(\frac{v_{n+1}}{v_n}\right)$$
$$= n\log\left(\frac{v_n + h(n)}{v_n}\right)$$
$$= n\log\left(1+\frac{h(n)}{v_n}\right)$$
$$= n\left[\frac{h(n)}{v_n} + o\left(\frac{1}{n}\right)\right]$$
(by Taylor Expansion)
$$= \frac{h(n)}{n^{\beta-1}}\frac{n^\beta}{v_n} + no\left(\frac{1}{n}\right)$$
$$\overset{n\to\infty}{\to} \beta S\cdot\frac{1}{S} + 0 = \beta. \tag{24}$$

■

The loss probability is closely related to the shape of the sample path, or how long Q_n stays in the overflow state. Before we give an illustrative example, we provide some notation. We define a *cycle* as this period, i.e., an interval between time instants when Q_n becomes zero. We let S_n^x denote the duration for which Q_n stays above threshold x in a cycle to which n belongs. Formally, let:

- $U_n := \sup\{k \le n : Q_{k-1} > 0, Q_k = 0\}$. (Start time of the current cycle to which n belongs.)
- $V_n := \inf\{k > n : Q_{k-1} > 0, Q_k = 0\}$. (Start time of the next cycle.)
- $W_n := V_n - U_n$. (Duration of a cycle to which n belongs.)
- $Z_n := V_n - n$. (Residual time to reach the end of cycle.)
- $S_n^x := \sum_{k=U_n}^{V_n-1} 1_{\{Q_k > x\}}$. (Duration for which $Q_k > x$ in a cycle containing n.)

Note that if $Q_n > 0$, Z_n is equal to the elapsed time to return to the empty-buffer (or zero) state. Since Q_n is stationary and ergodic, so are the above. Hence, their expectations are equal to time averages.

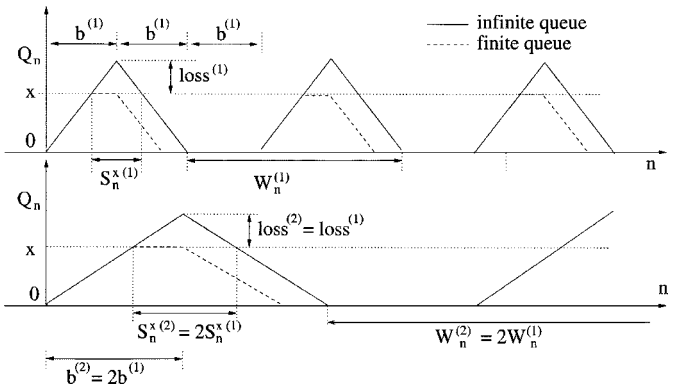


Fig. 10. Illustration of "same $\mathbb{P}\{Q > x\}$ but different $P_L(x)$."

Consider two systems whose sample paths look like those in Fig. 10. The sample paths are obtained when the input is a deterministic three-state source which generates fluid at rate $c+a$, $c-a$, and 0, at state 1, 2, and 3, respectively. The duration of each state is the same, say, b. Use the superscript (1) and (2) to represent values for the upper and the lower sample path. Set $a^{(1)} = 2a^{(2)}$ and $b^{(2)} = 2b^{(1)}$. Then, both cases have the same overflow probability. Now, consider a time interval from 0 to $3b^{(2)}$. The amount of fluid generated for that interval is clearly the same for both cases. But, the amount of loss in the upper case is exactly the twice of that in the lower case, hence, the upper case has the larger loss probability. We can infer from this that the loss probability is closely related to the length of S_n^x and the slope of the sample path. Since loss happens only when Q_n is greater than the buffer size x, we consider the condition that $Q_n > x$. Since it is difficult to know the distribution of S_n^x, and since S_n^x is determined by the sample path, we use a stochastic process defined as

$$Y_n = \sum_{k=1}^{n}\lambda_k + Q_0 - cn. \tag{25}$$

Here, we have chosen 0 as the origin, but, because of stationarity, the distribution of Y_n does not depend on the origin. Note that if $Q_0 > 0$, Y_n will be identical to Q_n until the end of cycle. We want to know the distribution of Y_n given $Q_0 > x$. Since Y_n is Gaussian, the distribution of Y_n can be characterized by the mean and the variance of Y_n. However, since Q_0 is the result of the entire history up to time 0 and the future is corre-

lated with the past, it is difficult to find an explicit expression of the mean and the variance of Y_n given $Q_0 > x$. Hence, we introduce upper-bound types of conditions on the mean and the variance of Y_n as (26) and (27). For notational simplicity, let $\mathbb{P}_x\{\cdot\} = \mathbb{P}\{\cdot | Q_0 > x\}$, and let $\mathbb{E}_x\{\cdot\}$ and $\mathrm{Var}_x\{\cdot\}$ be the expectation and the variance under $\mathbb{P}_x$, respectively.

We now state our main theorem.

Theorem 5: Assume condition (C1). Further assume that for any $\epsilon > 0$, there exist x_0, K, M, and α such that

$$\mathbb{E}_x\{Y_n\} \le (-\kappa + \epsilon)n \tag{26}$$

$$\mathrm{Var}_x\{Y_n\} \le Kn^{\beta} \tag{27}$$

for all $x \ge x_0$ and $n \ge Mx^{\alpha}$. Then

$$-\infty < \liminf_{x\to\infty} \frac{1}{\log x}\left(\log P_L(x) + \frac{m_x}{2}\right) \le \limsup_{x\to\infty} \frac{1}{\log x}\left(\log P_L(x) + \frac{m_x}{2}\right) < \infty. \tag{28}$$

Though the conditions of Theorem 5 look somewhat complex, they are expected to be satisfied by a large class of Gaussian processes. If the input process is i.i.d. with $C_\lambda(0) = S$ and $C_\lambda(l) = 0$ for $l \neq 0$, it can be easily checked that

$$\mathbb{E}_x\{Y_n\} \le -\kappa n + x$$
$$\mathrm{Var}_x\{Y_n\} = Sn$$

and (C1), (26), and (27) are satisfied with $\beta = 1$, $K = S$, $M = 1/\epsilon$, and $\alpha = 1$. It has been shown that Gaussian processes represented by the form of finite-ordered autoregressive moving average (ARMA) satisfy (26) and (27) [17]. Since the autocovariance function of a stable ARMA process is in the form of $C_\lambda(l) = \sum_{i=1}^{N} a_i p_i^{|l|}$ with $|p_i| < 1$, it satisfies (C1) with $\beta = 1$. So Theorem 5 is applicable to Gaussian ARMA processes.

More generally, $\mathbb{E}\{\sum_{k=1}^{n}\lambda_k\} - cn = -\kappa n$, and $\mathrm{Var}\{\sum_{k=1}^{n}\lambda_k\} \overset{n\to\infty}{\sim} Sn^{\beta}$ under (C1). Thus, for each x, $\mathbb{E}_x\{Y_n\} \overset{n\to\infty}{\sim} -\kappa n$ and $\mathrm{Var}_x\{Y_n\} \overset{n\to\infty}{\sim} Sn^{\beta}$, and we can find $K = K(\epsilon, x)$, $M = M(\epsilon, x)$, and $\alpha = \alpha(\epsilon, x)$ as small as possible. If $\sup_x K(\epsilon, x)$, $\sup_x M(\epsilon, x)$, and $\sup_x \alpha(\epsilon, x)$ are finite, then (26) and (27) hold. We conjecture that they are all finite for a large class of stationary Gaussian processes, and we are attempting to show it.

Note that the rightmost inequality ($\limsup$ part) in (28) holds without conditions (26) and (27), and it agrees with empirical observations that the tail probability curve provides an upper bound to the loss probability curve.

Before we prove the theorem, we first define the derivative of m_x with respect to x, m'_x. Recall (9), or

$$m_x = \frac{(x + \kappa n_x)^2}{v_{n_x}}.$$

Since n_x is an integer value, m_x is differentiable except for countably many x at which n_x has a jump. Let $D := \{x : m_x$ is not differentiable$\}$. Note that D has measure zero, and that the left and right limits of m'_x and m''_x exist for all $x \in D$. For simplicity, abuse notations by setting $m'_x = \lim_{z\downarrow x} m'_z$ and $m''_x = \lim_{z\downarrow x} m''_z$ for $x \in D$. The reason we set the (right) limit is that we will find the similarity relation (29) in Lemma 6, which is useful in proving Theorem 5. In fact, we may take the left limit to have the same asymptotic behavior. By building m'_x and m''_x in this way, it directly follows from Lemma 2 that $m'_x \sim ax^{1-\beta}$ and $m''_x \sim bx^{-\beta}$ for some constants $a > 0$ and b.

We now state three lemmas which are useful in proving the theorem. (Their proofs are in the Appendix.)

Lemma 6: Under hypotheses (H1) and (H2)

$$\int_x^{\infty} y^K e^{-(m_y/2)}\, dy \overset{x\to\infty}{\sim} \frac{2x^K}{m'_x} e^{-(m_x/2)} \tag{29}$$

where K is a constant.

Lemma 7: If $\mathbb{P}\{Q > x\} > 0$ and $\mathbb{E}\{Z|Q > x\} < \infty$ for all x

$$\int_x^{\infty} \frac{1}{2\mathbb{E}\{Z|Q > y\}}\, \mathbb{P}\{Q > y\}\, dy \le \overline{\lambda} P_L(x). \tag{30}$$

Lemma 8: Under conditions (26) and (27), $\mathbb{E}\{Z|Q > x\} = O(x^{\alpha})$ for some $\alpha \ge 1$.

Now, we are ready to prove Theorem 5.

Proof of Theorem 5: First of all, we find expressions in terms of $\mathbb{P}\{Q > x\}$ which are greater than or less than $\overline{\lambda} P_L(x)$. If $\mathbb{P}\{Q > x\} = 0$ for some x, it would contradict the asymptotic relation in Lemma 3. Hence, $\mathbb{P}\{Q > x\} > 0$ for all x. If $\mathbb{E}\{Z|Q > x\} = \infty$ for some x, it would contradict the asymptotic relation in Lemma 8. Hence, $\mathbb{E}\{Z|Q > x\} < \infty$ for all x. Thus, by Lemma 7 we have (30). Now, since $(Q_n - x)^+ = (Q_{n-1} + \lambda_n - c - x)^+$ from (5)

$$\begin{aligned}\overline{\lambda} P_L(x) &= \mathbb{E}\left\{\left(\hat{Q}_{n-1} + \lambda_n - c - x\right)^+\right\} \\ &\le \mathbb{E}\{(Q_{n-1} + \lambda_n - c - x)^+\} \\ &= \mathbb{E}\{(Q_n - x)^+\} \\ &= \int_x^{\infty} \mathbb{P}\{Q > y\}\, dy. \end{aligned} \tag{31}$$

By Lemma 4, (C1) implies (H1) and (H2). Hence, by Lemma 3, we have (21). Equation (21) means that there are x_0, K_1, and K_2 such that

$$\begin{aligned} e^{-(m_y/2) + K_1 \log y} &\le \mathbb{P}\{Q > y\} \\ &\le e^{-(m_y/2) + K_2 \log y}, \quad \forall y \ge x_0. \end{aligned} \tag{32}$$

Note that since $\mathbb{E}\{Z|Q > x\} = O(x^{\alpha})$ from Lemma 8, we can choose $K_3 > 0$ such that $\mathbb{E}\{Z|Q > x\} \le K_3 x^{\alpha}$ for all $x \ge x_0$. Combining with (30) and (31), integrate all sides of (32) to get

$$\begin{aligned} &\int_x^{\infty} \frac{1}{K_3 y^{\alpha}}\, y^{K_1} e^{-(m_y/2)}\, dy \\ &\quad \le \overline{\lambda} P_L(x) \le \int_x^{\infty} y^{K_2} e^{-(m_y/2)}\, dy, \quad \forall x \ge x_0. \end{aligned} \tag{33}$$

Since $m'_x \sim ax^{1-\beta}$ with the constant $a > 0$, by Lemma 6, there exist $x_1 \ge x_0$, $K_4 > 0$ and $K_5 > 0$ such that

$$\begin{aligned} &K_4 x^{K_1 - \alpha - 1 + \beta} e^{-(m_x/2)} \\ &\quad \le \int_x^{\infty} \frac{1}{K_3 y^{\alpha}}\, y^{K_1} e^{-(m_y/2)}\, dy, \qquad \forall x \ge x_1 \end{aligned} \tag{34}$$

and

$$\int_x^\infty y^{K_2} e^{-(m_y/2)}\,dy \le K_5 x^{K_2-1+\beta} e^{-(m_x/2)}, \qquad \forall x \ge x_1. \tag{35}$$

From (33)–(35)

$$K_4 x^{K_1-\alpha-1+\beta} e^{-(m_x/2)} \le \overline{\lambda} P_L(x) \le K_5 x^{K_2-1+\beta} e^{-(m_x/2)}, \quad \forall x \ge x_1.$$

Take logs and rearrange to get

$$\log\left(\frac{K_4}{\overline{\lambda}}\right) + (K_1-\alpha-1+\beta)\log x \le \log P_L(x) + \frac{m_x}{2} \le \log\left(\frac{K_5}{\overline{\lambda}}\right) + (K_2-1+\beta)\log x, \quad \forall x \ge x_1.$$

Divide by $\log x$ and take $x \to \infty$. Then, the theorem follows. ■

V. Applications to On-Line Measurements

In this section, we describe how to apply the MVA approach for the estimation of the loss probability, based on on-line measurements. In many practical situations, the characteristics of a flow may not be known beforehand or represented by a simple set of parameters. Hence, when we use a tool for the estimation of the loss probability, parameter values such as the moment generating function and the variance function should be evaluated from on-line measurements. Then, the question is what range of those parameters should be evaluated. If an estimation tool needs, for example, the evaluation of the moment generating function for the entire range of (θ, n), the tool may not be useful. This is fortunately not the case for the MVA approximation for loss.

Note that the MVA result has the form $\alpha e^{-m_x/2}$. The parameter m_x is a function of c, $\overline{\lambda}$, x, and $v(n)$, where $\overline{\lambda}$ and $v(n)$ are the mean and the variance of the input, i.e., $\overline{\lambda} = \mathbb{E}\{\lambda_n\}$ and $v(n) = \mathrm{Var}\{\sum_{\mathrm{k}=1}^{\mathrm{n}} \lambda_{\mathrm{k}}\}$. Hence, by measuring only the first two moments of the input we can estimate the loss probability. Recall that

$$m_x = \sup_n \frac{v(n)}{\left(c-\overline{\lambda}+x\right)^2}$$

and that $v(n)/\left(c-\overline{\lambda}+x\right)^2$ is maximized at $n = n_x$. This means that the result only depends on the value of $v(n)$ at $n = n_x$. This value of n_x corresponds to the most likely timescale over which loss occurs. This is called the *dominant time scale* (DTS) in the literature [43], [20]. Thus, the DTS provides us with a window over which to measure the variance function. It appears at first, however, that this approach may not work, because the DTS requires taking the maximum of the normalized variance over all n, which means that we would need to know $v(n)$ for all n beforehand. Thus, we are faced with a *chicken and egg* type of problem, i.e., which should we do first: measuring the variance function $v(n)$ of the input, or estimating the measurement window n_x. Fortunately, this type of cycle has recently been broken and a bound on the DTS can in fact be found through on-line measurements (see Theorem 1 and the algorithm in [44]). Thus, since our approximation is dependent on the DTS, we only need to estimate $v(n)$, for values of n up to a bound on the DTS (given in [44]), thereby making it amenable for on-line measurements.

VI. Concluding Remarks

We have proposed an approximation for the loss probability in a finite queue by making a simple mapping from the MVA estimate of the tail probability in the corresponding infinite queue. We show first via simulation results that our approximation is accurate for different input processes and a variety of buffer sizes and utilization. Since the loss probability is an important QoS measure of network traffic, this approximation will be useful in admission control and network design. Another feature of the approximation is that it is given in a single equation format and hence can easily be implemented in real-time. We have compared our approximation to existing methods including the effective bandwidth approximation, the Chernoff dominant eigenvalue approximation, and the many-sources asymptotic approximation of Likhanov and Mazumdar.

In this paper we also study the theoretical aspects of our approximation. In particular, we provide a strong asymptotic result that relates our approximation to the actual loss probability. We show that if our approximation were to diverge (with increasing buffer size) from the loss probability, it would do so slowly. For future work we plan on simplifying the conditions given in Theorem 5 and to extend the approximation result to a network of queues.

Appendix

Proof of Lemma 6: Let $f(x) = (m_x/2) - K\log x$. Since $m_x' \overset{x\to\infty}{\sim} a x^{1-\beta}$ and $\beta < 2$, $f'(x) \overset{x\to\infty}{\sim} m_x'/2$. Hence, to prove the lemma, it suffices to show that

$$\int_x^\infty e^{-f(y)}\,dy \overset{x\to\infty}{\sim} \frac{1}{f'(x)}\,e^{-f(x)}. \tag{36}$$

Let $D = \{x : m_x \text{ is not differentiable}\}$. For $x \notin D$,

$$\frac{d}{dy}\left(\frac{1}{f'(y)}e^{-f(y)}\right)\bigg|_{y=x} = -e^{-f(x)} - \frac{f''(x)}{f'(x)^2}\,e^{-f(x)}. \tag{37}$$

Since D has measure zero, $\int_{[x,\infty)-D}(\cdot)\,dy = \int_{[x,\infty)}(\cdot)\,dy$ and we may assign any values to $f'(x)$ and $f''(x)$ for all $x \in D$. Recall $m_x' = \lim_{z\downarrow x} m_z'$ and $m_x'' = \lim_{z\downarrow x} m_z''$ for $x \in D$. Set $f'(x) = \lim_{z\downarrow x} f'(x)$ and $f''(x) = \lim_{z\downarrow x} f''(x)$ for $x \in D$.

Now, let x be any value. Integrating both sides of (37) from x to ∞, we have

$$-\frac{1}{f'(x)}\,e^{-f(x)} + \lim_{x\to\infty}\frac{1}{f'(x)}\,e^{-f(x)} = -\frac{1}{f'(x)}\,e^{-f(x)} = -\int_x^\infty e^{-f(y)}\,dy - \int_x^\infty \frac{f''(y)}{f'(y)^2}\,e^{-f(y)}\,dy. \tag{38}$$

Note that $m_x' \sim a x^{1-\beta}$ and $m_x'' \sim b x^{-\beta}$ with constants $a > 0$ and b. Since $f'(x) = (m_x'/2) - Kx^{-1} \sim a x^{1-\beta}$ and $f''(x) =$

$(m_x''/2) + Kx^{-2} \sim bx^{-\beta}$, $\lim_{x\to\infty}(f''(x)/f'(x)^2) = 0$. Let $\epsilon \in (0, 1)$. We can find x_0 such that $|(f''(x)/f'(x)^2)| < \epsilon$ for all $x \geq x_0$. Then

$$\begin{aligned}
&\frac{1}{f'(x)}\,e^{-f(x)} - \epsilon\int_x^\infty e^{-f(y)}\,dy \\
&\quad\leq \int_x^\infty e^{-f(y)}\,dy \\
&\quad= \frac{1}{f'(x)}\,e^{-f(x)} - \int_x^\infty \frac{f''(y)}{f'(y)^2}\,e^{-f(y)}\,dy \\
&\quad\leq \frac{1}{f'(x)}\,e^{-f(x)} + \epsilon\int_x^\infty e^{-f(y)}\,dy, \quad \forall\, x \geq x_0 \qquad (39)
\end{aligned}$$

which means that

$$\begin{aligned}
&\frac{1}{1+\epsilon}\,\frac{1}{f'(x)}\,e^{-f(x)} \\
&\quad\leq \int_x^\infty e^{-f(y)}\,dy \\
&\quad\leq \frac{1}{1-\epsilon}\,\frac{1}{f'(x)}\,e^{-f(x)}, \quad \forall\, x \geq x_0 \qquad (40)
\end{aligned}$$

and the result follows. ∎

Proof of Lemma 7: Recall the notations:

- $U_n := \sup\{k \leq n : Q_{k-1} > 0, Q_k = 0\}$ (start time of the current cycle to which n belongs).
- $V_n := \inf\{k > n : Q_{k-1} > 0, Q_k = 0\}$ (start time of the next cycle).
- $W_n := V_n - U_n$ (duration of a cycle to which n belongs).
- $Z_n := V_n - n$ (residual time to reach the end of cycle).
- $S_n^x := \sum_{k=U_n}^{V_n-1} 1_{\{Q_k>x\}}$ (duration for which $Q_k > x$ in a cycle containing n).

Define one more:

- $R_n^x := \sum_{k=n}^{V_n-1} 1_{\{Q_k>x\}}$ (residual duration for which $Q_k > x$ in a cycle containing n).

Since Q_n is stationary and ergodic, so are the above. Hence, their expectations are equal to time averages. Since we are interested in the behavior of Q_n after loss happens, we consider the conditional expectations:

$$\mathbb{E}\{Z_n|Q_n > x\} = \lim_{k\to\infty} \frac{1}{\sum_{i=1}^k 1_{\{Q_i>x\}}} \sum_{i=1}^k Z_i 1_{\{Q_i>x\}} \qquad (41)$$

$$\mathbb{E}\{S_n^x|Q_n > x\} = \lim_{k\to\infty} \frac{1}{\sum_{i=1}^k 1_{\{Q_i>x\}}} \sum_{i=1}^k S_i^x 1_{\{Q_i>x\}} \qquad (42)$$

$$\mathbb{E}\{R_n^x|Q_n > x\} = \lim_{k\to\infty} \frac{1}{\sum_{i=1}^k 1_{\{Q_i>x\}}} \sum_{i=1}^k R_i^x 1_{\{Q_i>x\}}. \qquad (43)$$

Clearly, $\mathbb{E}\{R_n^x|Q_n > x\} \leq \mathbb{E}\{Z_n|Q_n > x\}$. And it can also be easily checked that $2\mathbb{E}\{R_n^x|Q_n > x\} \geq \mathbb{E}\{S_n^x|Q_n > x\}$, where the inequality is due to that n is discrete.[6] Since $\mathbb{E}\{\lambda\} < c$, there are infinitely many cycles for a sample path. Index cycles in the following manner:

- $V^{(1)} := V_1$, $U^{(1)} := U_1$, $V^{(i)} := V_{V^{(i-1)}+1}$, $U^{(i)} := V^{(i-1)}$, for $i > 1$,
- $A^{(i)} := \{n : U^{(i)} \leq n < V^{(i)}\}$.

Define:

- $S_x^{(i)} := \sum_{k\in A^{(i)}} I(Q_k > x)$, $i = 1, 2, 3, \ldots$,
- $$\hat{S}^x := \limsup_{m\to\infty} \frac{\sum_{i=1}^m S_x^{(i)}}{\sum_{i=1}^m I\left(S_x^{(i)} > 0\right)}.$$

Now, we prove the lemma in two steps:

- 1) Derive
$$P_L(x)\overline{\lambda} \geq \int_x^\infty \frac{1}{\hat{S}^y}\,\mathbb{P}\{Q > y\}\,dy.$$
- 2) Show $2\mathbb{E}\{Z_n|Q_n > x\} \geq \hat{S}^x$.

Step 1): The amount of loss in cycle i is greater than or equal to the difference between the maximum value of the queue level Q_n in cycle i and the buffer size x of the finite buffer queue, i.e.

$$\begin{aligned}
L^{(i)} &\geq \max_{k\in A^{(i)}} (Q_k - x)^+ \\
&= \int_x^\infty I\left(\max_{k\in A^{(i)}} Q_k > y\right) dy \\
&= \int_x^\infty I\left(S_y^{(i)} > 0\right) dy.
\end{aligned}$$

Take summation over i and divide by the total time, $\sum_{i=1}^m |A^{(i)}|$, where $|A^{(i)}|$ denotes the number of elements of $A^{(i)}$. Then

$$\begin{aligned}
\frac{\sum_{i=1}^m L^{(i)}}{\sum_{i=1}^m |A^{(i)}|} &= \left(\frac{\sum_{i=1}^m L^{(i)}}{\sum_{i=1}^m \sum_{k\in A^{(i)}} \lambda_k}\right)\left(\frac{\sum_{i=1}^m \sum_{k\in A^{(i)}} \lambda_k}{\sum_{i=1}^m |A^{(i)}|}\right) \\
&\geq \int_x^\infty \frac{\sum_{i=1}^m I\left(S_y^{(i)} > 0\right)}{\sum_{i=1}^m |A^{(i)}|}\,dy \\
&= \int_x^\infty \frac{\sum_{i=1}^m I\left(S_y^{(i)} > 0\right)}{\sum_{i=1}^m S_y^{(i)}} \frac{\sum_{i=1}^m S_y^{(i)}}{\sum_{i=1}^m |A^{(i)}|}\,dy
\end{aligned}$$

[6] Since n is discrete, for given n such that $Q_n > x$, R_n^x and S_n^x take (positive) integer values. If S_n is, for example 2, R_n can be either 1 or 2, and its expectation is 1.5 which is greater than 2/2.

$$\geq \int_x^\infty \left(\sup_{l\geq m} \frac{\sum_{i=1}^{l} S_y^{(i)}}{\sum_{i=1}^{l} I\left(S_y^{(i)}>0\right)} \right)^{-1} \frac{\sum_{i=1}^{m} S_y^{(i)}}{\sum_{i=1}^{m} |A^{(i)}|}\, dy. \tag{44}$$

Recalling (1) and (2)

$$\frac{\sum_{i=1}^{m} L^{(i)}}{\sum_{i=1}^{m} \sum_{k\in A^{(i)}} \lambda_k} \longrightarrow P_L(x)$$

$$\frac{\sum_{i=1}^{m} \sum_{k\in A^{(i)}} \lambda_k}{\sum_{i=1}^{m} |A^{(i)}|} \longrightarrow \overline{\lambda}$$

$$\frac{\sum_{i=1}^{m} S_y^{(i)}}{\sum_{i=1}^{m} |A^{(i)}|} \longrightarrow \mathbb{P}\{Q>x\}$$

$$\sup_{l\geq m} \frac{\sum_{i=1}^{l} S_y^{(i)}}{\sum_{i=1}^{l} I\left(S_y^{(i)}>0\right)} \longrightarrow \hat{S}^y$$

as $m \to \infty$. Since all components are nonnegative, by Fatou's Lemma, (44) becomes

$$P_L(x)\overline{\lambda} \geq \int_x^\infty \frac{1}{\hat{S}^y}\, \mathbb{P}\{Q>y\}\, dy.$$

Step 2): For better understanding, we first show

$$\limsup_{m\to\infty} \frac{1}{m} \sum_{i=1}^{m} S_x^{(i)} \leq \lim_{m\to\infty} \frac{1}{\sum_{i=1}^{m} S_x^{(i)}} \sum_{i=1}^{m} \left(S_x^{(i)}\right)^2. \tag{45}$$

Note that all components are nonnegative. Let

$$a_m := \frac{1}{m} \sum_{i=1}^{m} S_x^{(i)}$$

$$b_m := \frac{1}{\sum_{i=1}^{m} S_x^{(i)}} \sum_{i=1}^{m} \left(S_x^{(i)}\right)^2$$

$$a^* = \limsup a_m$$

and

$$b^* = \lim b_m.$$

For any $\epsilon > 0$, we can choose M such that $a_M - a^* < \epsilon$ and $|b^* - b_M| < \epsilon$. Then,

$$\begin{aligned} b^* - a^* &= b_M + (b^* - b_M) - a_M - (a^* - a_M) \\ &\geq (b_M - a_M) - 2\epsilon \geq -2\epsilon \end{aligned}$$

since

$$\begin{aligned} b_M - a_M &= \frac{\sum \left(S_x^{(i)}\right)^2}{\sum S_x^{(i)}} - \frac{\sum S_x^{(i)}}{M} \\ &= \frac{\sum_{i\neq j} \left(S_x^{(i)} - S_x^{(j)}\right)^2}{M \sum S_x^{(i)}} \geq 0. \end{aligned}$$

Since ϵ is arbitrary, we have $b^* \geq a^*$.

Now, we will verify that

$$\hat{S}^x \leq \lim_{m\to\infty} \frac{1}{\sum_{i=1}^{m} S_x^{(i)}} \sum_{i=1}^{m} \left(S_x^{(i)}\right)^2. \tag{46}$$

Construct a new sequence $\{T_x^{(i)}\}$ by removing zero-valued elements of $\{S_x^{(i)}\}$. Then, as in (45)

$$\limsup_{m\to\infty} \frac{1}{m} \sum_{i=1}^{m} T_x^{(i)} \leq \lim_{m\to\infty} \frac{1}{\sum_{i=1}^{m} T_x^{(i)}} \sum_{i=1}^{m} \left(T_x^{(i)}\right)^2. \tag{47}$$

Note that

$$\begin{aligned} \limsup_{m\to\infty} \frac{1}{m} \sum_{i=1}^{m} T_x^{(i)} &= \limsup_{m\to\infty} \frac{1}{\sum_{i=1}^{m} I\left(S_x^{(i)}>0\right)} \sum_{i=1}^{m} S_x^{(i)} \\ &= \hat{S}^x. \end{aligned} \tag{48}$$

Let $B_x^{(i)} := \{n : U^{(i)} \leq n < V^{(i)}, Q_n > x\}$. Since $S_j^x = S_x^{(i)}$ for all $j \in B_x^{(i)}$ and $|B_x^{(i)}| = S_x^{(i)}$

$$\begin{aligned} \mathbb{E}\{S_n^x | Q_n > x\} &= \lim_{k\to\infty} \frac{1}{\sum_{i=1}^{k} 1_{\{Q_i>x\}}} \sum_{i=1}^{k} S_i^x 1_{\{Q_i>x\}} \\ &= \lim_{k\to\infty} \frac{1}{\sum_{i=1}^{k} \sum_{j\in B_x^{(i)}} 1} \sum_{i=1}^{k} \sum_{j\in B_x^{(i)}} S_j^x \\ &= \lim_{k\to\infty} \frac{1}{\sum_{i=1}^{k} |B_x^{(i)}|} \sum_{i=1}^{k} \sum_{j\in B_x^{(i)}} S_x^{(i)} \\ &= \lim_{k\to\infty} \frac{1}{\sum_{i=1}^{k} S_x^{(i)}} \sum_{i=1}^{k} \left(S_x^{(i)}\right)^2 \\ &= \lim_{k\to\infty} \frac{1}{\sum_{i=1}^{k} T_x^{(i)}} \sum_{i=1}^{k} \left(T_x^{(i)}\right)^2. \end{aligned} \tag{49}$$

Combining (47)–(49), we have (46).

At last, we have

$$\begin{aligned}\hat{S}^x \le& \mathbb{E}\{S_n^x | Q_n > x\} \le 2\mathbb{E}\{R_n^x | Q_n > x\} \\ \le& 2\mathbb{E}\{Z_n | Q_n > x\} \end{aligned} \tag{50}$$

from which (30) follows. ∎

Proof of Lemma 8: Define:

- $\Phi(x, n) := \mathbb{P}_x\{Y_n > 0\}$;
- $\boldsymbol{V}(x_0, \alpha, M) := \{(x, n) : x \ge x_0, n \ge Mx^\alpha\}$.

The proof will be done in two steps:

- 1) Find $x_1 > x_0$ such that $\Phi(x, n) \le n^{-2}$ for all $(x, n) \in \boldsymbol{V}(x_1, \alpha, M)$.
- 2) Using 1), show that $\mathbb{E}\{Z | Q_0 > x\}$ is $O(x^\alpha)$.

Step 1): Let ϵ be so small that $-\kappa + \epsilon < 0$. Then, we choose x_0, M, and α satisfying (26) and (27). Let $m(n) := \mathbb{E}_x\{Y_n\}$ and $v(n) := \mathrm{Var}_x\{Y_n\}$. Then, the moment generating function of Gaussian Y_n is given by $e^{\theta m(n) + (1/2)\theta^2 v(n)}$. From (26) and (27), $m(n) \le (-\kappa + \epsilon)n$ and $v(n) \le Kn^\beta$ for all $(x, n) \in \boldsymbol{V}(x_0, \alpha, M)$. Thus

$$\begin{aligned}\Phi(x, n) \le& \mathbb{E}_x\{e^{\theta Y_n}\} \qquad \text{(Chernoff bound)} \\ =& e^{\theta m(n) + (1/2)\theta^2 v(n)} \\ \le& e^{-\theta\kappa' n + (K/2)\theta^2 n^\beta}, \quad \forall (x, n) \in \boldsymbol{V}(x_0, \alpha, M), \end{aligned} \tag{51}$$

where $\kappa' = \kappa - \epsilon$. Let $\theta = n^{-1+\delta}$ with $\delta \in (0, 2 - \beta)$. Then, for all $(x, n) \in \boldsymbol{V}(x_0, \alpha, M)$

$$\Phi(x, n) \le \exp\left(-\kappa' n^\delta + \frac{K}{2} n^{2\delta - (2-\beta)}\right). \tag{52}$$

Note that $\delta > 2\delta - (2 - \beta)$. Since the coefficient of the leading term, $-\kappa'$, is negative and its order, δ, is positive, we have for all $(x, n) \in \boldsymbol{V}(x_0, \alpha, M)$ as $n \to \infty$

$$n^2 \Phi(x, n) \le n^2 \exp\left(-\kappa' n^\delta + \frac{K}{2} n^{2\delta - (2-\beta)}\right) \to 0. \tag{53}$$

Note that in (53) κ', K, and δ are fixed constants for all $(x, n) \in \boldsymbol{V}(x_0, \alpha, M)$. Thus, there exists x_2 such that $\Phi(x, n) \le n^{-2}$ for all $(x, n) \in \boldsymbol{V}(x_0, \alpha, M)$ with $n \ge x_2$. Now, we choose $x_1 > x_0$ such that $Mx_1^\alpha \ge x_2$. Then, $\Phi(x, n) \le n^{-2}$ for all $(x, n) \in \boldsymbol{V}(x_1, \alpha, M)$.

Step 2): Consider

$$\begin{aligned}\mathbb{P}_x\{Y_n \le 0\} =& \mathbb{P}_x\{Y_n \le 0 | Z \le n\}\mathbb{P}_x\{Z \le n\} \\ &+ \mathbb{P}_x\{Y_n \le 0 | Z > n\}\mathbb{P}_x\{Z > n\}. \end{aligned} \tag{54}$$

From the definition of Z, $Z > n$ implies $Y_n > 0$. Thus, $\mathbb{P}_x\{Y_n \le 0 | Z > n\} = 0$. Therefore, we have

$$\begin{aligned}\mathbb{P}_x\{Y_n \le 0\} =& \mathbb{P}_x\{Y_n \le 0 | Z \le n\}\mathbb{P}_x\{Z \le n\} \\ \le& \mathbb{P}_x\{Z \le n\}, \end{aligned}$$

or

$$\mathbb{P}_x\{Y_n > 0\} \ge \mathbb{P}_x\{Z > n\}. \tag{55}$$

Obviously, $\mathbb{P}_x\{Y_n > 0\} = \Phi(x, n) \le 1$. Let $x \ge x_1$. Then, as shown in Step 3), $\Phi(x, n) \le n^{-2}$ for all $n \ge Mx^\alpha$. Applying this and (55)

$$\begin{aligned}\mathbb{E}_x\{Z\} =& \sum_{n=0}^{\infty} \mathbb{P}_x\{Z > n\} \\ \le& \sum_{n=0}^{\infty} \mathbb{P}_x\{Y_n > 0\} \\ =& \sum_{n=0}^{\infty} \Phi(x, n) \\ =& \sum_{n=0}^{\lceil Mx^\alpha \rceil} \Phi(x, n) + \sum_{n=\lceil Mx^\alpha \rceil + 1}^{\infty} \Phi(x, n) \\ \le& \sum_{n=0}^{\lceil Mx^\alpha \rceil} 1 + \sum_{n=\lceil Mx^\alpha \rceil + 1}^{\infty} n^{-2}, \\ \overset{x \to \infty}{\sim}& Mx^\alpha, \end{aligned} \tag{56}$$

where $\lceil x \rceil$ denotes the smallest integer which is greater than or equal to x. Since $\mathbb{E}_x\{Z\}$ is nonnegative, $\mathbb{E}_x\{Z\} = O(x^\alpha)$. ∎

REFERENCES

[1] D. Anick, D. Mitra, and M. M. Sondhi, "Stochastic theory of a data handling system with multiple sources," *Bell Systems Tech. J.*, vol. 61, pp. 1871–1894, Oct. 1982.

[2] J. Abate, G. L. Choudhury, and W. Whitt, "Asymptotics for steady-state tail probabilities in structured Markov queueing models," *Commun. Statist. Stochastic Models*, vol. 10, no. 1, pp. 99–143, 1994.

[3] R. G. Addie and M. Zukerman, "An approximation for performance evaluation of stationary single server queues," *IEEE Trans. Commun.*, vol. 42, pp. 3150–3160, Dec. 1994.

[4] C.-S. Chang, "Stability, queue length, and delay of deterministic and stochastic queueing networks," *IEEE Trans. Automat. Contr.*, vol. 39, pp. 913–931, May 1994.

[5] N. G. Duffield and N. O'Connell, "Large deviations and overflow probabilities for the general single server queue, with application," *Proc. Cambridge Philosoph. Soc.*, vol. 118, pp. 363–374, 1995.

[6] G. L. Choudhury, D. M. Lucantoni, and W. Whitt, "Squeezing the most out of ATM," *IEEE Trans. Commun.*, vol. 44, pp. 203–217, Feb. 1996.

[7] P. W. Glynn and W. Whitt, "Logarithmic asymptotics for steady-state tail probabilities in a single-server queue," *J. Appl. Probabil.*, pp. 131–155, 1994.

[8] A. Baiocchi, N. Melazzi, M. Listani, A. Roveri, and R. Winkler, "Loss performance analysis of an ATM multiplexer loaded with high speed ON–OFF sources," *IEEE J. Select. Areas Commun.*, vol. 9, pp. 388–393, Apr. 1991.

[9] N. Likhanov and R. R. Mazumdar, "Cell-loss asymptotics in buffers fed with a large number of independent stationary sources," in *Proc. IEEE Infocom*, San Francisco, CA, 1998, p. .

[10] N. B. Shroff and M. Schwartz, "Improved loss calculations at an ATM multiplexer," *IEEE/ACM Trans. Networking*, vol. 6, pp. 411–422, Aug. 1998.

[11] C. Courcoubetis, V. A. Siris, and R. Weber, "Investigation of cell scale and burst scale effects on the cell loss probability using large deviations," presented at the 15th U.K. Workshop Performance Engineering of Computer and Telecommunication Systems, Ilkley, U.K., July 1997.

[12] R. M. Loynes, "The stability of a queue with nonindependent inter-arrival and service times," *Proc. Cambridge Philosoph. Soc.*, vol. 58, pp. 497–520, 1962.

[13] J. W. Cohen, *The Single Server Queue*. Amsterdam, The Netherlands: North-Holland, 1969.

[14] D. P. Heyman and W. Whitt, "Limits for queues as the waiting room grows," *Queueing Systems*, vol. 5, pp. 381–392, 1989.

[15] A. Brandt, P. Franken, and B. Lisek, *Stationary Stochastic Models*. New York: Wiley, 1990.

[16] A. P. Zwart, "A fluid queue with a finite buffer and subexponential input," *Adv. Appl. Probabil.*, vol. 32, pp. 221–243, Mar. 2000.
[17] H. S. Kim and N. B. Shroff. (2000) On the asymptotic relationship between the overflow probability in an infinite queue and the loss probability in a finite queue. Tech. Rep., Purdue Univ., West Lafayette, IN. [Online]. Available: http://yara.ecn.purdue.edu/~newsgrp/publication/hkim/tr00.ps
[18] J. Choe and N. B. Shroff, "A central limit theorem based approach to analyze queue behavior in ATM networks," in *Proc. 15th Int. Teletraffic Congr.*, Washington, DC, 1997, pp. 1129–1138.
[19] ——, "A new method to determine the queue length distribution at an ATM multiplexer," in *Proc. IEEE Infocom*, Kobe, Japan, 1997, pp. 550–557.
[20] ——, "A central limit theorem based approach for analyzing queue behavior in high-speed networks," *IEEE/ACM Trans. Networking*, vol. 6, pp. 659–671, Oct. 1998.
[21] ——, "On the supremum distribution of integrated stationary Gaussian processes with negative linear drift," *Adv. Appl. Probabil.*, vol. 31, no. 1, pp. 135–157, Mar. 1999.
[22] B. Maglaris, D. Anastassiou, P. Sen, G. Karlsson, and J. D. Robbins, "Performance models of statistical multiplexing in packet video communication," *IEEE Trans. Commun.*, vol. 36, pp. 834–843, July 1988.
[23] I. Norros, "On the use of fractal Brownian motion in the theory of connectionless networks," *IEEE J. Select. Areas Commun.*, vol. 13, pp. 953–962, Aug. 1995.
[24] N. B. Shroff and M. Schwartz, "Improved loss calculations at an ATM multiplexer," in *Proc. IEEE Infocom*, vol. 2, San Francisco, CA, Mar. 1996, pp. 561–568.
[25] J. Beran, R. Sherman, M. S. Taqqu, and W. Willinger, "Long-range dependence in variable-bit-rate video traffic," *IEEE Trans. Commun.*, vol. 43, no. 2–4, pp. 1566–1579, Feb.–April 1995.
[26] W. E. Leland, M. S. Taqqu, W. Willinger, and D. V. Wilson, "On the self-similar nature of ethernet traffic (extended version)," *IEEE/ACM Trans. Networking*, vol. 2, pp. 1–15, Feb. 1994.
[27] J. Choe and N. B. Shroff, "Queueing analysis of high-speed multiplexers including long-range dependent arrival processes," in *Proc. IEEE Infocom*, New York, Mar. 1999.
[28] ——, "Use of supremum distribution of Gaussian processes in queueing analysis with long-range dependence and self-similarity," *Stochastic Models*, vol. 16, no. 2, Feb. 2000.
[29] I. Sidhu and S. Jordan, "Multiplexing gains in bit stream multiplexors," *IEEE/ACM Trans. Networking*, vol. 3, pp. 785–797, Dec. 1995.
[30] D. D. Botvich and N. G. Duffield, "Large deviations, the shape of the loss curve, and economies of scale in large multiplexers," *Queueing Systems*, vol. 20, pp. 293–320, 1995.
[31] N. G. Duffield, "Economies of scale in queues with sources having power-law large deviation scaling," *J. Appl. Probabil.*, vol. 33, pp. 840–857, 1996.
[32] A. Simonian and J. Guilbert, "Large deviations approximation for fluid queues fed by a large number of on–off sources," *IEEE J. Select. Areas Commun.*, vol. 13, pp. 1017–1027, Aug. 1994.
[33] C.-S. Chang, P. Heidelberger, S. Juneja, and P. Shahabuddin, "Effective bandwidth and fast simulation of ATM intree networks," *Perform.Eval.*, vol. 20, pp. 45–65, 1994.
[34] J. N. Daigle and J. D. Langford, "Models for analysis of packet voice communication systems," *IEEE J. Select. Areas Commun.*, vol. 4, pp. 847–855, Sept. 1986.
[35] K. Sriram and W. Whitt, "Characterizing superposition arrival processes in packet multiplexer for voice and data," *IEEE J. Select. Areas Commun.*, vol. 4, pp. 833–846, Sept. 1986.
[36] E. W. Knightly, "Second moment resource allocation in multi-service networks," in *Proc. ACM SIGMETRICS*, Seattle, WA, 1997, pp. 181–191.
[37] D. Tse, R. G. Gallager, and J. N. Tsitsiklis, "Statistical multiplexing of multiple time-scale Markov streams," *IEEE J. Select. Areas Commun.*, vol. 13, pp. 1028–1039, Aug. 1995.
[38] A. Elwalid, D. Heyman, T. V. Lakshman, D. Mitra, and A. Weiss, "Fundamental bounds and approximations for ATM multiplexers with applications to video teleconferencing," *IEEE J. Select. Areas Commun.*, vol. 13, pp. 1004–1016, Aug. 1995.
[39] T. Lee, K. Lai, and S. Duann, "Design of a real-time call admission controller for ATM networks," *IEEE/ACM Trans. Networking*, vol. 4, pp. 758–765, Oct. 1996.
[40] A. I. Elwalid and D. Mitra, "Effective bandwidth of general Markovian traffic sources and admission control of high speed networks," *IEEE/ACM Trans. Networking*, vol. 1, pp. 329–343, June 1993.
[41] J. Choe, "Queueing analysis of high-speed networks with Gaussian traffic models," Ph.D. dissertation, Purdue Univ., West Lafayette, IN, Aug. 1998.
[42] B. Tsybakov and N. D. Georganas, "Self-similar processes in communications networks," *IEEE Trans. Inform. Theory*, vol. 44, pp. 1713–1725, Sept. 1998.
[43] M. Montgomery and G. De Veciana, "On the relevance of time scales in performance oriented traffic characterization," in *Proc. IEEE Infocom*, San Francisco, CA, 1996, pp. 513–520.
[44] D. Eun and N. B. Shroff. The measurement-analytic framework for QoS estimation based on the dominant time scale. presented at Proc. IEEE Infocom 2001. [Online]. Available: http://shay.ecn.purdue.edu/~eun/infocom01.ps.

Han S. Kim received the B.S. degree and the M.S. degree in electronics engineering from the Seoul National University, Seoul, Korea, in 1990, 1992, respectively. From 1992 to 1997, he was with LG Electronics Inc., Seoul, Korea, as a research staff. Since 1998, he has been at Purdue University, West Lafayette, IN, as a Ph.D. student in the School of Electrical and Computer Engineering. His research interests include network traffic modeling and analysis, admission control, resource allocation, and network management.

Ness B. Shroff (S'91–M'93–SM'01) received the B.S. degree from the University of Southern California, Los Angeles, the M.S.E. degree from the University of Pennsylvania, Philadelphia, and the M.Phil and Ph.D. degrees from Columbia University, New York.

He is currently an Associate Professor in the School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN. During his doctoral study, he worked at AT&T Bell Labs in 1991 and Bell Communications Research in 1992, on problems involving fault management in telephone networks. His current research interests are in high-speed broadband and wireless communication networks, especially issues related to performance modeling and analysis, routing, network management, scheduling, and control in such networks. He also works on problems related to source coding, vector quantization, and error concealment.

Dr. Shroff has received research and equipment grants to conduct fundamental work in broadband and wireless networks, and quantization from the National Science Foundation, AT&T, Hewlett Packard, Intel, LG Electronics, Indiana Department of Transportation, the Indiana 21st Century Fund, and the Purdue Research Foundation. He received the NSF CAREER Award from the National Science Foundation in 1996. He has served on the technical program committees of various conferences and on NSF review panels. He was the Conference Chair for the 14th Annual IEEE Computer Communications Workshop (CCW) and is Program Co-Chair for the High-Speed Networking Symposium at Globecom 2000. He is currently on the editorial board of the IEEE/ACM TRANSACTIONS ON NETWORKING and the *Computer Networks* journal. He is a past editor of IEEE COMMUNICATION LETTERS.