A Central-Limit-Theorem-Based Approach for Analyzing Queue Behavior in High-Speed Networks

Jinwoo Choe, *Student Member, IEEE,* and Ness B. Shroff, *Member, IEEE*

***Abstract*— In this paper, we study $\mathbb{P}(Q > x)$, the tail of the steady-state queue length distribution at a high-speed multiplexer. In particular, we focus on the case when the aggregate traffic to the multiplexer can be characterized by a stationary Gaussian process. We provide two asymptotic upper bounds for the tail probability and an asymptotic result that emphasizes the importance of the dominant time scale and the maximum variance. One of our bounds is in a single-exponential form and can be used to calculate an upper bound to the asymptotic constant. However, we show that this bound, being of a single-exponential form, may not accurately capture the tail probability. Our asymptotic result on the importance of the maximum variance and our extensive numerical study on a known lower bound motivate the development of our second asymptotic upper bound. This bound is expressed in terms of the maximum variance of a Gaussian process, and enables the accurate estimation of the tail probability over a wide range of queue lengths. We apply our results to Gaussian as well as multiplexed non-Gaussian input sources, and validate their performance via simulations. Wherever possible, we have conducted our simulation study using *importance sampling* in order to improve its reliability and to effectively capture rare events. Our analytical study is based on *extreme value theory*, and therefore different from the approaches using traditional Markovian and Large Deviations techniques.**

***Index Terms*— Asymptotic upper bound, Gaussian process, queue lenght distribution, strong asymptotics.**

I. Introduction

ADVANCES in lightwave communication technology have enabled high-speed networks, such as the *asynchronous transfer mode* (ATM) networks, to support various real-time applications. Statistical multiplexing is very important in such networks, since it increases network efficiency by allowing a large number of applications to share network resources (e.g., buffer space and link capacity). However, when these resources are shared, there also exists the possibility of excessive congestion, which could impact the quality of the underlying applications. Therefore, a network has to be designed and controlled based on certain measures that reflect the degree of the expected congestion in the network. A fundamental measure of congestion that we study in this paper is $\mathbb{P}(Q > x)$, the tail of the steady-state buffer occupancy (queue length) distribution at a multiplexer.

The tail probability $\mathbb{P}(Q > x)$ has been extensively studied, but usually can be computed exactly only for a limited class of queuing systems. Further, even for traffic sources (such as the *Markov arrival processes* (MAP) or *Markov modulated fluid* (MMF) Processes) for which exact analytical techniques have been developed [14], [22], one quickly runs into classical computational infeasibility problems when the number of multiplexed traffic sources is increased [13], [31]. At the same time, analyzing the queuing system with many multiplexed sources is extremely important, since real networks are expected to support a large number of heterogeneous network applications. To address this problem, a large-scale effort has been devoted to the study of the asymptotic behavior of the tail probability, and a number of approximations for $\mathbb{P}(Q > x)$ have been developed (see [27] for a recent overview of queuing analysis in broadband networks). We next briefly overview related work on the asymptotics of $\mathbb{P}(Q > x)$.

Large deviation techniques have been developed on general mathematical settings and are used to investigate the asymptotic behavior of $\mathbb{P}(Q > x)$. For instance, in [19], the following asymptotic log-similarity ($\overset{\log}{\sim}$) relation has been obtained for $\mathbb{P}(Q > x)$ in considerable generality:[1]

$$\mathbb{P}(Q > x) \overset{\log}{\sim} e^{-\eta x}. \tag{1}$$

Here $f(x) \overset{\log}{\sim} g(x)$ if $\log f(x) \sim \log g(x)$, and $f(x) \sim g(x)$ means $\lim_{x\to\infty} f(x)/g(x) = 1$. The positive constant η in (1) is typically called the *asymptotic decay rate* and can be easily obtained even when the number of traffic sources being multiplexed is very large. Therefore, this result has lead researchers to propose the well known *effective bandwidth* (EB) approximation $\mathbb{P}(Q > x) \approx e^{-\eta x}$ (e.g., see [8] and references therein for more about the EB approximation and its theoretical foundation). However, the great generality of the large deviation techniques comes at a cost: the asymptotic relation in (1) captures only the leading (fastest decaying) term in $\log \mathbb{P}(Q > x)$. For example, there are an infinite number of functions such as $e^{-\eta x+\sqrt{x}}$ and $x^{10}e^{-\eta x}$, which

Manuscript received June 6, 1997; revised June 5, 1998; approved by IEEE/ACM TRANSACTIONS ON NETWORKING Editor R. Guerin. This work was supported in part by the National Science Foundation under CAREER Grant NCR-9624525 and Grant CDA-9422250, and the Purdue Research Foundation under Grant 690-1285-2479.

The authors are with the School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN 47907–1285 USA.

Publisher Item Identifier S 1063-6692(98)07006-X.

[1] This result has been extended to the queues serving long-range dependent input processes (see [15]) in which case, the tail probability may not be asymptotically exponential (even in a log-similar sense). This paper focuses on Gaussian processes but does not cover long-range dependency. Readers that are interested in our work are referred to a more recent study of the tail probability for long-range dependent Gaussian processes [10].

are significantly different from $e^{-\eta x}$ but can replace $e^{-\eta x}$ in (1) to result in another valid log-similar relation.

To alleviate the poor "resolution" of log-similarity, a stronger form of asymptotics has also been developed for different classes of queuing systems. These asymptotics show that (1) can be significantly strengthened to obtain a similarity ($\sim$) relation (e.g., see [1], [2], [20], [32]); i.e.,

$$\mathbb{P}(Q > x) \sim Ce^{-\eta x}. \tag{2}$$

Here, C is a positive constant called the *asymptotic constant.* From this stronger asymptotic relation, the *asymptotic approximation* $\mathbb{P}(Q > x) \approx Ce^{-\eta x}$, has been suggested for large values of x (e.g., see [1], [2], [13], [20]). Unlike the EB approximation (which can also be obtained by setting $C = 1$ above), it has been shown that the asymptotic approximation does account for statistical multiplexing. The reason is that the effect of statistical multiplexing is captured by the asymptotic constant C [13], [31], and not by the asymptotic decay rate η. Unfortunately, unlike the asymptotic decay rate η, the exact value of the asymptotic constant C cannot usually be determined (especially when a large number of traffic sources are multiplexed). Hence, methods have been developed to approximate C for special cases (e.g., see [2], [13], [16], [31]).

In this paper, we focus on the case when the input process is stationary Gaussian. Gaussian process modeling is useful for two main reasons. First, Gaussian processes have several appealing properties. For example, independent Gaussian processes are closed under superposition, and any stationary Gaussian process can be completely specified by its mean and autocovariance. Therefore, unlike the case of MMF processes, analyzing a queue with a large number of Gaussian input processes is no more difficult than analyzing a queue with a single Gaussian input process. Second, and more importantly, the large bandwidth (compared to the bandwidth required by a typical network application) of high-speed networks make it a natural approximation for the aggregate input process. Due to the huge capacity of network links, hundreds or even thousands of network applications are likely to be served by a multiplexer. Therefore, even when the traffic from each individual application cannot be characterized by a Gaussian process, by appealing to the *central limit theorem,* the aggregate traffic to the multiplexer can be effectively modeled as a Gaussian process.

Such queues (fed by a stationary Gaussian input process) have recently received some attention (e.g., see [2], [9], [25], [26]). We already know from [19], that the log-similarity relation (1) holds for Gaussian processes. The excellent work by Addie and Zuckerman [2] strengthens this result by showing that for fairly general discrete-time Gaussian sources, the tail probability is in the form of (2). They also suggest possible approximations of the asymptotic constant C. In [26], Norros provides an approximation to determine the tail probability for the special case of *fractal Brownian motion.* In this case, the asymptotic behavior of the tail probability is not in the form of (2).

We will provide two asymptotic upper bounds for $\mathbb{P}(Q > x)$ for a large class of Gaussian processes for which (2) holds. Our approach is quite novel: it is based on *extreme value theory* for Gaussian processes [4] and is different from traditional Markovian or large deviation techniques. One of our bounds is of a single exponential form and results in an accurate *upper bound* to the asymptotic constant C. For the reason mentioned earlier, this bound (as an accurate estimate for the asymptotic constant) is important in effectively exploiting the statistical multiplexing gain. Further, since the upper bound is obtained as a simple expression in terms of the autocovariance function of the input process, it gives us important insights into the relationship between the correlation structure of an input process and its queuing behavior. In spite of the theoretical value of our single-exponential asymptotic upper bound, we show that it suffers from the same limitation inherent in all single-exponential based approximations for $\mathbb{P}(Q > x)$; when the tail probability converges to its asymptote slowly, a single exponential approximation may fail to accurately approximate $\mathbb{P}(Q > x)$ even for fairly large values of x. To address this problem, we introduce another asymptotic upper bound which is asymptotically similar to the first bound, but also accurately captures the tail probability over a wide range of queue lengths x. The development of the second asymptotic upper bound is motivated by our past numerical studies on a well known lower bound[2] and a theoretical result (Theorem 2). This theoretical result also serves to emphasize the importance of the dominant time scale in queuing analysis for Gaussian sources. We further provide an extensive numerical study involving importance sampling and actual video traces to demonstrate the accuracy of our analytical results.

Here, we should distinguish our work in this paper from some results in the literature. All of the above discussion (including the work in this paper) is about "x-asymptotics" i.e., the asymptotic behavior of $\mathbb{P}(Q > x)$, as the queue length x increases. There has been recent work that focuses on the asymptotic behavior of $\mathbb{P}(Q > x)$ when the number of sources, the queue length, and the service rate are all proportionally sent to infinity (e.g., [6], [25]). We classify these studies as M-asymptotics, where M represents the number of sources in the system. In particular, Montgomery and De Veciana [25] have significantly strengthened the corresponding log-similarity relation in [6] using the Bahadur–Rao asymptotics, and obtained asymptotic bounds for the tail probability. However, note that M-asymptotics considers a limit in a different direction from that in x-asymptotics. Therefore, results in M-asymptotics cannot be extended to x-asymptotics (and *vice versa*) unless very strong properties such as uniformity of convergence can be shown (which is usually not the case). Hence, the results in this paper belong to a different category, from those in M-asymptotics.

As an important final note, due to space limitations, we do not provide any proofs to the theoretical results in this paper. Interested readers are referred to our technical report [11].

II. Preliminaries

A. Fluid Queue Model

We model a high-speed statistical multiplexer by an infinite buffer fluid queue shown in Fig. 1. The fluid queue consists of

[2] As will be described in Section IV, approximations equivalent to this lower bound have already been suggested (e.g., see [25], [27]).

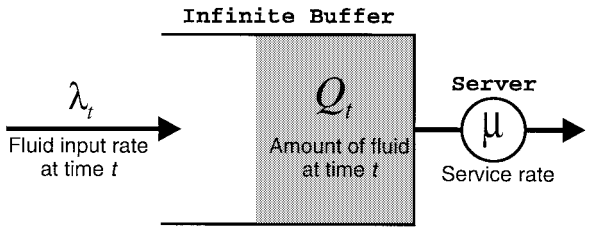

Fig. 1. A typical fluid queue model.

a server that drains the fluid from the buffer at a constant rate μ, and a fluid input that fills the buffer at a rate λ_t. The fluid input λ_t corresponds to the aggregate arrival process to a high-speed multiplexer, and μ corresponds to the rate at which fixed size packets (such as ATM cells) are transmitted onto the link. Consequently, Q_t, the amount of fluid in the buffer at time t, represents the number of cells in the multiplexer.

Depending on the index set T, from which the time index t takes its value, a fluid queue is classified as either a continuous-time fluid queue ($T = (-\infty, \infty)$) or a discrete-time fluid queue ($T = \{\cdots, -1, 0, 1, \cdots\}$). In this paper, we only consider discrete-time fluid queues, although equivalent results can also be obtained for the continuous-time case [12].

In a discrete-time fluid queue, the evolution of Q_n, the amount of fluid in the buffer, can be expressed by Lindley's equation:

$$Q_n = (Q_{n-1} + \gamma_n)^+ \tag{3}$$

where $\gamma_n := \lambda_n - \mu$ is the net amount of fluid input at time n and $(x)^+ := \max\{0, x\}$. In [24], it has been shown under some mild assumptions (such as the stationarity and ergodicity of γ_n and the stability condition, i.e., $\mathbb{E}\{\gamma_n\} < 0$), that the distribution of Q_n determined by (3) converges to a unique limiting distribution (the steady-state queue distribution) as n goes to infinity, regardless of the initial condition Q_0. In addition, it has been shown that the supremum distribution of $\{X_n: n = 0, 1, \cdots\}$ defined by $X_n := \sum_{m=1}^{n} \gamma_{-m}$, is equal to the steady-state queue length distribution, i.e.,

$$\mathbb{P}(Q > x) = \mathbb{P}\left(\sup_{n \geq 0} X_n > x\right). \tag{4}$$

This relation, which originally comes from [24], has played a key role in obtaining a number of important results on the steady-state queue length (or waiting time) distribution.

From here on, throughout this paper, we focus on the cases for which the aggregate arrival process λ_n (and hence γ_n) can be characterized by a stationary Gaussian process.

B. Important Notations and Definitions

Let $C_\gamma(l)$ denote the autocovariance function of the stationary Gaussian net input process $\gamma_n = \lambda_n - \mu$ (note that $C_\gamma(l) = C_\lambda(l)$ since we set the service rate to a constant μ). It is easy to see from the definition of X_n, that it is also a Gaussian process. The mean and autocovariance function of X_n can be computed in terms of $\kappa := -\mathbb{E}\{\gamma_0\}$ and $C_\gamma(l)$ as $\mathbb{E}\{X_n\} = -\kappa n$, and $C_X(n_1, n_2) = \sum_{l_1=1}^{n_1} \sum_{l_2=1}^{n_2} C_\gamma(l_2 - l_1)$. By a change of variables $l = l_2 - l_1$, the variance of X_n can be expressed as a weighted sum of $C_\gamma(l)$, i.e., $\mathrm{Var}\{X_n\} = nC_\gamma(0) + 2\sum_{l=1}^{n-1}(n-l)C_\gamma(l)$.

Note that $\mathrm{Var}\{X_n\}$ can also be expressed in terms of $\mathbf{I}_{\mathrm{dc}}(n) := \mathrm{Var}\{\sum_{m=1}^{n} \lambda_n\}/\mathbb{E}\{\sum_{m=1}^{n} \lambda_n\}$, the (generalized) *index of dispersion for counts* (IDC) by the relation $\mathrm{Var}\{X_n\} = n(\mu - \kappa)\mathbf{I}_{\mathrm{dc}}(n)$. Assuming that the net input process γ_n is stationary Gaussian, its distribution is completely determined by either κ and $\mathbf{I}_{\mathrm{dc}}(n)$, or κ and $\mathrm{Var}\{X_n\}$. Therefore, this paper also falls into the classification of queuing analysis based on the mean and the index of dispersion of the input traffic (e.g., [18], [34]).

For notational simplicity, for each $x > 0$, we define a new stochastic process $Y_n^{(x)} := [\sqrt{x}(X_n + \kappa n)]/(x + \kappa n)$. It then follows that for any $x > 0$ and any $n \in \{0, 1, 2, \cdots\}$

$$X_n > x \quad \text{if and only if} \quad Y_n^{(x)} > \sqrt{x}. \tag{5}$$

Hence, from (4), we have $\mathbb{P}(Q > x) = \mathbb{P}(\sup_{n \geq 0} Y_n^{(x)} > \sqrt{x})$. Note that for each x, $Y_n^{(x)}$ is a centered Gaussian process, and its autocovariance function $C_{Y^{(x)}}$ is given by

$$C_{Y^{(x)}}(n_1, n_2) = \frac{xC_X(n_1, n_2)}{(x + \kappa n_1)(x + \kappa n_2)}.$$

Further, $\sigma_{x,n}^2$, the variance of $Y_n^{(x)}$, can be written as

$$\sigma_{x,n}^2 = \frac{x\mathrm{Var}\{X_n\}}{(x + \kappa n)^2} = \frac{x\left(nC_\gamma(0) + 2\sum_{l=1}^{n-1}(n-l)C_\gamma(l)\right)}{(x + \kappa n)^2}. \tag{6}$$

Henceforth, we let $\langle w \rangle_\Theta$ denote $\sup_{\theta \in \Theta} w_\theta$. Moreover, we do not specify the index range Θ when it includes the entire domain of w_θ. For example, $\langle \sigma_x^2 \rangle$ represents the supremum of $\sigma_{x,n}^2 = \mathrm{Var}\{Y_n^{(x)}\}$ over $n \in \{0, 1, 2, \cdots\}$ (the index omitted in $\langle \cdot \rangle$), and $\langle Y^{(x)} \rangle_{[a,b]}$ represents the supremum of $Y_n^{(x)}$ over $n \in [a, b]$. We now list three important conditions on $C_\gamma(l)$, and state three important propositions (we provide detailed proofs in [11]) which will be referred to later in the paper:

$$\sum_{l=-\infty}^{\infty} |C_\gamma(l)| < \infty \quad \text{and} \quad \sum_{l=-\infty}^{\infty} C_\gamma(l) > 0 \tag{C1}$$

$$\sum_{l=-\infty}^{\infty} |lC_\gamma(l)| < \infty \tag{C2}$$

$$\sum_{l=1}^{m} lC_\gamma(l) + \sum_{l=m+1}^{\infty} mC_\gamma(l) > 0, \qquad \forall m = 1, 2, \cdots,$$

and

$$\sum_{l=1}^{\infty} lC_\gamma(l) > 0. \tag{C3}$$

Proposition 1: Let k_i and l_i be two nonnegative sequences such that $k_i, l_i \to \infty$ and $k_i/l_i \to \alpha \geq 1$ as $i \to \infty$. Then, under condition (C1),

$$\lim_{i \to \infty} \frac{C_X(k_i, l_i)}{l_i} = \lim_{i \to \infty} \frac{C_X(l_i, k_i)}{l_i} = S$$

where $S := \sum_{l=-\infty}^{\infty} C_\gamma(l)$. In particular, $\lim_{n \to \infty} \mathrm{Var}\{X_n\}/n = S$.

Proposition 2: Define $\hat{n}_x$ to be the time at which $\sigma^2_{x,n}$ attains its maximum $\langle\sigma^2_x\rangle$. Then, under condition (C1),

$$\hat{n}_x \sim \frac{x}{\kappa}.$$

Proposition 3: Under condition (C1), $\lim_{x\to\infty}\langle\sigma^2_x\rangle = S/4\kappa$.

It should be mentioned that (C1)–(C3) characterize a fairly large class of Gaussian processes. Condition (C1) is mainly on the absolute summability of the autocovariance function of the input process. Hence, a sufficient condition for (C1) [assuming $\sum_{l=-\infty}^{\infty} C_\lambda(l) > 0$] is that there exists an $\epsilon > 1$ such that $C_\lambda(l) < l^{-\epsilon}$ for all sufficiently large l. It should be noted that condition (C1) can be thought of as the boundary between the processes that exhibit long-range dependence and those that do not (see [5], [23] for the definition and properties of long-range dependence and/or self-similarity). In other words, under this condition the tail probability satisfies (2) with $\eta = 2\kappa/S$ and some finite constant C [2].

Condition (C2) is on the absolute summability of a weighted autocovariance function of the input process. This condition is somewhat more restrictive than (C1), and satisfied if there exists an $\epsilon > 2$ such that $C_\lambda(l) < l^{-\epsilon}$, for all sufficiently large l.

While (C1) and (C2) are related to the decay rate of an autocovariance function, condition (C3) is related to its shape and sign. Roughly speaking, (C3) is satisfied when $C_\lambda(l)$, the autocovariance function of an input process, is positive for most values of l. The class of input processes characterized by (C3) is very important for the analysis of network delay, since positive autocovariance is related to the bursty nature of an input process, which in turn is the main cause of network congestion.

III. Single Exponential Asymptotic Upper Bound

In this section, we introduce our first asymptotic upper bound for $\mathbb{P}(Q > x)$ expressed as an exponential function of x, and illustrate its theoretical importance. We say that $f(x)$ asymptotically bounds $g(x)$ from above if $\limsup_{x\to\infty} g(x)/f(x) \le 1$. We also briefly discuss its performance as an approximation for $\mathbb{P}(Q > x)$ through numerical examples.

It should be noted here that Simonian [33] has derived an elegant upper bound in an integral form for general continuous-time fluid queues fed by input processes having density function. However, in spite of its significant theoretical value, the upper bound usually results in a fairly complicated expression when it is evaluated for a specific fluid queue. Moreover, the asymptotic behavior of this upper bound has only been shown to be exponential for the *Ornstein–Uhlenbeck* process. For more general processes we do not even know if the bound is asymptotically log-similar to the tail probability, thus limiting its practical value.

In contrast, the asymptotic upper bound for $\mathbb{P}(Q > x)$ that we introduce in this section is in a simple exponential form which can easily be obtained from the mean and autocovariance of the net input Gaussian process. Although it is not a global upper bound, but an asymptotic upper bound, it is of both theoretical and practical importance, as will be discussed shortly.

This section proceeds as follows. We first make some interesting observations by time-scaling the stochastic process $Y^{(x)}_n$. These observations provide some insight on the behavior of $\mathbb{P}(Q > x)$ and point us in the development of our asymptotic upper bound.

A. Interpretation of Time-Scaling $Y^{(x)}_n$

Consider a continuous-time stochastic process $\tilde{Y}^{(x)}_t$ defined for each $x > 0$ as $\tilde{Y}^{(x)}_t := Y^{(x)}_{\lfloor xt/\kappa\rfloor}$, where $\lfloor z\rfloor$ denotes the largest integer that is smaller than or equal to z. The stochastic process $\tilde{Y}^{(x)}_t$ is simply an interpolated (by holding its value for a period of length κ/x) and scaled (in time) version of $Y^{(x)}_n$, that is enforced to attain its maximum variance around $t = 1$, as $x \to \infty$ (see Proposition 2). From the definition of $\tilde{Y}^{(x)}_t$, the following can easily be verified:

$$\mathbb{P}(Q > x) = \mathbb{P}\left(\langle\tilde{Y}^{(x)}\rangle > \sqrt{x}\right) \tag{7}$$

$$\lim_{x\to\infty} C_{\tilde{Y}^{(x)}}(t_1, t_2) = \frac{S\min\{t_1, t_2\}}{\kappa(1+t_1)(1+t_2)} \quad \text{(from Proposition 1).} \tag{8}$$

Since $\tilde{Y}^{(x)}_t$ is a centered Gaussian process for each $x > 0$, (8) implies that, as $x \to \infty$,

$$\tilde{Y}^{(x)}_t \to U_t := \frac{\sqrt{S}\,B_t}{\sqrt{\kappa}(1+t)} \qquad \text{in distribution} \tag{9}$$

where B_t is the standard Brownian motion process.

Now, we briefly move our attention to continuous-time fluid queues. For continuous-time fluid queues, continuous-time stochastic processes X_t, $Y^{(x)}_t$, and $\tilde{Y}^{(x)}_t$ can be defined in an analogous way to their discrete-time counterparts:

$$X_t := \Gamma_0 - \Gamma_{-t},$$

$$Y^{(x)}_t := \frac{\sqrt{x}\,(X_t + \kappa t)}{x + \kappa t}$$

and

$$\tilde{Y}^{(x)}_t := Y^{(x)}_{xt/\kappa}.$$

Here, Γ_t is a stochastic process with stationary increments and negative drift such that $\Gamma_t - \Gamma_s (s \le t)$ represents the net input into a fluid queue during the interval $(s, t]$, and $\kappa := -[(\mathbb{E}\{\Gamma_t - \Gamma_s\})/(t-s)]$. Remember that the results [including (7)] obtained for discrete-time fluid queues can also be derived for continuous-time fluid queues [12]. Also, note that if we set $\Gamma_t = \sqrt{S}\,B_t - \kappa t$ (which corresponds to an uncorrelated input process) $\tilde{Y}^{(x)}_t$ would have the same distribution as U_t. This fact together with (7) and (9) indicates that as x increases, $\mathbb{P}(Q > x)$ behaves as if the fluid queue is driven by a completely uncorrelated input process, regardless of the correlation structure of the actual input process.

This phenomenon can be intuitively interpreted as follows. From Proposition 2, $\hat{n}_x$, the time at which X_n ($Y^{(x)}_n$) is most likely to be larger than x ($\sqrt{x}$) increases linearly with x. Therefore, as x increases, $\hat{n}_x$ eventually becomes significantly larger than the time scale over which the net input process

is correlated. As a result, the effect of the correlated input process is negligible on the time scale of $\hat{n}_x$, and $Y_n^{(x)}$ behaves as if the input is an uncorrelated Gaussian process (with the same value of S as the original input process). For instance, let χ_n be an i.i.d. Gaussian process and let $\zeta_n = 0.5\chi_n + 0.3\chi_{n-1} + 0.2\chi_{n-2}$. Then, although χ_n is not correlated, ζ_n is a correlated process. However, if we compare the two partial sums, $\sum_{m=1}^{n} \chi_m$ and $\sum_{m=1}^{n} \zeta_m$ over a much larger time scale (say $n > 100$) than the time scale over which ζ_n is correlated, the difference, $0.5(\chi_0 - \chi_n) + 0.2(\chi_{-1} - \chi_{n-1})$ between these sums becomes very minor. Therefore, we can expect that these two partial sums will exhibit very similar stochastic behavior for such large values of n.

The above discussion suggests the following simple approximation for the tail probability:

$$\begin{aligned}\mathbb{P}(Q > x) &= \mathbb{P}\left(\langle \tilde{Y}^{(x)} \rangle > \sqrt{x}\right), \qquad \text{from (7)} \\ &\approx \mathbb{P}\left(\langle U \rangle > \sqrt{x}\right), \qquad \text{from (9)} \\ &= \mathbb{P}\left(\left\{B_t > \sqrt{\frac{\kappa x}{S}}\,(t+1), \qquad \text{ever}\right\}\right) \\ &= \exp\left(-\frac{2\kappa x}{S}\right), \qquad \text{e.g., see [29, p. 199].} \quad (10)\end{aligned}$$

Since η in (1) and (2) has been shown to be $2\kappa/S$ [2], [19], (10) corresponds to the famous EB approximation. This means that to go beyond the EB approximation and obtain some information about the asymptotic constant in (2), more than the limiting distribution of $\tilde{Y}_t^{(x)}$ has to be considered. The asymptotic upper bound that we now introduce, can be obtained by capturing the way in which the variance of $\tilde{Y}_t^{(x)}$ converges to its limiting value.

B. Single-Exponential Asymptotic Upper Bound

By observing how the variance of $\tilde{Y}_t^{(x)}$ converges to that of U_t around the time ($\kappa\hat{n}_t/x \approx 1$, from Proposition 2) at which the variance of $\tilde{Y}_t^{(x)}$ attains its maximum, we get the following theorem.

Theorem 1: Under conditions (C1)–(C3),

$$\limsup_{x\to\infty} \exp\left(\frac{2\kappa x}{S}\right)\mathbb{P}(Q > x) \le \exp\left(-\frac{2\kappa^2 D}{S^2}\right)$$

where $D := 2\sum_{l=1}^{\infty} lC_\gamma(l)$.

Proof: Refer to [11, Theorem 3.2]. ∎

Theorem 1 gives us an exponential asymptotic upper bound $\exp[-(2\kappa/S)(x + \kappa D/S)]$ to the tail probability $\mathbb{P}(Q > x)$. Further, since it has been shown under condition (C1) that (2) holds for stationary Gaussian input processes with $\eta = 2\kappa/S$ [2], Theorem 1 also provides us with an upper bound $\exp[-(2\kappa^2 D/S^2)]$ to the asymptotic constant C. The asymptotic upper bound accounts for statistical multiplexing in the sense that the bound for the asymptotic constant decreases exponentially when more sources are multiplexed. For instance, consider a fluid queuing system serving M identical input processes with an infinite buffer and a fixed service rate μ per input, and let $\mathbb{P}(Q^M > x)$ denote the corresponding tail probability. Then, the bound for the asymptotic constant of $\mathbb{P}(Q^M > x)$ can be written as $\exp[-(2\kappa^2 DM/S^2)]$ where κ, S, and D are defined by the first two moments of a single input process and the service rate μ per input. Note that the bound decreases exponentially as M increases. Therefore, if we quantitatively define statistical multiplexing gain as the reciprocal of the asymptotic constant, then this gain increases *at least* exponentially with the system size. This result, in fact, supports the behavior of the asymptotic constant that has been observed in empirical studies (e.g., see [13, eq. (1.6)]).

The form of the upper bound to the asymptotic constant gives us more insight into the queuing behavior for stationary Gaussian input processes. It is well known that S, in conjunction with κ, determines the asymptotic decay rate η given in (2) [2], [19]. Further, the limiting value of the IDC of an input process [i.e., $\lim_{n\to\infty} \mathbf{I}_{\rm dc}(n) = S/(\mu - \kappa)$] can also be expressed in terms of S [3]. Therefore S can be thought of as a measure of the "burstiness" of the input process, which is invariant to filtering or finite time-shifting of the arrival process. For example, let $a_m \in [0, 1]$ be a sequence that sums to 1, and consider a linear smoothing system which delays the a_m portion of the input at time n by $m \ge 0$. Then, the output process λ'_n can be expressed as a convolution of a_m and the input process λ_n, i.e., $\lambda'_n = \sum_{m=0}^{\infty} a_m \lambda_{n-m}$. From this relation, the autocovariance function of λ'_n can be computed as $C_{\lambda'}(l) = \sum_{m_1=0}^{\infty}\sum_{m_2=0}^{\infty} a_{m_1}a_{m_2}C_\lambda(l + m_1 - m_2)$. Hence, we have

$$\begin{aligned}\sum_{l=-\infty}^{\infty} C_{\lambda'}(l) &= \sum_{m_1=0}^{\infty} a_{m_1} \sum_{m_2=0}^{\infty} a_{m_2} \sum_{l=-\infty}^{\infty} C_\lambda(l) \\ &= \sum_{l=-\infty}^{\infty} C_\lambda(l).\end{aligned}$$

In other words, since the system does not impose an infinite amount of delay (that is, $\sum_{m=0}^{\infty} a_m = 1$), the autocovariance function of the input process and that of the output process have the same sum. On the other hand, $\sum_{l=0}^{\infty} lC_\lambda(l)$ could be quite different from $\sum_{l=0}^{\infty} lC_{\lambda'}(l)$. In other words, the parameter D is not invariant to filtering or finite time-shifting, and many autocovariance functions with the same S may have very different values of D. Now, consider two nonnegative autocovariance functions $C_1(l)$ and $C_2(l)$ having the same sum S. The autocovariance function $C_1(l)$ has most of its mass distributed close to $l = 0$, while $C_2(l)$ has its mass spread over a wider range of l. In this case, it is obvious from the definition of D, that $C_1(l)$ will have a smaller value of D than $C_2(l)$. In other words, for the same amount of total burstiness in the arrival process, the more the burstiness is spread over time, the larger is the corresponding value of D. Hence, from our bound to the asymptotic constant, the larger is the eventual statistical multiplexing gain. This implies that for a given constraint on the tail probability, by spreading the burstiness over time (e.g., the familiar smoothing concept [30]), we can get better statistical multiplexing gain. In the following section, we will show just how dramatic the difference in this gain can be for two different Gaussian processes having the same value of S.

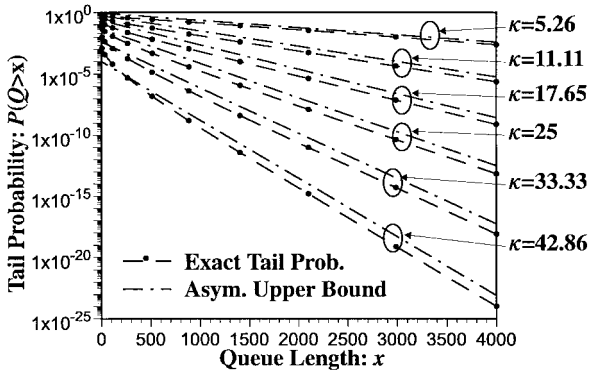

Fig. 2. The exact tail probability and the asymptotic upper bound for a Gaussian input process with autocovariance function $C_\lambda(l) = 200 \times 0.95^{|l|}$.

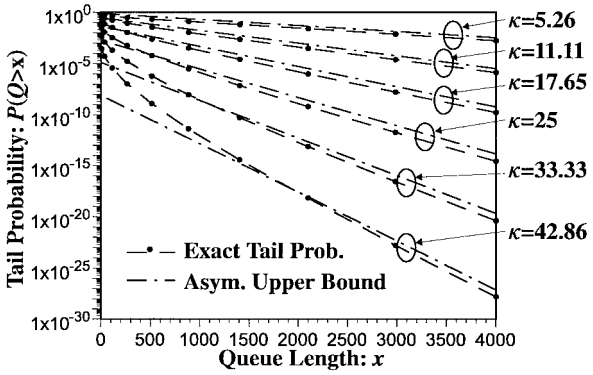

Fig. 3. The exact tail probability and the asymptotic upper bound for a Gaussian input process with $C_\lambda(l) = 100 \times 0.9^{|l|} + 60 \times 0.98^{|l|}$.

C. Numerical Examples and Discussion

In this section, we experimentally investigate the performance of the asymptotic upper bound as an approximation to the tail probability. To validate our results numerically, we use the *Importance Sampling* simulation technique described in [7] (see [21] for a general overview of Importance Sampling techniques). We have calculated 95% confidence intervals for each tail probability estimated via simulation by the method of batch mean. However, to not unnecessarily clutter the figures *we only show confidence intervals when they are larger than ±20% of the estimated tail probability.*

Example 1: In this example, we consider fluid queues fed by two different Gaussian input processes. In particular, in Figs. 2 and 3, we show the exact tail probability and the asymptotic upper bound for two Gaussian input processes with the autocovariance functions $200 \times 0.95^{|l|}$ and $100 \times 0.9^{|l|} + 60 \times 0.98^{|l|}$, respectively, for six different values of κ. Note that these autocovariance functions are nonnegative and vanish exponentially as l increases, so that they satisfy conditions (C1)–(C3). Therefore, from Theorem 1, an exponential asymptotic upper bound for the tail probability can be computed for these two Gaussian sources. As one can see in Fig. 2, for large x, the asymptotic upper bound parallels the tail probability for all values of κ. This is not a surprising result because both the asymptotic upper bound and the tail probability are asymptotically exponential with the same decay rate. Therefore, the logarithmic error between the bound and the tail probability will eventually converge to a finite value. Further note that the bound matches the simulation results quite well. This indicates that the limiting error will be fairly small, and that $\exp[-(2\kappa^2 D/S^2)]$, the upper bound for the asymptotic constant is an accurate estimate of the asymptotic constant. The same observations can be made in Fig. 3; the asymptotic upper bound parallels the tail probability as x increases and the difference between the bound and the exact tail probability is less than an order of magnitude for large enough values of x. However, in Fig. 3, the asymptotic upper bound fails to approximate the tail probability for small queue lengths (<500) for $\kappa = 33.33$, 42.86. This is because the tail probability in Fig. 3 converges to its exponential asymptote slowly, while the tail probability in Fig. 2 converges to its asymptote fairly fast, and forms a nearly straight line. Note that the autocovariance function of the Gaussian input used in Fig. 3 consists of two power terms with different decay rates. Hence, the input is correlated at different time scales, which typically results in a slower convergence of the tail probability to its asymptote. In the following example, a far more significant effect of this multiple time-scale correlation is demonstrated.

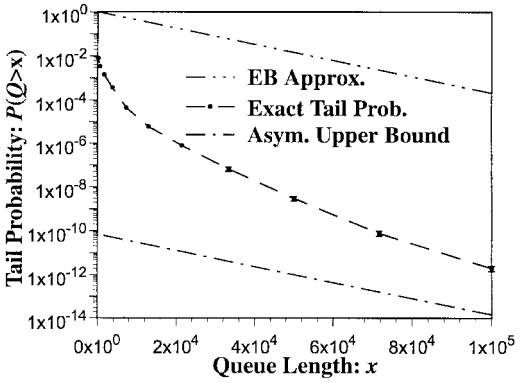

Fig. 4. The exact tail probability, the EB approximation, and the asymptotic upper bound for a Gaussian input process with $C_\lambda(l) = 104 \times 0.99^{|l|} + 64.14 \times 0.999^{|l|} + 31.86 \times 0.9999^{|l|}$ when $\kappa = 33.33$.

Example 2: In this example we consider a fluid queue fed by a Gaussian input process with autocovariance function $C_\lambda(l) = 104 \times 0.99^{|l|} + 64.14 \times 0.999^{|l|} + 31.86 \times 0.9999^{|l|}$. As can be observed, the autocovariance function is a sum of three weighted powers with very different decay rates. This means that the source is correlated at very different time scales. In Fig. 4, the asymptotic upper bound, the EB approximation and simulation results are shown for $\kappa = 33.33$. Note that the slope of the simulation curve significantly differs from that of the EB approximation (or the asymptotic upper bound) even at $x = 10^5$. This implies that the tail probability is not close to its asymptote over the entire range of queue lengths shown in the figure. Even though we cannot calculate the exact asymptote given in (2), we know that it has to be below the asymptotic upper bound. Therefore, in this case, neither the EB approximation nor the asymptotic approximation can accurately estimate the tail probability even for very large values of x. For example, for the queue length as large as 20 000, the EB approximation overestimates the exact tail

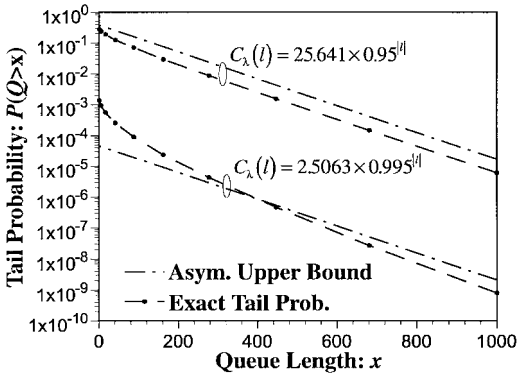

Fig. 5. The exact tail probability and the asymptotic upper bound for two Gaussian input processes with $C_\lambda(l) = 25.641 \times 0.95^{|l|}$, $C_\lambda(l) = 2.5063 \times 0.995^{|l|}$, and $\kappa = 5$.

probability by five orders of magnitude, while the asymptotic approximation underestimates the exact tail probability by at least five orders of magnitude. This also implies that even though the asymptotic upper bound provides a close upper bound to the asymptotic constant (this is found to be true in this case as well by examining larger values of x), since it is in a single exponential form, it may not provide a useful estimate of $\mathbb{P}(Q > x)$ for probabilities of interest. Further, even by using current multi-term exponential approximation techniques, it is difficult to accurately capture the tail probability for these cases [13]. The slow convergence of the tail probability to its asymptote is often observed when the source is correlated at multiple time scales. Multiple time-scale correlation in general occurs when heterogeneous sources are multiplexed. Also certain traffic sources (for example, MPEG and JPEG encoded video) are themselves correlated at different time scales. Since high-speed networks are expected to support many different types of traffic, each of which has its own correlation pattern, the network traffic is very likely to be correlated at multiple time scales. Therefore, it is important to be able to analyze the queue behavior for such traffic. In Section IV, we will introduce our second asymptotic upper bound based on the maximum variance $\langle \sigma_x^2 \rangle$ which will be useful even when the traffic is correlated at different time scales.

Example 3: In this example, we show that the asymptotic constant and the statistical multiplexing gain could be very different even for stationary Gaussian input processes having the same autocovariance sum S. Consider two autocovariance functions, $C_1(l) = 25.641 \times 0.95^{|l|}$ and $C_2(l) = 2.5063 \times 0.995^{|l|}$, both of which sum up to $S = 1000$ and satisfy conditions (C1)–(C3). Although these functions have the same values of S, as one can see from their decay rate (as $l \to \infty$), $C_2(l)$ is spread over a wider range of l than $C_1(l)$. Therefore, $C_2(l)$ has a significantly larger value of D than $C_1(l)$ [19 487.16 for $C_1(l)$ versus 199 501.48 for $C_2(l)$]. Hence, as we discussed in the previous section, the asymptotic constant (for the same value of κ) for the Gaussian input process with autocovariance $C_2(l)$ is expected to be smaller than that for the Gaussian input process with autocovariance $C_1(l)$. In Fig. 5, we show the exact tail probability and the asymptotic upper bound for two Gaussian input processes with autocovariance $C_1(l)$ and $C_2(l)$ when $\kappa = 5$. The asymptotic constant is accurately estimated by its upper bound as in the previous examples, and the asymptotic constant for the autocovariance function $C_2(l)$ is smaller than that for $C_1(l)$ (by almost 4 orders of magnitude!). Further, note that the statistical multiplexing gain as a function of M (the system scale, i.e., the number of sources, when the capacity is also proportionally increased) increases as fast as $\exp[(2\kappa^2 D/S^2)M]$. Therefore, as the system scale increases, the (logarithmic) difference between the asymptotic constants for these two Gaussian input processes will also increase very fast.

The above example can also be related to the effect of smoothing in the following way. The Gaussian process with autocovariance $C_2(l)$ can be thought of as the output of a linear smoothing system discussed in the previous section fed by the Gaussian process with autocovariance $C_1(l)$ for appropriately chosen coefficients $a_m (m = 0, 1, \cdots)$. Therefore, this example illustrates that smoothing certain types of network traffic which are correlated over a relatively short time scale, can significantly reduce network congestion. On the other hand, for some traffic types, such as JPEG-encoded video traffic, which are intrinsically correlated over very long time scales, smoothing over a small number of time frames will only marginally change the value of D and hence will not effectively reduce network congestion. For the case of real video traffic this type of effect has already been observed (e.g., [30]).

IV. Maximum Variance Asymptotic Upper Bound

We begin this section by studying the importance of $\hat{n}_x$, the time scale at which $\sigma_{x,n}^2$ attains its maximum, and a well known lower bound which motivates the development of our second asymptotic upper bound.

A. Dominant Time-Scale $\hat{n}_x$ and a Known Lower Bound

For a general (including non-Gaussian) stationary ergodic net input process γ_n, it can be shown that $\mathbb{P}(X_n > x) \to 0$, as $n \to \infty$. Therefore, there must exist a finite value of $n = \hat{n}_x$ at which the function $\mathbb{P}(X_n > x)$ attains its maximum. From (4) we get the following trivial lower bound:

$$\mathbb{P}(Q > x) \geq \sup_{n \geq 0} \mathbb{P}(X_n > x) = \mathbb{P}(X_{\hat{n}_x} > x). \tag{11}$$

At first glance, it appears that this simple lower bound is probably loose, since it is the probability that X_n is greater than x at only one point $n = \hat{n}_x$ in the index set $\{0, 1, 2, \cdots\}$ made of infinite elements. However, in many studies, it has been found that $\mathbb{P}(Q > x) = \mathbb{P}(\langle X \rangle > x)$ is largely dominated by $\mathbb{P}(X_{\hat{n}_x} > x)$, the probability that X_n exceeds x where it is most likely to happen (i.e., at $\hat{n}_x$). For instance, the lim inf-part of many asymptotic results have been derived using this lower bound (e.g., see [6], [15], [19], [25]). Further, in many cases, this lower bound has been found to be log-similar ($\overset{\log}{\sim}$) to the tail probability as x (or M) goes to infinity. From (5), remember that for Gaussian processes, $\hat{n}_x$, the dominant time scale is also the time at which $\sigma_{x,n}^2$ attains its maximum value $\langle \sigma_x^2 \rangle$. We will now introduce an asymptotic

result for Gaussian input processes that has been used in the derivation of Theorem 1, and further illustrates the importance of $\hat{n}_x$ in studying the asymptotic behavior of $\mathbb{P}(Q > x)$.

Theorem 2: Under condition (C1), for any $\alpha > 1$,[3]

$$\lim_{x\to\infty} \frac{\mathbb{P}(\langle X\rangle_{[x/\alpha\kappa,\, \alpha x/\kappa]} > x)}{\mathbb{P}(\langle X\rangle > x)} = 1.$$

Proof: Refer to Theorem 3.1 in [11]. ■

From Proposition 2, note that for arbitrary $\alpha > 1$, the interval $[\hat{n}_x/\alpha,\, \alpha\hat{n}_x]$ (and hence $\hat{n}_x$ itself) will eventually be contained in $[x/2\alpha\kappa,\, 2\alpha x/\kappa]$ as x increases. Therefore, Theorem 2 implies that for any $\alpha > 1$,

$$\lim_{x\to\infty} \mathbb{P}(\langle X\rangle_{[\hat{n}_x/\alpha,\, \alpha\hat{n}_x]} > x | \langle X\rangle > x) = 1. \tag{12}$$

In other words, as x increases, $\mathbb{P}(Q > x) = \mathbb{P}(\langle X\rangle > x)$ is essentially determined on a relatively small interval around the maximum variance time $\hat{n}_x$. Also, (12) can be interpreted as a theoretical verification of the qualitative statement "rare events take place only in the most probable way" [15], [26].

Observe that $\mathbb{P}(\langle X\rangle_{[\hat{n}_x/\alpha,\, \alpha\hat{n}_x]} > x)$ with $\alpha = 1$ corresponds to the lower bound, (for Gaussian input processes). This lower bound can be written in terms of $\Psi(z) := 1/\sqrt{2\pi}\int_z^\infty \exp[-(y^2/2)]\,dy$ (the tail function of the standard Gaussian distribution) as

$$\mathbb{P}(Q > x) \geq \Psi\left(\sqrt{\frac{x}{\langle\sigma_x^2\rangle}}\right). \tag{13}$$

Note that the lower bound is virtually equivalent to the approximation for the tail probability suggested in [25], [27] (the approximation in [25], [27] corresponds to the middle term in (16) which is almost the same as the lower bound). Since (12) holds for any arbitrary α greater than 1, it suggests that even if the lower bound $\Psi(\sqrt{x/\langle\sigma_x^2\rangle})$ were to asymptotically diverge from the exact tail probability, it would do so very slowly. In fact, through extensive numerical studies [9], [11], we have found that our lower bound accurately captures $\mathbb{P}(Q > x)$ even for small values of x. For illustration, in Fig. 6, we consider the same multiple time-scale source of Example 2. Unlike the earlier asymptotic upper bound, the lower bound closely tracks the tail probability over the entire range of queue lengths shown. This is a very important feature of the lower bound which no single exponential approximation can possess (as was illustrated in Example 2). On the other hand, since for a very large class of Gaussian input processes, the tail probability is asymptotically exponential, *our asymptotic upper bound is asymptotically tight in the sense that the (logarithmic) difference between the exact tail probability and the bound is bounded.* In contrast, as we will show later, the lower bound does in fact asymptotically diverge from the exact tail (albeit very slowly). Hence, in the next section, we will provide another asymptotic upper bound that has the nice properties of both the lower bound, and the single-exponential asymptotic upper bound.

[3] Recently, this theorem has been generalized and significantly strengthened [10]. However, since the improved version has been derived (as yet) for only continuous-time Gaussian processes, we do not provide it here. Moreover, the theorem in its current form has been used to derive all of the main results in this paper [11].

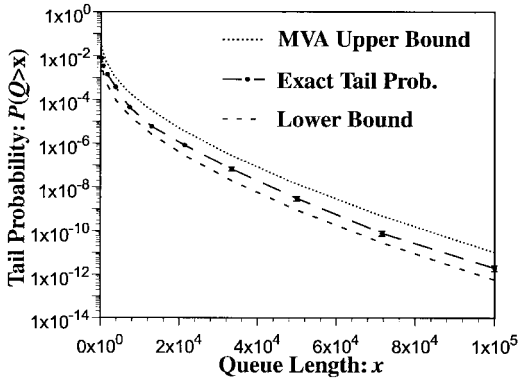

Fig. 6. The exact tail probability, the lower bound, and the MVA upper bound for a Gaussian input process with autocovariance function $C_\lambda(l) = 104 \times 0.99^{|l|} + 64.14 \times 0.999^{|l|} + 31.86 \times 0.9999^{|l|}$ when $\kappa = 33.33$.

B. Maximum Variance Asymptotic Upper Bound

In this section, we will introduce an asymptotic upper bound that, like the lower bound, will be based on the maximum variance of a Gaussian process. Recall that the lower bound is a simple (standard Gaussian tail distribution) function of $\sqrt{x/\langle\sigma_x^2\rangle}$. From Theorem 2, and the fact that the lower bound matches the shape of the tail probability curve, we can infer that the term $x/\langle\sigma_x^2\rangle$, as a function of x, contains key information about the behavior of the tail probability before it closely converges to its asymptote. Our idea is to find a function $q(z)$ which resembles $\Psi(z)$ such that $q(\sqrt{x/\langle\sigma_x^2\rangle})$ is similar ($\sim$) to the asymptotic upper bound $\exp[-(2\kappa/S)\,(x + \kappa D/S)]$. In this way, $q(\sqrt{x/\langle\sigma_x^2\rangle})$ would asymptotically bound the exact tail probability from above, and also closely track the shape of the tail probability curve. In the following theorem, which is based on Theorem 1, we find such an asymptotic upper bound.

Theorem 3: Under conditions (C1) and (C2)

$$\exp[-(x/2\langle\sigma_x^2\rangle)] \sim \exp\{-[(2\kappa/S)(x + \kappa D/S)]\}.$$

Therefore, with an additional condition (C3), $\exp[-(x/2\langle\sigma_x^2\rangle)]$ asymptotically bounds $\mathbb{P}(Q > x)$.

Proof: Refer to [11, Proposition 4.1].

We call this new bound the *maximum variance asymptotic* (MVA) upper bound. Note that the MVA upper bound, as a function of $z = \sqrt{x/\langle\sigma_x^2\rangle}$, can be written as $q(z) = \exp[-(z^2/2)]$. Further, from a well-known bound for $\Psi(z)$ [17], i.e.,

$$\frac{1 - z^{-2}}{\sqrt{2\pi}}\, z^{-1} \exp\left(-\frac{z^2}{2}\right) \leq \Psi(z) \frac{1}{\sqrt{2\pi}}\, z^{-1} \exp\left(-\frac{z^2}{2}\right) \quad \forall z > 0 \tag{14}$$

we have

$$\Psi(z) \sim \frac{\exp\left(-\frac{z^2}{2}\right)}{\sqrt{2\pi}\, z} = \frac{q(z)}{\sqrt{2\pi}\, z}. \tag{15}$$

Note [from (14)] that the above similarity comes into effect very fast as z increases, and $\Psi(z) \approx \exp[-(z^2/2)]/\sqrt{2\pi}\, z$

even for fairly small values (>2) of z. Therefore, the major difference between $\Psi(z)$ and $\exp[-(z^2/2)]$ is the multiplicative term $1/\sqrt{2\pi}\,z$ on the right-hand side of (15). This term is very slowly decreasing (as z increases) compared to the remaining part $\exp[-(z^2/2)]$. Therefore, the shape of the MVA upper bound curve should almost be the same as that of the lower bound. Also, in a sense this MVA upper bound is obtained by "lifting" the lower bound in such a way that it becomes a tight asymptotic upper bound. Hence, unlike the asymptotic upper bound in Section III, we expect that the MVA upper bound will bound the tail probability even for very small values of queue lengths as if it were a global upper bound. This prediction has been verified through simulations [11]. In addition to the asymptotic tightness of the MVA upper bound, this is another property of the MVA upper bound which makes it more useful than the lower bound (since conservative, rather than optimistic, engineering is often desirable for network dimensioning and control).

A direct result of Theorem 3 is that under conditions (C1)–(C3):

$$\Psi\left(\sqrt{\frac{x}{\langle\sigma_x^2\rangle}}\right) \sim \sqrt{\frac{\langle\sigma_x^2\rangle}{2\pi x}}\,\exp\left(-\frac{x}{2\langle\sigma_x^2\rangle}\right) \sim \sqrt{\frac{S}{8\pi\kappa x}}\cdot\exp\left\{-\left[\frac{2\kappa}{S}\left(x+\frac{\kappa D}{S}\right)\right]\right\}. \quad (16)$$

Note that the second similarity is from Proposition 3 and Theorem 3. From (16), it is now clear that the lower bound is not asymptotically exponential, and hence cannot be similar to the exact tail probability. However, the leading term $\sqrt{S/8\pi\kappa x}$ decreases slowly compared to the remaining term $\exp[-(2\kappa/S)\,(x+\kappa D/S)]$, as $x\to\infty$. For this reason, the divergence of the lower bound from the tail probability was nearly unrecognizable in all our numerical studies [9], [11]. Perhaps the following observation will shed further light on this issue.

The (logarithmic) difference

$$-(x/2\langle\sigma_x^2\rangle) - \log\,\Psi(\sqrt{x/\langle\sigma_x^2\rangle})$$

between the MVA upper bound and the lower bound is actually a function of $\sqrt{x/\langle\sigma_x^2\rangle}$, that can be closely approximated by $\frac{1}{2}\,(\log 2\pi x/\langle\sigma_x^2\rangle)$. Therefore, the difference between these bounds cannot be arbitrary but can be determined from either the MVA upper bound or the lower bound, as illustrated in Fig. 7. In the figure, the difference between the two bounds is only about an order of magnitude even when the MVA upper bound is as small as 10^{-20}. Therefore, Fig. 6 suggests that the MVA upper bound and lower bound may provide a narrow envelope that bounds the exact tail probability in the typical range of interest. This is also suggested in Fig. 6 earlier, where we plot the lower bound and the MVA upper bound for a Gaussian input process correlated at multiple time scales. Note that the lower bound and the MVA upper bound encapsulate the tail probability over the entire range of queue lengths. Since both bounds are based on the maximum variance, neither suffers from the slow convergence of the tail probability to its asymptote. Similar experimental studies

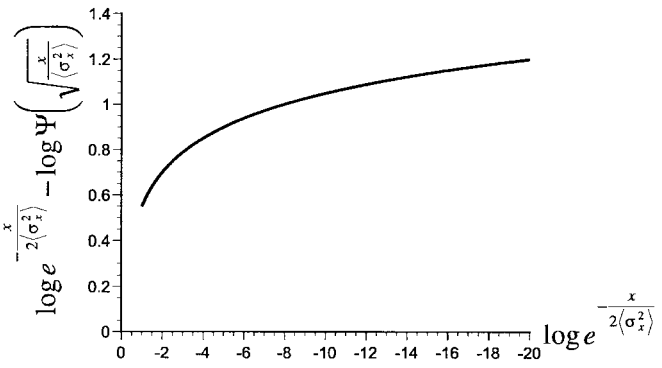

Fig. 7. The difference $\log\,\exp[-(x/2\langle\sigma_x^2\rangle)] - \log\,\Psi(\sqrt{x/\langle\sigma_x^2\rangle})$ versus the MVA upper bound $\log\,\exp[-(x/2\langle\sigma_x^2\rangle)]$.

have demonstrated that: 1) the tail probability almost never escapes from the envelope constructed by the bounds, as long as conditions (C1)–(C3) are satisfied and 2) that both the lower bound and the asymptotic upper bound can approximate tail probabilities as small as 10^{-20} with errors less than or close to an order of magnitude.

As a final remark of the section, it is interesting to note that the approximation for $\mathbb{P}(Q>x)$ based on the large deviation M-asymptotics result by Botvich and Duffield [6], results in the same expression as the MVA upper bound, when applied to Gaussian fluid queues. Remember that the M-asymptotics result in [25] improved upon the result in [6] (from log-similarity to nearly similarity), and an approximation based on these stronger asymptotics was suggested (which is equivalent to the lower bound). This tells us that the approximation that satisfies only the weaker asymptotics in M-asymptotics [6], now satisfies the stronger asymptotics in x-asymptotics (and vice versa). As mentioned in Section I, this is because x-asymptotics and M-asymptotics consider asymptotic properties of $\mathbb{P}(Q>x)$ in different limiting regimes.

V. Applications for General Input Processes

The numerical examples provided in Sections III and IV were for stationary Gaussian input processes. Further, both the asymptotic upper bounds described in the previous sections are valid under three conditions (C1)–(C3). In this section, we investigate and discuss the accuracy of the lower bound and the MVA upper bound as an approximation for the tail probability when conditions (C1)–(C3) are violated, and also when the aggregate input process is itself not Gaussian.

A. General Gaussian Process

The relation (11) is very generally true, and the lower bound $\Psi(\sqrt{x/\langle\sigma_x^2\rangle})$ is valid as long as the input process is stationary Gaussian. On the other hand, both the asymptotic upper bounds in Sections III and IV, require conditions (C1)–(C3).

As mentioned in Section II, when condition (C1) is violated, the input process shows long-range dependence, and the corresponding tail probability may not even be asymptotically exponential [15]. However, as long as the input process is stationary and ergodic, the (finite) maximum variance $\langle\sigma_x^2\rangle$ can be found and used to compute the lower bound and

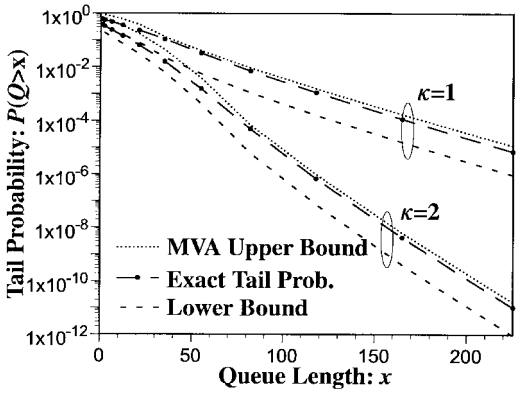

Fig. 8. The exact tail probability, the lower bound, and the MVA upper bound for a Gaussian input process with autocovariance function $C_\lambda(l) = 10 \times 0.9^{|l|} \cos \pi l/12 + 0.1 \times 0.99^{|l|}$ and $\kappa = 1, 2$.

the MVA upper bound. In fact, in [26],[4] an approximation for the tail probability, equivalent to the MVA upper bound, has been used for the special case of Fractal Brownian motion, and empirically found to be fairly accurate. Our own numerical investigations with long-range dependent sources [which violate both conditions (C1) and (C2)] have resulted in the same conclusion. Further, in more recent work using extreme value theory, we have shown (a significantly stronger result than the Large Deviation results) that for a very large class of long-range dependent (and other) Gaussian processes, the MVA upper bound diverges very slowly (or not at all) from the exact tail [10]. However, in this paper we will not explicitly focus on numerically studying long-range dependent processes, but instead will provide examples using actual traces of video traffic (which is often considered to exhibit self-similar behavior).

Even though any nonnegative autocovariance function satisfies condition (C3), it should be noted that some types of network applications (such as MPEG video) generate network traffic in a fairly periodic fashion. This may result in a large enough negative component of the autocovariance function to violate condition (C3). Thus, in the following example, we investigate the performance of the lower bound and the MVA upper bound for input processes that do not satisfy condition (C3).

Example 4: In Fig. 8, we show the exact tail probability, the lower bound, and the MVA upper bound for a Gaussian input process whose autocovariance function is given by $C_\lambda(l) = 10 \times 0.9^{|l|} \cos(\pi l/12) + 0.1 \times 0.99^{|l|}$. One can easily check that this autocovariance function does not satisfy condition (C3). Hence, the MVA upper bound in this example may not be an asymptotic upper bound. However, note that both the lower bound and the MVA upper bound still accurately match the tail probability curve. In particular, note how both these approximations are able to track even minor transitions of the exact tail curve from concavity to convexity. This again emphasizes the importance of the maximum variance $\langle \sigma_x^2 \rangle$.

[4] In this paper, the tail probability was approximated by the lower bound given in (13), but the lower bound itself was evaluated through another approximation $\Psi(z) \approx \exp[-(z^2/2)]$. As a consequence, the resultant estimate of $\mathbb{P}(Q > x)$ actually corresponds to our MVA upper bound.

Further, the MVA upper bound seems to be asymptotically close to the tail probability. This suggests that the bound $\exp[-(2\kappa^2 D/S^2)]$ to the asymptotic constant C in (2) may be used to accurately approximate it even when (C3) is violated, or when D has a negative value. This may be true in part because the expression $\exp[-(2\kappa^2 D/S^2)]$ has important properties that the asymptotic constant is known to have, such as: 1) if the input process is i.i.d. Gaussian, then $D = 0$ and the asymptotic upper bound simply becomes $\exp[-(2\kappa x/S)]$ which is a well-know bound for the level crossing probability of a random walk with drift (see [29, p. 236]) and 2) also, D can have a negative value, only when the autocovariance function of the input process takes large negative values (i.e., when the input process is significantly periodic and less bursty than i.i.d. input processes). If D takes on a negative value, then $\exp[-(2\kappa^2 D/S^2)]$ is greater than 1, and will increase exponentially with the size of the system (as explained in Section III). This indicates that for strongly periodic input processes, there will be no gain in statistical multiplexing the traffic; an observation which is well known for certain types of periodic input traffic [13], [31].

In the following section, we altogether weaken the Gaussian assumption on the input process, and use the lower and the MVA upper bounds to approximate the tail probability of fluid queues with a large number of non-Gaussian input processes.

B. Applications to Voice and Video Traffic

As mentioned in Section I, the huge capacity of high-speed network links motivates the Gaussian characterization of the aggregate traffic to a multiplexer. For example, FORE SYSTEMS has already built commercial ATM switches to support *OC-12* (622.08 Mb/s) lines, and ATM networks with *OC-24* (1.2 Gb/s) lines are already operational (at Cambridge University). Due to the huge capacity of a single ATM link, hundreds or even thousands of network applications are expected to share an ATM link; an *OC-3* (155.52 Mb/s) line can accommodate over 6800 voice calls (assuming 16-Kb/s mean bit-rate) and an *OC-12* line over 300 MPEG video calls (assuming 1.5-Mb/s mean bit-rate) both at a utilization of $\rho := \mathbb{E}\{\lambda_0\}/\mu = 0.8$. These numbers seem to be large enough for the central limit theorem to be applied, and to characterize the aggregate input process by a Gaussian process. Through empirical evidence we have found that a few hundred sources are generally sufficient for the Gaussian approximation to be quite good (e.g., see [9]).

In this section, we illustrate the effectiveness of the Gaussian characterization and the applicability of the lower and the MVA upper bounds for general traffic models. Our examples focus on voice and video traffic models. It should be emphasized that since we have weakened the Gaussian assumption, both the lower and MVA upper bounds cannot strictly be thought of as bounds, but are approximations, even if the various conditions on the autocovariance function of the aggregate input process were satisfied. However, as will be illustrated by the numerical examples, as long as the Gaussian model is reasonably good, these analytical approximations do behave like real bounds over the tail probabilities of interest.

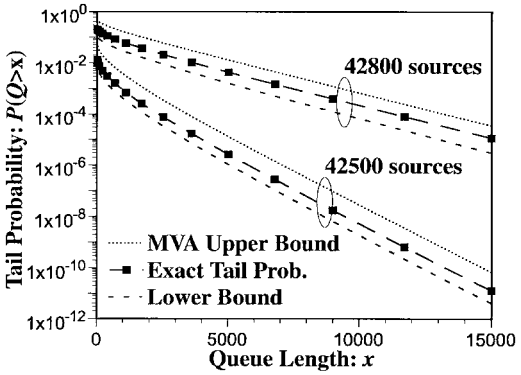

Fig. 9. The exact tail probability, the lower bound and the MVA upper bound for a multiplexer serving 42 500 and 42 800 voice traffic sources. The output link capacity is set to 622.08 Mb/s (*OC-12* line).

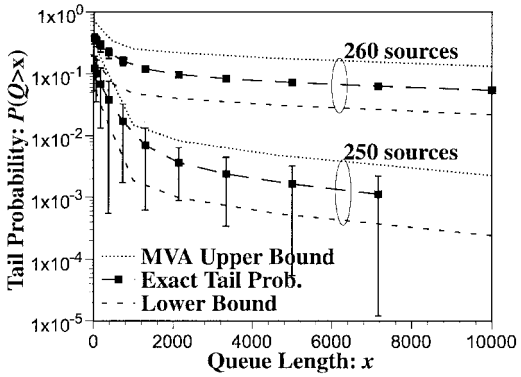

Fig. 10. The exact tail probability, the lower bound and the MVA upper bound for a multiplexer serving 250 and 260 real MPEG sources. The output link capacity is set to 155.52 Mb/s (*OC-3* line).

In the next few examples, we demonstrate the utility of the MVA upper bound and lower bound in analyzing the tail probability at a multiplexer for different cases. In each case, the sources are fed into a multiplexer being served by an *OC-3* (155.52 Mb/s) or *OC-12* (622 Mb/s) line. To save space, we refer to our technical report [11] for the detailed specifications of the traffic source models that we use in this section.

1) Voice Traffic Sources:

Example 5: The typical behavior of efficiently encoded voice traffic is that it alternates between "active" and "inactive" states. Hence, Markov modulated on–off processes have frequently been used to model voice traffic (e.g., see [34]). For our experiment, we assume a 10 ms slot size and use a discrete-time on-off MMF process as a voice traffic source model obtained by discretizing the continuous-time MMF voice traffic source model used in [31]. In Fig. 9, we show the exact tail, the lower bound and the MVA upper bound for 42 500 and 42 800 voice sources served by an *OC-12* (622.08 Mb/s) line. As one can see in the figure, the simulation results are accurately captured between the lower bound and the MVA upper bound.

2) Video Traffic Sources: In general, the stochastic characteristics of a video traffic source changes with the type of video application which the source represents. For instance, a video traffic source that mainly transmits movies is likely to have different characteristics from that of a video source that transmits news programs. Further, the video coding schemes employed to reduce the required bandwidth can also significantly affect the stochastic characteristics of the generated video traffic. Therefore, the detailed modeling of such diverse video traffic sources may neither be an easy nor an efficient way of characterizing these sources. From this viewpoint, traffic characterization based only on the first two moments (mean and autocovariance or mean and IDC) has advantages over the characterization based on explicit stochastic modeling, since they can be directly measured from the source. In the previous example involving a non-Gaussian voice traffic source model, the first two moments of the traffic sources have been analytically obtained from the source model. In the next example, we will show that from the measured mean and autocovariance of a real video trace, the queue length distribution can also be accurately computed.

3) Example 6: In this example, we use real MPEG video (frame-size) traces generated by Rose [28]. To simulate MPEG-encoded video traffic, 16 different MPEG coded traces of 40 000 frames are concatenated into one trace of 640 000 frames, and the frame sizes are read out sequentially from this trace starting at a random position in the trace. Since all the concatenated frame-size traces are from video sequences captured at 25 frames/s, the total length (640 000 frames) of the concatenated frame-size trace corresponds to more than 7 h of play time. Since the trace is very long, by simply assigning a random starting position to each simulated MPEG video traffic source, we generate a large number of MPEG video traffic sources. Since we assume a 10-ms slot size in this example, each frame size should be read out over 4 slots. We assume that each frame is transmitted uniformly over a frame period (40 ms or equivalently four slots). In Fig. 10, the lower bound and the MVA upper bound for 250 and 260 MPEG video sources served at 3667 cells/slot (*OC-3* line) are compared to the exact tail probabilities. The mean and autocovariance function of the simulated MPEG source are measured directly from the concatenated frame-size trace, and used for our approximation technique. Since we are now using real frame-size traces to simulate MPEG encoded video sources, the importance sampling technique cannot be used for this experiment and, hence, the simulation results show larger confidence intervals. Nevertheless, as one can see in the figure, both the lower bound and the MVA upper bound again seem to encapsulate the exact tail probability within an order of magnitude.

4) Example 7: In this example, we use a frame-size trace of the JPEG-encoded movie "Star Wars" to simulate real video sources. Also, we design a simple JPEG video traffic source model based on the mean and autocovariance function measured directly from the frame-size trace. We then use the model to obtain our bounds and another set of simulation results. Many types of video traffic have been found to be heavily correlated over multiple time scales or even thought to exhibit self-similar behavior over a certain time-period (e.g., see [5]). To capture this multiple time-scale correlation of video traffic, we model the JPEG video traffic source as the superposition of 3 two-state MMF processes with very different mean state sojourn times. More precisely, this

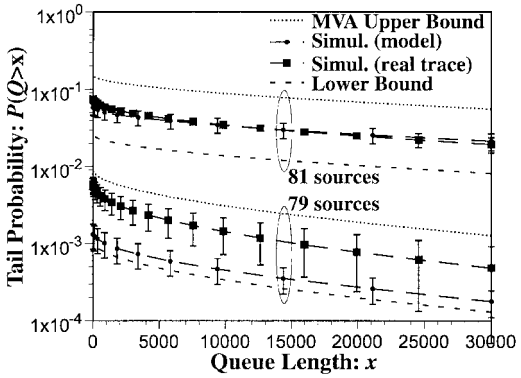

Fig. 11. Simulation results, the lower bound, and the MVA upper bound for a multiplexer serving 79 and 81 JPEG-encoded movie "Star Wars" through an *OC-12* output link.

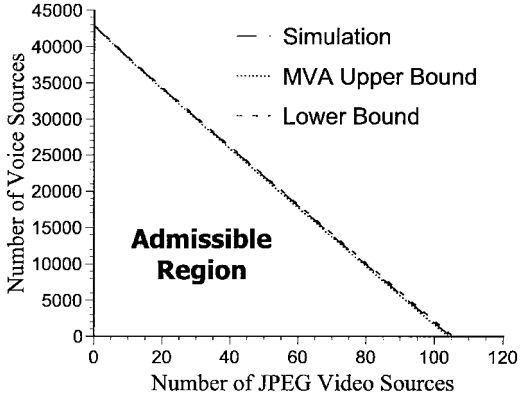

Fig. 12. Admissible combinations of voice and JPEG-encoded video calls for an *OC-12* link with 20 000 cell buffers, computed by simulation, the lower bound, and the MVA upper bound. The maximum tolerable tail probability (φ) is set to 10^{-6}.

source model is obtained by matching the autocovariance function measured from the frame-size trace using the *least-square* method. The main purpose of designing a model for JPEG traffic is to demonstrate that the queuing behavior of a traffic source can be captured by a relatively simple stochastic model of the traffic source, especially when the number of multiplexed traffic sources is large. In Fig. 11, we show simulation results, the lower bound, and the MVA upper bound for a multiplexer serving 79 and 81 JPEG traffic sources through an *OC-12* line. The time slot size is set to 8.333 ms. Since the frame-size trace is from video sequences captured at 30 frames/s, each frame-size is read out over four slots. As in the previous example, we assume that a frame is uniformly transmitted over four slots. As one can see in the figure, the two simulation results (one using the real frame-size trace and the other using the model) are encompassed within the lower and MVA upper bounds.

5) Admission Control—Voice and Video: An important application of our analytical results is for admission control. We assume that a new call is admitted to an ATM multiplexer with buffer size B if the resulting tail probability $\mathbb{P}(Q > x = B)$ is less than some φ. Hence, φ corresponds to the maximum tolerable tail probability for a call to be admitted.

6) Example 8: In Fig. 12, we show the admissible region for voice and JPEG-encoded video calls computed by simulation, and via our maximum variance based bounds. The maximum tolerable tail probability φ and the buffer size B are set to 10^{-6} and 20 000 cells, respectively. Again, we assume that an *OC-12* line serves the multiplexer. Since the required constraint φ is quite small, we use simple stochastic models for both voice and JPEG video traffic sources in order to employ the importance sampling technique. While we use the same traffic source model that is used in Example 5, we use a JPEG video traffic model that is somewhat different from the model used in Example 7 (in order to simulate smaller tail probabilities than given in Fig. 11). It is interesting to note that in Fig. 12, the admissible region computed by simulation, the lower bound, and the MVA upper bound are so close that it is almost difficult to distinguish their boundaries. In fact, the lower bound overestimates and the MVA upper bound underestimates the maximum admissible number of calls by less than 1% in terms of utilization.

VI. Conclusion

In this paper, we provide two asymptotic upper bounds to analyze the tail of the steady-state distribution $\mathbb{P}(Q > x)$ at a high-speed multiplexer. We model the multiplexer as an infinite buffer fluid queue and characterize the aggregate input process as a Gaussian stochastic process. This enables us to avoid the classical state explosion problem that occurs when many traffic sources are multiplexed.

For a Gaussian input process satisfying fairly general conditions, we provide an exponential asymptotic upper bound (Theorem 1) $\exp[-(2\kappa/S)(x + \kappa D/S)]$ to the tail probability $\mathbb{P}(Q > x)$ using key results in extreme value theory. This asymptotic upper bound in turn results in a theoretical contribution to the extreme value literature. The asymptotic upper bound also results in an upper bound to the asymptotic constant.

We develop another result (Theorem 2) which emphasizes the importance of the maximum variance $\langle\sigma_x^2\rangle$, and provides theoretical grounding for a well-known lower bound. Building upon our exponential asymptotic upper bound and Theorem 2, we also develop an asymptotic (MVA) upper bound $\exp[-(x/2\langle\sigma_x^2\rangle)]$ (Theorem 3), based on the maximum variance $\langle\sigma_x^2\rangle$. Through an extensive and systematic numerical study, we find that both the lower bound and the MVA upper bound accurately approximate the tail probability as long as the input process can be effectively characterized by a Gaussian process. We also illustrate that our analysis of the tail probabilities results in very efficient admission control.

In this paper, we have provided results only for the discrete-time fluid queues in which the fluid arrival and service take place only at discrete times. Equivalent results for the continuous-time fluid queue have already been derived and are available in [12]. We find that Gaussian modeling of the input traffic provides significant simplicity and has great potential, and are currently investigating ways to extend the analysis to a network end-to-end.

Acknowledgment

The authors would like to thank the anonymous reviewers for their careful reading of this paper and for bringing to the

authors' attention various important references in the large deviations literature that the authors were previously unaware of. This paper has been improved due to their input.

REFERENCES

[1] J. Abate, G. L. Choudhury, and W. Whitt, "Exponential approximations for tail probabilities in queues—I: Waiting times," *Oper. Res.*, vol. 43, no. 5, pp. 885–901, 1995.

[2] 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.

[3] R. G. Addie, M. Zukerman, and T. Neame, "Fractal traffic: Measurements, modeling, and performance evaluation," in *Proc. IEEE INFOCOM*, 1995, pp. 977–984.

[4] R. J. Adler, *An Introduction to Continuity, Extrema, and Related Topics for General Gaussian Processes.* Hayward, CA: Inst. Math. Statist., 1990.

[5] J. Beran, R. Sherman, M. S. Taqqu, and W. Willinger, "Long-range dependence in variable-bit-rate video traffic," *IEEE Trans. Commun.*, vol. 43, pp. 1566–1579, Feb.–Apr. 1995.

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

[7] 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.

[8] C.-S. Chang and J. A. Thomas, "Effective bandwidth in high-speed digital networks," *IEEE J. Select. Areas Commun.*, vol. 13, pp. 1091–1114, Aug. 1995.

[9] J. Choe and N. B. Shroff, "A new method to determine the queue length distribution at an ATM multiplexer," in *Proc. IEEE INFOCOM*, 1997, pp. 550–557.

[10] ———, "Supremum distribution of Gaussian processes and queueing analysis including long-range dependence and self-similarity," *Stochastic Models*, Purdue Univ., West Lafayette, IN, Tech. Rep., 1997, submitted for publication.

[11] ———, "A central limit theorem based approach for analyzing queue behavior in high-speed networks," Purdue Univ., West Lafayette, IN, Tech. Rep., 1998.

[12] ———, "On the supremum distribution of integrated stationary Gaussian processes with negative linear drift," *Adv. Appl. Prob.*, to be published.

[13] 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.

[14] L. Donatiello and R. Nelson, *Models and Techniques for Performance Evaluation of Computer and Communication Systems.* New York: Springer-Verlag, 1993.

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

[16] A. Elwalid, D. Heyman, T. 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.

[17] W. Feller, *An Introduction to Probability Theory and its Applications I.* New York: Wiley, 1968.

[18] K. W. Fendick and W. Whitt, "Measurements and approximations to describe the offered traffic and predict the average workload in a single-server queue," *Proc. IEEE*, vol. 77, pp. 171–194, Jan. 1989.

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

[20] R. Guerin, H. Ahmadi, and M. Naghshineh, "Equivalent capacity and its application to bandwidth allocation in high-speed networks," *IEEE J. Select. Areas Commun.*, vol. 9, pp. 968–981, Sept. 1991.

[21] P. Heidelberger, "Fast simulation of rare events in queueing and reliability models," *ACM Trans. Modeling and Computer Simulation*, vol. 5, pp. 43–85, Jan. 1995.

[22] L. Kosten, "Liquid models for a type of information buffer problems," *DELFT Progress Rep.*, vol. 11, pp. 71–86, July 1986.

[23] W. E. Leland, M. 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.

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

[25] M. Montgomery and G. De Veciana, "On the relevance of time scales in performance oriented traffic characterization," in *Proc. IEEE INFOCOM*, 1996, pp. 513–520.

[26] 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.

[27] J. Roberts, U. Mocci, and J. Virtamo, *Broadband Network Teletraffic, Final Report of Action COST 242.* New York: Springer-Verlag, 1996.

[28] O. Rose, "Statistical properties of MPEG video traffic and their impact on traffic modeling in ATM systems," in *Proc. the 20th Conference on Local Computer Networks*, Minneapolis, MN, Oct. 1995, pp. 397–406.

[29] S. M. Ross, *Stochastic Processes.* New York: Wiley, 1983.

[30] N. B. Shroff, "Traffic modeling and analysis in high speed ATM networks," Ph.D. dissertation, Graduate School of Arts and Sciences, Columbia Univ., New York, NY, 1995.

[31] N. B. Shroff and M. Schwartz, "Improved loss calculations at an ATM multiplexer," *IEEE/ACM Trans. Networking*, vol. 6, pp. 411–421, Aug. 1998.

[32] A. Simonian, "Stationary analysis of a fluid queue with input rate varying as an Ornstein–Uhlenbeck process," *SIAM J. Appl. Math.*, vol. 51, pp. 828–842, 1991.

[33] ———, "Transient and stationary distributions for fluid queues and input processes with a density," *SIAM J. Appl. Math.*, vol. 51, pp. 1732–1739, 1991.

[34] 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.

Jinwoo Choe (S'96) received the B.S. and M.S. degrees from Seoul National University in Seoul, Korea in 1990 and 1992, respectively. He is currently pursuing a doctoral degree in the School of Electrical and Computer Engineering at Purdue University. His research areas of interest encompass network analysis, traffic modeling, and rare event simulations.

Ness B. Shroff (S'90–M'94), for photograph and biography, see p. 421 of the August 1998 issue of this TRANSACTIONS.