9-1-1997

New Bounds and Approximations using Extreme Value Theory for the Queue Length Distribution in High-Speed Networks

Jinwoo Choe
Purdue University School of Electrical and Computer Engineering

Ness B. Shroff
Purdue University School of Electrical and Computer Engineering

New Bounds and Approximations Using Extreme Value Theory for the Queue Length Distribution in High-Speed Networks

Jinwoo Choe
Ness B. Shroff

TR-ECE 97-9
September 1997

School of Electrical
and Computer Engineering
Purdue University
West Lafayette, Indiana 47907-1285

New Bounds and Approximations using Extreme Value Theory for the Queue Length Distribution in High-Speed Networks*

Jinwoo Choe Ness B. Shroff

School of Electrical and Computer Engineering
1285 Electrical Engineering Building
Purdue University
West Lafayette, IN 47907-1285

*This research was supported in part by NSF CAREER Grant NCR-9624525, NSF Grant CDA-9422250, and the Purdue Research Foundation grant 690-1285-2479.

TABLE OF CONTENTS

LIST OF FIGURES

Abstract

Statistical multiplexing is very important in high-speed ATM type of networks, since it allows applications to efficiently share valuable network resources. However, statistical multiplexing can also lead to congestion which must be effectively controlled in order to provide users satisfactory quality of service.

In this report we study a fundamental measure of network congestion, the tail of the steady state queue length distribution at an ATM multiplexer. In particular, we focus on the case when the aggregate traffic to the multiplexer can be characterized by a Gaussian process. In our approach, an ATM multiplexer is modeled by a fluid queue serving a large number of input processes. We derive asymptotic upper bounds to $P(\{Q > x))$ the tail of the queue length distribution, and provide several numerical examples to illustrate the tightness of the bounds. Our study is based on Extreme *Value* Theory, and therefore different from the popular Markovian and Large Deviation techniques.

1. Introduction

Advances in lightwave communication technology have enable networks to support various real-time applications. Statistical multiplexing is extremely important in such high-speed networks, since it increases network efficiency by allowing a large number of applications to share network resources. However, when these resources, such as buffer space and link capacity, are shared, there also exists the possibility of excessive congestion, which could impact the quality of the underlying delay-sensitive real-time applications. A fundamental measure of congestion is $P(\{Q > x\})$, the tail of the steady state queue length distribution in an infinite buffer system. We will study $P(\{Q > x\})$ in the context of an ATM multiplexer (shown in Figure 1.1). The accurate computation of the tail probability $P(\{Q > x\})$ is important for the control and design of these high-speed networks. For example, $P(\{Q > B\})$ is often used to approximate the loss probability in the corresponding finite buffer system with buffer size B.

Commercial ATM switches already support 622 Mbps link speeds and gigabit-per-second switches are expected to appear soon. Therefore, most ATM multiplexers are expected to serve a large number of heterogeneous traffic sources, and the analysis of the corresponding queueing system becomes increasingly important. Computing the queue length distribution at an ATM multiplexer has been a challenging problem, and many analytical techniques have been developed. The rich theory of Markov processes has been found to be especially useful for the analysis of statistical multiplexers (e.g. [24, 27]). However, the computational complexity of these techniques increases rapidly with the number of states, and the number of states needed to model the aggregate traffic increases exponentially with the number of traffic sources being multiplexed. For this reason, in the literature, a number of approximation techniques have been suggested [1, 2, 8, 10, 12, 15, 16, 20, 21, 22, 28].

It has been shown under a variety of Cramer type assumptions (exponentially bounded marginals and autocorrelations of the arrival process) that $P(\{Q > x\})$, the tail of the queue length distribution of an infinite buffer queue, is asymptotically exponential [1, 2, 15]; that is,

$$P(\{Q > x\}) \sim Ce^{-\eta x} \qquad \text{as} \quad x \to \infty. \tag{1.1}$$

Here η is a positive constant called the asymptotic decay rate,* C is a positive constant called the asymptotic constant, and $f(x) \sim g(x)$ means that $\lim_{x\to\infty} f(x)/g(x) = 1$. Therefore, the asymptote $Ce^{-\eta x}$ has been a natural choice to approximate the tail probability for large values of x. This approximation is often called the asymptotic approximation. For a large class of queueing systems, computing the asymptotic decay rate η is quite straightforward even with a large number of arrival processes. However, the asymptotic constant C can be exactly determined only for a limited class of queueing systems. Furthermore,

*Using Large **Deviation** techniques, it has been shown for more general classes of arrival processes, that the limit $\lim_{x\to\infty} -\frac{1}{x}\log P(\{Q > x\})$ exists [19]. Obviously, (1.1) implies $\eta = \lim_{x\to\infty} -\frac{1}{x}\log P(\{Q > x\})$. Therefore, for greater generality, we define the asymptotic decay rate η as $\lim_{x\to\infty} -\frac{1}{x}\log P(\{Q > x\})$, whenever the limit exists. Note that the tail probability does not have to be asymptotically exponential for the asymptotic decay rate to be well defined.

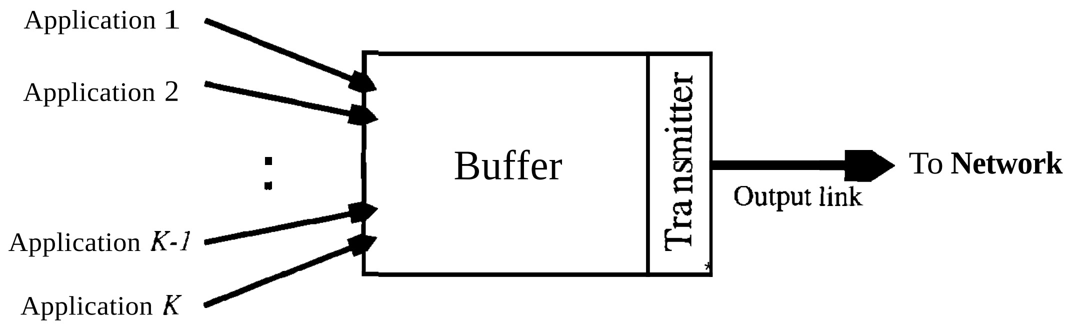

Figure 1.1: A typical ATM multiplexer serving K network applications. K applications are sharing the buffer and the network link of the ATM multiplexer.

even for these queueing systems, it is usually computationally difficult to exactly compute C when the queue serves a large number of arrival processes. Consequently, the following simpler approximation has been proposed (by setting the asymptotic constant C to 1).

$$P(\{Q > x\}) \approx e^{-\eta x}. \qquad (1.2)$$

This approximation is the well known *Effective Bandwidth* (EB) approximation, which has been suggested for use in admission control [8, 16, 20, 21, 22]. In recent papers, however, it has been found that the EB approximation does not account for statistical multiplexing gain and could thus be quite conservative [12, 28]. Therefore, there is renewed interest in the asymptotic approximation, and methods have been developed to approximate the asymptotic constant C for special cases [2, 13, 14].

Our goal is to compute the tail probability $P(\{Q > x\})$ from the first two moments of the traffic sources. In particular, we focus on the cases when the aggregate traffic to an ATM multiplexer can be characterized by a stationary Gaussian process appealing the *Central Limit Theorem* [18]. We develop an asymptotic upper bound of the form shown in (1.1) to the tail probability $P(\{Q > x\})$. Since for very general Gaussian arrival processes, it has been shown that (1.1) holds, this also implies that we find an upper bound to the asymptotic constant C. A good bound on the asymptotic constant is important, as mentioned above. However, researchers have recently contended that even the asymptotic approximation may be a poor estimate in the range of tail probabilities of interest. For example, this has been shown in [12] for On-Off arrival processes and in [9] for stationary Gaussian input processes that are correlated at multiple time scales. Hence, based on this upper bound and an earlier lower bound result [9, 10], we derive another asymptotic upper bound which is shown to track the tail probability very closely for a wide range of queue lengths x.

The remainder of this report is constructed as follows. In Chapter 2, we introduce a queueing model for ATM multiplexers and provide important definitions and facts related to the model. Also, in this chapter, we briefly discuss the Gaussian characterization of the aggregate traffic. In Chapter 3 we derive asymptotic upper bounds for the tail probability $P(\{Q > x\})$ of the queueing model fed by some classes of Gaussian input processes, and suggest an approximation for more general Gaussian processes. In Chapter 4 we first investigate the performance of our bounds through numerical examples. We also apply our approximation to the ATM multiplexer to demonstrate the practical applicability of the technique. Finally, in Chapter 5, we bring this report to a conclusion and briefly discuss future research directions.

2. Problem Modeling

2.1 Fluid Queue

Fluid queues have often been suggested as good models for the analysis of statistical multiplexers [15]. We model an ATM multiplexer by a discrete-time fluid queue shown in Figure 2.1. The fluid queue consists of an infinite buffer, a server that drains fluid from the buffer at rate μ, and K independent fluid inputs that fill the buffer at rate $\lambda_n^{(k)}$ $(k = 1, 2, \ldots, K)$. The fluid buffer and server correspond to the cell buffer and the high-speed network link of the ATM multiplexer, respectively. Conceptually, the K inputs fill the fluid buffer in much the same way as K applications load a statistical multiplexer, and the fluid server drains the buffer at a constant rate μ, in much the same way as a network link empties the buffer by transmitting cells at a fixed rate. Consequently, Q_n, the amount of fluid in the buffer at time n, is closely related to the number of cells in the multiplexer.

The evolution of the Q_n, the amount of fluid in the buffer, can be expressed by the following famous (Lindley's) equation:

$$Q_n = (Q_{n-1} + \lambda_n - \mu)^+, \tag{2.1}$$

where $\lambda_n := \sum_{k=1}^{K} \lambda_n^{(k)}$ and $(x)^+ := \max\{0, x\}$.

It has been shown under some mild assumptions (such as the stationarity and ergodicity of λ_n and the stability condition; that is, $\bar{\lambda} := E\{\lambda_n\} < \mu$), that the distribution of Q_n determined by (2.1) and an initial condition Q_0, converges to a unique limiting distribution (the steady state queue distribution) as n goes to infinity, regardless of the initial condition [23]. In addition, it has been shown that the marginal distribution of the stationary stochastic process $\tilde{Q}_n$ defined by $\tilde{Q}_n := \sup_{m \le n} \sum_{i=m+1}^{n} (\lambda_i - \mu)$, is equal to the steady state queue length distribution [23, 29]. Therefore, if we define a stochastic process ξ_n and a constant κ as

$$\xi_n := \lambda_{-n} - \bar{\lambda} \quad \text{and} \quad \kappa := \mu - \bar{\lambda}, \tag{2.2}$$

then the suprema distribution of the stochastic process X_n defined by

$$X_n := \sum_{m=1}^{n} \xi_m - \kappa n, \tag{2.3}$$

is, in fact, the steady state queue distribution. In other words,

$$P(\{Q > x\}) = P(\{\tilde{Q}_{-1} > x\}) = P(\{\sup_{n \ge 0} X_n > x\}). \tag{2.4}$$

Relation (2.4), which comes originally from [23, 29], is very important in our study of the steady state queue distribution. We will study the suprema distribution $P(\{\sup_{n \ge 0} X_n > x\})$ to obtain asymptotic upper bounds to $P(\{Q > x\})$. It can be easily checked that ξ_n is a centered (zero mean) process and its autocovariance function C_ξ is the sum of the autocovariance functions of K independent input processes;

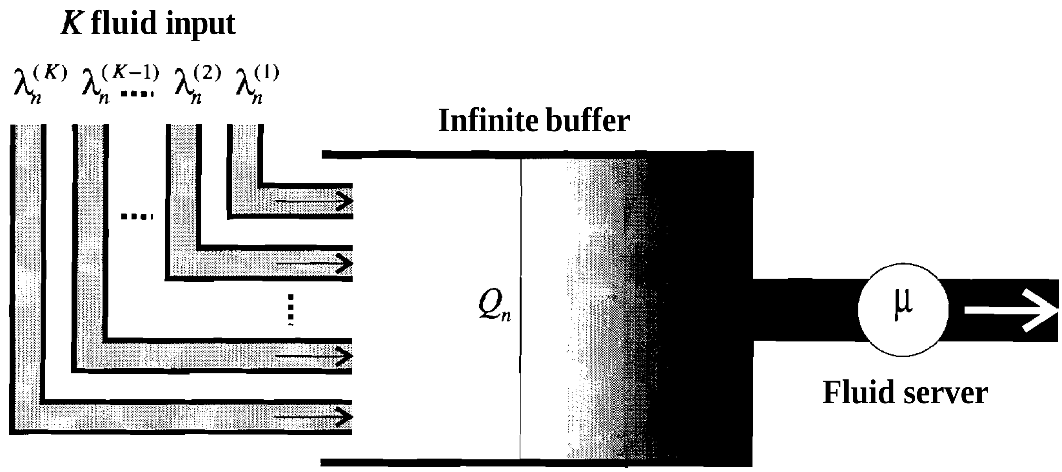

Figure 2.1: Fluid queue with K inputs and one server.

that is,

$$C_\xi(l) = \sum_{k=1}^{K} C_k(l), \tag{2.5}$$

where $C_k(l)$ is the autocovariance function of the k-th input.

From now on, let $\langle w \rangle_\Theta$ denote $\sup_{\theta \in \Theta} w_\theta$. We do not specify the index range Θ when it includes the entire domain of w_θ. Also, we make the following weak assumptions on the process ξ_n and the constant

(A1) ξ_n (or equivalently λ_n) is stationary and ergodic.

(A2) $\kappa < 0$ (or equivalently $\bar{\lambda} < \mu$, the stability condition for a single-server queue).

Next, we briefly motivate the Gaussian characterization of the aggregate traffic to the ATM multiplexer.

2.2 Central Limit Theorem in ATM networks

As mentioned earlier, we expect a large number of traffic sources to be multiplexed in ATM networks. For example, FORE SYSTEMS has already built commercial ATM switches to support *OC-12* (622.08 Mbps) lines, and ATM networks with *OC-24* (1.2 Gbps) lines are already operational (at Cambridge University). Due to the huge capacity of a single ATM link, we can expect that hundreds or even thousands of applications will share an ATM link; an *OC-3* (155.52 Mbps) line can accommodate over 7700 voice calls (32 Kbps during active period) and an *OC-12* line over 300 MPEG video calls both at a utilization of $\rho := \bar{\lambda}/\mu = 0.8$. These numbers seem to be large enough for the Central Limit Theorem to be applied to characterize the aggregate traffic to an ATM multiplexer by a Gaussian process [9, 10]; even though how precisely the Gaussian process reflects the aggregate traffic may depend on

- How many input processes are multiplexed.

- How input processes are distributed.

Unfortunately, it is not easy to analytically investigate the effect of these two major factors to the performance of the Gaussian characterization of the aggregate traffic. This is mainly because the effect of the Gaussian characterization of the aggregate traffic in queueing problems is not easily analyzable. Therefore, in our previous research [9, 10] we have studied the effect of these factors via numerical experiments, and found that several hundred independent inputs are usually sufficient for the Gaussian approximation to work well.

3. Asymptotic Upper Bound on $P(\{\langle X\rangle > x))$

In this chapter, we study the suprema distribution of the stochastic process X_n described by (2.3). We first introduce a few fundamental results from *Extreme Value Theory.* Then, in Section 3.2, we obtain several preliminary results, and from these results, in Section 3.3, we derive asymptotic upper bounds* to the tail probability $P(\{\langle X\rangle > x))$. Throughout this chapter, remember that $P(\{\langle X\rangle > x\}) = P(\{Q > x))$ when the stochastic process ξ_n and the constant κ are defined by (2.2). Therefore, through the study of the stochastic process X,, we will study properties of the steady state queue length distribution.

3.1 Extreme Value Theory

The following two inequalities from the Extreme Value Theory play key roles in our study of the suprema distribution of X, [3].

Theorem 1 (Borell's Inequality) *Let $\{\zeta_t : t \in T)$ be a centered Gaussian process with sample path bounded a.s.; that is $\langle\zeta\rangle < \infty$ a.s. Then $E\{\langle\zeta\rangle\}$ is finite and for all $x > E\{\langle\zeta\rangle\}$,*

$$P(\{\langle\zeta\rangle > x\}) \leq 2e^{-\frac{(x-E\{\langle\zeta\rangle\})^2}{2\langle\sigma^2\rangle}},$$

where $(a^2) := \sup_{t\in T} E\{\zeta_t^2\}$.

Theorem 2 (Slepian's Inequality) *Let ζ and ς be two centered Gaussian processes on an index set T with sample path bounded a.s. If $E\{\zeta_t^2\} = E\{\varsigma_t^2\}$ and $E\{(\zeta_s - \zeta_t)^2\} \leq E\{(\varsigma_s - \varsigma_t)^2\}$ for all $s,t \in T$, then for all* x

$$P(\{\langle\zeta\rangle > x\}) \leq P(\{\langle\varsigma\rangle > x\}).$$

Even though these two theorems are not presented in their strongest form (many variations and improvements can be found in the literature; e.g. [3, 26]), they are usually easier to apply because of their simple preconditions.

Also, to obtain an upper bound to the tail probability $P(\{\langle\zeta\rangle > x))$ using Borell's inequality, we will need the following theorem [3, Corollary 4.151, which provides us a way to bound $E\{\langle\zeta\rangle\}$, the mean of the supremum of ζ_t.

Theorem 3 *Let $\{\zeta_t : t \in T)$ be a centered Gaussian process and define a pseudo-metric d on T as $d(t_1,t_2) := \sqrt{E\{(\zeta_{t_1} - \zeta_{t_2})^2\}}$ (note that d is not a metric, since $d(t_1,t_2) = 0$ does not necessarily imply $t_1 = t_2$). Also, let $N(\epsilon)$ be the minimum number of closed d-balls of radius ϵ needed to cover T, then there exists a universal constant L such that*

$$E\{\langle\zeta\rangle\} \leq L\int_0^\infty \sqrt{\log N(\epsilon)}d\epsilon.$$

*We say f(X) asymptotically bounds $g(x)$ from above, if $\limsup_{x\to\infty} \frac{g(x)}{f(x)} \leq 1$.

3.2 Preliminaries

In this section we set the stage for our study of the suprema distribution of X_n. Those readers that are not interested in the detailed proofs can safely skip to Theorems 9.

We will derive asymptotic upper bounds to the tail probability $P(\{\langle X\rangle > x\})$ when the centered Gaussian process ξ_n given in (2.2) satisfies the following conditions:

(C1) $C_\xi(l)$ is absolutely summable and $\sum_{l=-\infty}^{\infty} C_\xi(l) > 0$.

(C2) $lC_\xi(l)$ is absolutely summable.

(C3) $\sum_{l=1}^{m} lC_\xi(l) + \sum_{l=m+1}^{\infty} mC_\xi(l) > 0$ for all m $= 1, 2, \ldots$ and $\sum_{l=1}^{\infty} lC_\xi(l) > 0$.

From (2.3), the mean and autocovariance function of X_n can be computed as

$$E\{X_n\} = -\kappa n, \quad \text{and} \tag{3.1}$$

$$C_X(n_1,n_2) := E\{(X_{n_1}+\kappa n_1)(X_{n_2}+\kappa n_2)\} = \sum_{m_1=1}^{n_1}\sum_{m_2=1}^{n_2} C_\xi(m_2-m_1). \tag{3.2}$$

In the following proposition, we show several important properties of the variance and the autocovariance function of X_n. Before we do that, however, we first define three parameters S, D, and $\tilde{D}$ which will be extensively used in the report.

$$S := \sum_{l=-\infty}^{\infty} C_\xi(l),$$

$$D := 2\sum_{l=0}^{\infty} lC_\xi(l), \quad \text{and}$$

$$\tilde{D} := 2\sum_{l=0}^{\infty} l|C_\xi(l)|.$$

Proposition 4

(a) For n > 1, $\frac{Var\{X_n\}}{n} - \frac{Var\{X_{n-1}\}}{n-1} = \frac{2}{n(n-1)}\sum_{m=1}^{n-1} mC_\xi(m)$.

(b) $C_X(n_1,n_2) = \frac{1}{2}\left(Var\{X_{n_1}\} + Var\{X_{n_2}\} - Var\{X_{|n_1-n_2|}\}\right)$

(c) Under condition (C1), $\lim_{n\to\infty} \frac{Var\{X_n\}}{n} = S$.

(d) Under conditions (C1) and (C2), $\left|\frac{Var\{X_{n_1}\}}{n_1} - \frac{Var\{X_{n_2}\}}{n_2}\right| \leq \frac{\tilde{D}|n_1-n_2|}{n_1 n_2}$ *for all* $n_1, n_2 > 0$, and $\lim_{n\to\infty} n\left(S - \frac{Var\{X_n\}}{n}\right) = D$.

(e) *Under conditions* (C1)-(C3), $\frac{Var\{X_n\}}{n} < S$ and there *exists* an n_o *such* that for all n $\geq n_o$, $\frac{Var\{X_n\}}{n} = \sup_{0<m\leq n} \frac{Var\{X_m\}}{m}$.

Proof of Proposition 4 : (a) From (3.2), we have

$$\frac{Var\{X_n\}}{n} = \frac{1}{n}\sum_{m_1=1}^{n}\sum_{m_2=1}^{n} C_\xi(m_2-m_1)$$

$$= C_\xi(0) + 2\sum_{m=1}^{n-1}(1-\frac{m}{n})C_\xi(m) \quad \text{(by changing variables m} = m_2 - m_1). \tag{3.3}$$

Therefore, for n > 1,

$$\frac{Var\{X_n\}}{n} - \frac{Var\{X_{n-1}\}}{n-1} = 2\sum_{m=1}^{n-1}(1-\frac{m}{n})C_\xi(m) - 2\sum_{m=1}^{n-2}(1-\frac{m}{n-1})C_\xi(m) = \frac{2}{n(n-1)}\sum_{m=1}^{n-1} mC_\xi(m).$$

(b) Without loss of generality (W.L.O.G) assume $n_2 > n_1$. Then,

$$\begin{aligned}
2C_X(n_1,n_2) &= \sum_{m_1=1}^{n_1}\sum_{m_2=1}^{n_2} C_\xi(m_2-m_1) + \sum_{m_1=1}^{n_2}\sum_{m_2=1}^{n_1} C_\xi(m_2-m_1) \\
&= \sum_{m_1=1}^{n_2}\sum_{m_2=1}^{n_2} C_\xi(m_2-m_1) - \sum_{m_1=n_1+1}^{n_2}\sum_{m_2=1}^{n_2} C_\xi(m_2-m_1) \\
&\quad + \sum_{m_1=1}^{n_1}\sum_{m_2=1}^{n_1} C_\xi(m_2-m_1) + \sum_{m_1=n_1+1}^{n_2}\sum_{m_2=1}^{n_1} C_\xi(m_2-m_1) \\
&= \sum_{m_1=1}^{n_2}\sum_{m_2=1}^{n_2} C_\xi(m_2-m_1) + \sum_{m_1=1}^{n_1}\sum_{m_2=1}^{n_1} C_\xi(m_2-m_1) - \\
&\quad \sum_{m_1=1}^{n_2-n_1}\sum_{m_2=1}^{n_2-n_1} C_\xi(m_2-m_1) \\
&= Var\{X_{n_2}\} + Var\{X_{n_1}\} - Var\{X_{n_2-n_1}\}.
\end{aligned}$$

(c) Let $h_n(m)$ be defined for $m = 0, 1, 2, \ldots$ as

$$h_n(m) := \begin{cases} \frac{1}{2}C_\xi(0) & \text{if m} = 0, \\ (1-\frac{m}{n})C_\xi(m) & \text{if } m = 1, 2, \ldots, n-1, \\ 0 & \text{otherwise.} \end{cases}$$

Then, it follows from (3.3) that $\frac{Var\{X_n\}}{n} = 2\sum_{m=0}^{\infty} h_n(m)$. On the other hand, from the definition of h_n(m) we know that $|h_n(m)| \le |C_\xi(\text{m})|$ and

$$\lim_{n\to\infty} h_n(m) = \begin{cases} \frac{1}{2}C_\xi(0) & \text{if } m = 0, \\ C_\xi(\text{m}) & \text{otherwise.} \end{cases}$$

Therefore, from condition (C1) and the Dominated *Convergence* Theorem (DCT) [5], we have

$$\lim_{n\to\infty} \frac{Var\{X_n\}}{n} = \sum_{\text{m=O}}^{\infty} \lim_{n\to\infty} h_n(m) = C_\xi(0) + 2\sum_{m=1}^{\infty} C_\xi(m) = S.$$

(d) W.L.O.G. assume $n_2 > n_1 > 0$. From (3.3), we have

$$\begin{aligned}
\frac{Var\{X_{n_2}\}}{n_2} - \frac{Var\{X_{n_1}\}}{n_1} &= 2\left(\sum_{m=1}^{n_2-1}(1-\frac{m}{n_2})C_\xi(m) - \sum_{m=1}^{n_1-1}(1-\frac{m}{n_1})C_\xi(m)\right) \\
&= \frac{2(n_2-n_1)}{n_1 n_2}\left(\sum_{m=1}^{n_1-1} mC_\xi(m) + \sum_{m=n_1}^{n_2-1}\frac{n_1(n_2-m)}{n_2-n_1}C_\xi(m)\right)
\end{aligned}$$

Since $0 \le \frac{n_1(n_2-m)}{n_2-n_1} \le$ m for m $= n_1, n_1 + 1, \ldots, n_2 - 1$, it follows that

$$\left|\frac{Var\{X_{n_2}\}}{n_2} - \frac{Var\{X_{n_1}\}}{n_1}\right| \le \frac{2(n_2-n_1)}{n_1 n_2}\left(\sum_{m=1}^{n_1-1} m|C_\xi(m)| + \sum_{m=n_1}^{n_2-1}\frac{n_1(n_2-m)}{n_2-n_1}|C_\xi(m)|\right)$$

$$\leq \frac{2(n_2-n_1)}{n_1n_2}\sum_{m=1}^{n_2-1} m|C_\xi(m)| \leq \frac{\tilde{D}(n_2-n_1)}{n_1n_2}$$

Now, let $h_n(m)$ be defined as

$$h_n(\text{m}) := \begin{cases} mC_\xi(m) & \text{if } m = 0,1,\ldots,n, \\ nC_\xi(\text{m}) & \text{otherwise.} \end{cases}$$

Then, from **(3.3)** and the definition of S,

$$n\left(S - \frac{Var\{X_n\}}{n}\right) = 2n\left(\sum_{m=1}^{\infty} C_\xi(m) - \sum_{m=1}^{n-1}(1-\frac{m}{n})C_\xi(m)\right) = 2\sum_{m=1}^{\infty} h_n(m).$$

Again, we know that $h_n(m) \to mC_\xi(m)$ as $n \to \infty$ and $|h_n(m)| \leq m|C_\xi(m)|$. Therefore, from condition (C2) and DCT, $\lim_{n\to\infty} n\left(S - \frac{Var\{X_n\}}{n}\right) = 2\sum_{m=1}^{\infty} mC_\xi(m) = D$.

(e) From (3.3) and the definition of S,

$$\begin{aligned} n\left(S - \frac{Var\{X_n\}}{n}\right) &= 2n\left(\sum_{m=1}^{\infty} C_\xi(m) - \sum_{m=1}^{n-1}(1-\frac{m}{n})C_\xi(m)\right) \\ &= 2\left(\sum_{m=1}^{n} mC_\xi(m) + \sum_{m=n+1}^{\infty} nC_\xi(m)\right) > 0 \quad \text{(from condition (C3))}. \end{aligned}$$

Therefore, $\frac{Var\{X_n\}}{n} < S$ for all $n > 0$. From conditions (C2) and (C3), it follows that $\lim_{n\to\infty}\sum_{m=1}^{n} mC_\xi(m) = \sum_{m=1}^{\infty} mC_\xi(m) > 0$. This with (a) implies that there exists an $n_1 > 0$ such that $\frac{Var\{X_n\}}{n} - \frac{Var\{X_{n-1}\}}{n-1} > 0$ for all $n \geq n_1$; that is, $\frac{Var\{X_n\}}{n}$ is an increasing function for $n \geq n_1$. Now, let $c := \sup_{0<m\leq n_1} \frac{Var\{X_m\}}{m}$, then $c < S$, and from (c) there exists an $n_o \geq n_1$ such that $\frac{Var\{X_{n_o}\}}{n_o} > c$. Let $n \geq n_o$, then for $m \leq n_1$,

$$\begin{aligned} \frac{Var\{X_m\}}{m} \leq c &\leq \frac{Var\{X_{n_o}\}}{n_o} \quad \text{(from the definition of } n_o) \\ &\leq \frac{Var\{X_n\}}{n} \quad \text{(because } \tfrac{Var\{X_n\}}{n} \text{ is increasing for } n \geq n_1). \end{aligned}$$

Also, since $\frac{Var\{X_n\}}{n}$ is increasing for $n \geq n_1$, $\frac{Var\{X_m\}}{m} \leq \frac{Var\{X_n\}}{n}$ for $m \in (n_1, n)$. Therefore, for all $n \geq n_o$, $\frac{Var\{X_n\}}{n} = \sup_{0<m\leq n}\frac{Var\{X_m\}}{m}$. Q.E.D.

These properties will be extensively used in our study of the suprema distribution of X_n. For simplicity, we now define a new Gaussian process $Y_n^{(x)}$ for each $x > 0$ as

$$Y_n^{(x)} := \frac{\sqrt{x}(X_n + \kappa n)}{x + \kappa n} = \frac{\sqrt{x}\sum_{m=1}^{n}\xi_m}{x+\kappa n} \tag{3.4}$$

The following lemma is a direct result of the definition of $Y_n^{(x)}$, and plays a key role in obtaining bounds to the tail probability $P(\{Q > x\})$.

Lemma 5 For any $n \in \{0,1,2,\ldots)$, $X_n > x$ if and only if $Y_n^{(x)} > \sqrt{x}$

From the definition of $Y_n^{(x)}$, it immediately follows that $Y_n^{(x)}$ has zero mean, and its autocovariance function $C_Y^{(x)}$ is given by

$$C_Y^{(x)}(n_1, n_2) = \frac{x C_X(n_1, n_2)}{(x + \kappa n_1)(x + \kappa n_2)}. \tag{3.5}$$

Now, let $\sigma_{x,n}^2$ be the variance of $Y_n^{(x)}$, then it can be expressed in terms of $Var\{X_n\}$ as

$$\sigma_{x,n}^2 = \frac{x Var\{X_n\}}{(x + \kappa n)^2}. \tag{3.6}$$

From Proposition 4(c), it follows that $\sigma_{x,n}^2 \to 0$ as $n \to \infty$. Therefore, $\sigma_{x,n}^2$ should attain its maximum $\langle \sigma_x^2 \rangle$ at some finite value of $n = \hat{n}_x$ (note that $\langle \sigma_x^2 \rangle$ denotes the supremum of $\sigma_{x,n}^2$ over the time index n). In the following proposition (Proposition 6) we show an important property of $\hat{n}_x$. Before we introduce this proposition, for notational simplicity, we define a function $g(n)$ for $n = 0, 1, 2, \ldots$ as

$$g(n) := \begin{cases} 0 & \text{if } n = 0, \\ \frac{Var\{X_n\}}{Sn} & \text{otherwise.} \end{cases}$$

Then we can write the variance of $Y_n^{(x)}$ in terms of function $g(n)$ as

$$\sigma_{x,n}^2 = \frac{Sxn}{(x + \kappa n)^2} g(n). \tag{3.7}$$

Proposition 6 *Under condition* (C1), $\hat{n}_x \sim \frac{x}{\kappa}$ *as* $x \to \infty$. *Further, under conditions* (C1) *and* (C2), $\lim_{x\to\infty} \frac{\hat{n}_x - \frac{x}{\kappa}}{x^\epsilon} = 0$ *for* all $\epsilon > 0$.

Proof of Proposition 6 : From Proposition 4(c), we have $\lim_{n\to\infty} g(n) = 1$. Let $G := \sup_{n\geq 0} g(n)$ and n, be the nonnegative integer at which $\frac{xn}{(x+\kappa n)^2}$ attains its maximum. Then, it follows that G is finite and not less than 1, and $|\dot{n}_x - \frac{x}{\kappa}| \leq 1$. Since $\sigma_{x,n}^2$ attains its maximum at $n = \hat{n}_x$,

$$\frac{Sx\dot{n}_x g(\dot{n}_x)}{(x + \kappa\dot{n}_x)^2} = \sigma_{x,\dot{n}_x}^2 \leq \sigma_{x,\hat{n}_x}^2 = \frac{Sx\hat{n}_x g(\hat{n}_x)}{(x + \kappa\hat{n}_x)^2} \leq \frac{Sx\hat{n}_x G}{(x + \kappa\hat{n}_x)^2}. \tag{3.8}$$

By solving (3.8) for $\hat{n}_x$, we have

$$\begin{aligned} &\left(\left(\frac{G(\frac{x}{\kappa} + \dot{n}_x)^2}{2\dot{n}_x \frac{x}{\kappa} g(\dot{n}_x)} - 1 \right) - 2\sqrt{\frac{G(\frac{x}{\kappa} + \dot{n}_x)^2}{4\dot{n}_x \frac{x}{\kappa} g(\dot{n}_x)} \left(\frac{G(\frac{x}{\kappa} + \dot{n}_x)^2}{4\dot{n}_x \frac{x}{\kappa} g(\dot{n}_x)} - 1 \right)} \right) \frac{x}{\kappa} \leq \hat{n}_x \\ &\leq \left(\left(\frac{G(\frac{x}{\kappa} + \dot{n}_x)^2}{2\dot{n}_x \frac{x}{\kappa} g(\dot{n}_x)} - 1 \right) + 2\sqrt{\frac{G(\frac{x}{\kappa} + \dot{n}_x)^2}{4\dot{n}_x \frac{x}{\kappa} g(\dot{n}_x)} \left(\frac{G(\frac{x}{\kappa} + \dot{n}_x)^2}{4\dot{n}_x \frac{x}{\kappa} g(\dot{n}_x)} - 1 \right)} \right) \frac{x}{\kappa}. \end{aligned}$$

Since $\frac{\kappa \dot{n}_x}{x}, g(\dot{n}_x) \to 1$ as $x \to \infty$, this inequality implies that $\hat{n}_x \to \infty$ (consequently, $g(\hat{n}_x) \to 1$) as $x \to \infty$.

Since $\frac{Sxn}{(x+\kappa n)^2}$ attains its maximum at $n = \dot{n}_x$, we know from (3.8) that $g(\dot{n}_x) \leq g(\hat{n}_x)$, and that the following relation should hold.

$$\begin{aligned} &\left(\left(\frac{g(\hat{n}_x)(\frac{x}{\kappa} + \dot{n}_x)^2}{2\dot{n}_x \frac{x}{\kappa} g(\dot{n}_x)} - 1 \right) - 2\sqrt{\frac{g(\hat{n}_x)(\frac{x}{\kappa} + \dot{n}_x)^2}{4\dot{n}_x \frac{x}{\kappa} g(\dot{n}_x)} \left(\frac{g(\hat{n}_x)(\frac{x}{\kappa} + \dot{n}_x)^2}{4\dot{n}_x \frac{x}{\kappa} g(\dot{n}_x)} - 1 \right)} \right) \frac{x}{\kappa} \leq \hat{n}_x \\ &\leq \left(\left(\frac{g(\hat{n}_x)(\frac{x}{\kappa} + \dot{n}_x)^2}{2\dot{n}_x \frac{x}{\kappa} g(\dot{n}_x)} - 1 \right) + 2\sqrt{\frac{g(\hat{n}_x)(\frac{x}{\kappa} + \dot{n}_x)^2}{4\dot{n}_x \frac{x}{\kappa} g(\dot{n}_x)} \left(\frac{g(\hat{n}_x)(\frac{x}{\kappa} + \dot{n}_x)^2}{4\dot{n}_x \frac{x}{\kappa} g(\dot{n}_x)} - 1 \right)} \right) \frac{x}{\kappa}. \end{aligned} \tag{3.9}$$

Since both $g(\dot{n}_x)$ and $g(\hat{n}_x)$ approach 1 as x $\to \infty$, this inequality implies that

$$\lim_{x\to\infty} \frac{\kappa \hat{n}_x}{x} = 1. \tag{3.10}$$

Thus, we have proven the first part of the proposition. Now, assume $C_\xi(l)$ satisfies conditions (C1) and (C2). From Proposition 4(d), note that

$$|g(n_1) - g(n_2)| \le \frac{\tilde{D}|n_2 - n_1|}{S n_1 n_2}. \tag{3.11}$$

From (3.9), it follows that

$$\left|\hat{n}_x - \frac{x}{\kappa}\right| \le \frac{2x}{\kappa}\left(\left|\frac{g(\hat{n}_x)(\frac{x}{\kappa}+\dot{n}_x)^2}{4\dot{n}_x\frac{x}{\kappa}g(\hat{n}_x)} - 1\right| + \sqrt{\frac{g(\hat{n}_x)(\frac{x}{\kappa}+\dot{n}_x)^2}{4\dot{n}_x\frac{x}{\kappa}g(\hat{n}_x)}\left|\frac{g(\hat{n}_x)(\frac{x}{\kappa}+\dot{n}_x)^2}{4\dot{n}_x\frac{x}{\kappa}g(\hat{n}_x)} - 1\right|}\right). \tag{3.12}$$

On the other hand,

$$\left|\frac{g(\hat{n}_x)(\frac{x}{\kappa}+\dot{n}_x)^2}{4\dot{n}_x\frac{x}{\kappa}g(\hat{n}_x)} - 1\right| \le \frac{1}{g(\dot{n}_x)}\left(\frac{(\frac{x}{\kappa}+\dot{n}_x)^2}{4\dot{n}_x\frac{x}{\kappa}}|g(\hat{n}_x) - g(\dot{n}_x)| + g(\dot{n}_x)\frac{(\frac{x}{\kappa}-\dot{n}_x)^2}{4\dot{n}_x\frac{x}{\kappa}}\right). \tag{3.13}$$

Since $\frac{(\frac{x}{\kappa}+\dot{n}_x)^2}{4\dot{n}_x\frac{x}{\kappa}}$, $\frac{\kappa\dot{n}_x}{x}$ and $g(\dot{n}_x)$ approach 1 as x $\to\infty$ and since $|\frac{x}{\kappa} - \dot{n}_x| \le 1$, it follows from (3.13) that for sufficiently large x,

$$\left|\frac{g(\hat{n}_x)(\frac{x}{\kappa}+\dot{n}_x)^2}{4\dot{n}_x\frac{x}{\kappa}g(\hat{n}_x)} - 1\right| \le 2|g(\hat{n}_x) - g(\dot{n}_x)| + \frac{Is^2}{x^2}. \tag{3.14}$$

Therefore, from (3.12) and (3.14), for sufficiently large x, we have

$$\begin{aligned}
\left|\hat{n}_x - \frac{x}{\kappa}\right| &\le \frac{4|g(\hat{n}_1) - g(\dot{n}_x)|}{\frac{\kappa}{x}} + \frac{2\kappa}{x} + \sqrt{16\frac{|g(\hat{n}_x) - g(\dot{n}_x)|}{\frac{\kappa^2}{x^2}} + 8} \\
&\quad \text{(from the fact that } \tfrac{g(\hat{n}_x)(\frac{x}{\kappa}+\dot{n}_x)^2}{4\dot{n}_x\frac{x}{\kappa}g(\hat{n}_x)} \to 1 \text{ as } x\to\infty) \\
&\le \frac{4\tilde{D}|\hat{n}_x - \dot{n}_x|}{S\frac{\kappa}{x}\hat{n}_x\dot{n}_x} + \frac{21s}{x} + \sqrt{16\frac{\tilde{D}|\hat{n}_x - \dot{n}_x|}{S\hat{n}_x\dot{n}_x\frac{\kappa^2}{x^2}} + 8} \quad \text{(from (3.11))} \\
&\le 1 + \frac{2\kappa}{x} + \sqrt{32\frac{\tilde{D}|\hat{n}_x - \dot{n}_x|}{S} + 8} \qquad (3.15)\\
&\quad \text{(since } \tfrac{|\hat{n}_x - \dot{n}_x|}{n_x} \to 0 \text{ and } \tfrac{\hat{n}_x\kappa}{x}, \tfrac{\dot{n}_x\kappa}{x} \to 1 \text{ as } x\to\infty).
\end{aligned}$$

Now, assume that $\lim_{\infty} \frac{\hat{n}_x - \frac{x}{\kappa}}{x^\epsilon} = 0$ for some $\epsilon > 0$ (from (3.10), we already know that this holds for any $\epsilon > 1$). Then, since $|\hat{n}_x - \dot{n}_x| \le |\hat{n}_x - \frac{x}{\kappa}| + 1$, from (3.15) we have

$$\left|\frac{\hat{n}_x - \frac{x}{\kappa}}{x^{\frac{\epsilon}{2}}}\right| \le x^{-\frac{\epsilon}{2}} + \frac{2\kappa}{x^{1+\frac{\epsilon}{2}}} + \sqrt{32\frac{\tilde{D}|\hat{n}_x - \dot{n}_x|}{Sx^\epsilon} + \frac{8}{x^\epsilon}} \to 0 \quad \text{as } x\to\infty.$$

Hence, $\lim_{x\to\infty} \frac{\hat{n}_x - \frac{x}{\kappa}}{x^{\frac{\epsilon}{2}}} = 0$. Thus it follows by induction that $\lim_{\infty} \frac{\hat{n}_x - \frac{x}{\kappa}}{x^\epsilon} = 0$, for all $\epsilon > 0$. *Q.E.D.*

The following proposition describes the asymptotic behavior of $\langle\sigma_x^2\rangle$ is a direct result of Proposition 4(c) and Proposition 6.

Proposition 7 *Under condition* (C1), $\lim_{x\to\infty}\langle\sigma_x^2\rangle = \frac{S}{4\kappa}$.

Proof of Proposition 7 : From (3.6), we have $\langle\sigma_x^2\rangle = \frac{x Var\{X_{\hat{n}_x}\}}{(x+\kappa\hat{n}_x)^2} = \frac{1}{\kappa}\frac{Var\{X_{\hat{n}_x}\}}{\hat{n}_x}\frac{\kappa\hat{n}_x}{x}\frac{1}{(1+\frac{\kappa\hat{n}_x}{x})^2}$. However, we know that $\frac{Var\{X_{\hat{n}_x}\}}{\hat{n}_x} \to S$ (Proposition 4(c)) and $\frac{\kappa\hat{n}_x}{x} \to 1$ (Proposition 6), as x $\to$ m. Thus, $\lim_{x\to\infty}\langle\sigma_x^2\rangle = \frac{S}{4\kappa}$. *Q.E.D.*

3.3 Main Results

In our previous papers [9, 10], we have derived a lower bound to $P(\{\langle X\rangle > x\})$ in terms of maximum variance $\langle\sigma_x^2\rangle$ from the following simple relation.

$$
\begin{aligned}
P(\{\langle X\rangle > x\}) &\geq P(\{X_{\hat{n}_x} > x\}) \\
&= P(\{Y_{\hat{n}_x}^{(x)} > \sqrt{x}\}) = \Psi\left(\frac{\sqrt{x}}{\langle\sigma_x\rangle}\right) \quad \text{(from Lemma 5).}
\end{aligned}
\tag{3.16}
$$

Here $\Psi(x) := \frac{1}{\sqrt{2\pi}}\int_x^\infty e^{-\frac{y^2}{2}}dy$ is the tail of the standard Gaussian distribution.

Using the following well known inequality for function Ψ [17],

$$
\frac{1-y^{-2}}{\sqrt{2\pi}}y^{-1}e^{-\frac{y^2}{2}} \leq \Psi(y) \leq \frac{1}{\sqrt{2\pi}}y^{-1}e^{-\frac{y^2}{2}} \quad \text{for all } y > 0,
\tag{3.17}
$$

we have shown that the lower bound is in fact asymptotically similar to the tail probability $P(\{\langle X\rangle > x\})$ in the logarithmic sense (also see [9]); that is,

$$
\log\Psi\left(\frac{\sqrt{x}}{\langle\sigma_x\rangle}\right) \sim \log P(\{\langle Y^{(x)}\rangle > \sqrt{x}\}) = \log P(\{\langle X\rangle > x\}) \quad \text{as } x\to\infty.
$$

This lower bound has been used to approximate the tail probability $P(\{\langle X\rangle > x\})$ (or equivalently, to approximate $P(\{Q > x\})$) in [9, 10]. In these papers, it has been shown through many numerical examples, that the lower bound accurately approximates the tail probability for a very wide range of values of x. Noting that the lower bound is the probability that X_n is greater than x at only one point $\hat{n}_x$ out of the whole index set $\{0,1,2\ldots\}$, the reader may wonder how this lower bound can be so close to the tail probability. However, in the Extreme Value Theory for Gaussian processes, the maximum variance of a centered Gaussian process with nonconstant variance, has been frequently emphasized as a very important factor in studying the suprema distribution of the Gaussian process (also as can be seen in Borell's inequality) [3, 4, 26, 30]. In addition, it has been found in various forms, that the behavior of a centered Gaussian process around the index, at which the maximum variance is attained, essentially determines its suprema distribution. Therefore, it seems natural to expect that there should be a neighborhood F_x around $\hat{n}_x$ (or around $\frac{x}{\kappa}$) such that $P(\{\langle Y^{(x)}\rangle_{F_x} > \sqrt{x}\}) \sim P(\{\langle Y^{(x)}\rangle > \sqrt{x}\})$ as x $\to$ m. The following theorem validates our expectation and will be used to obtain an asymptotic upper bound to $P(\{\langle X\rangle > x\})$.

Theorem 8 *Under condition* (C1)*, for any* a >1,

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

Proof of Theorem *8* : To prove the theorem, it suffices to show that

$$\lim_{x\to\infty}\frac{P(\{\langle Y^{(x)}\rangle_{[\frac{x}{\alpha\kappa},\frac{\alpha x}{\kappa}]^c}>\sqrt{x}\})}{P(\{\langle Y^{(x)}\rangle>\sqrt{x}\})}=0$$

for all $a>1$, where A^c denotes the complementary set of A.

Let $a>1$. Since $g(n)\to 1$ as n $\to\infty$, there exists an n_o such that $g(n)\le\frac{\alpha+1}{2\sqrt{\alpha}}$ for all n $\ge n_o$. Now, let G := $\sup_{n\ge 0}g(n)$, then there exists an x, $>\alpha\kappa n_o$ such that $\frac{Sxn_oG}{(x+\kappa n_o)^2}\le\frac{S\sqrt{\alpha}}{2\kappa(\alpha+1)}$ for all x $\ge x_o$. Since $\frac{SxnG}{(x+\kappa n)^2}$ is an increasing function for n $\le\frac{x}{\kappa}$, this (in conjunction with (3.7)) implies that,

$$\sigma^2_{x,n}\le\frac{SxnG}{(x+\kappa n)^2}\le\frac{S\sqrt{\alpha}}{2\kappa(\alpha+1)}\quad\text{for all } x\ge \text{x, and n}\le n_o. \tag{3.18}$$

It can easily be shown that $\frac{Sxn}{(x+\kappa n)^2}\le\frac{S\alpha}{\kappa(\alpha+1)^2}$ for n $\in[\frac{x}{\alpha\kappa},\frac{\alpha x}{\kappa}]^c$. Therefore, from the definition of n_o, we have

$$\sigma^2_{x,n}=\frac{Sxng(n)}{(x+\kappa n)^2}\le\frac{S\sqrt{\alpha}}{2\kappa(\alpha+1)}\quad\text{for } x\ge x_o \text{ and } n\in(n_o,\tfrac{x}{\alpha\kappa})\cup(\tfrac{\alpha x}{\kappa},\infty). \tag{3.19}$$

Now from (3.18) and (3.19), it follows that

$$\langle\sigma^2_x\rangle_{[\frac{x}{\alpha\kappa},\frac{\alpha x}{\kappa}]^c}\le\frac{S\sqrt{\alpha}}{2\kappa(\alpha+1)}\quad\text{for all x}\ge x_o. \tag{3.20}$$

We now define a pseudo-metric $d^{(x)}$ on $\{0,1,2,\ldots\}$ as $d^{(x)}(n_1,n_2):=\sqrt{E\{(Y^{(x)}_{n_2}-Y^{(x)}_{n_1})^2\}}$. Also, let $B^{(x)}_\epsilon(n):=\{$m : $d^{(x)}(n,m)\le$ a) be a $d^{(x)}$-ball of radius ϵ centered at n, and $N^{(x)}(\epsilon)$ be the minimum number of $d^{(x)}$-balls of radius of ϵ needed to cover $\{0,1,2,\ldots)$. Since $Var\{Y^{(x)}_n\}\le\frac{SGxn}{(x+\kappa n)^2}\le\frac{SG}{4\kappa}$ and since $Y^{(x)}_0=0$, $B^{(x)}_\epsilon(0)$ covers $\{0,1,2,\ldots)$ when a $\ge\sqrt{\frac{SG}{4\kappa}}$. Therefore, for all x >0,

$$N^{(x)}(\epsilon)=1\quad\text{for }\epsilon\ge\sqrt{\frac{SG}{4\kappa}}. \tag{3.21}$$

Now, assume that a $<\sqrt{\frac{SG}{4\kappa}}$ and $n_2>n_1$. Then,

$$\begin{aligned}
d^{(x)}(n_1,n_2) &= \sqrt{E\left\{\left(\tfrac{\sqrt{x}(X_{n_2}+\kappa n_2)}{x+\kappa n_2}-\tfrac{\sqrt{x}(X_{n_1}+\kappa n_1)}{x+\kappa n_1}\right)^2\right\}}\\
&= \sqrt{E\left\{\left(\tfrac{\sqrt{x}(X_{n_2}+\kappa n_2)}{x+\kappa n_2}-\tfrac{\sqrt{x}(X_{n_1}+\kappa n_1)}{x+\kappa n_2}+\tfrac{\sqrt{x}(X_{n_1}+\kappa n_1)}{x+\kappa n_2}-\tfrac{\sqrt{x}(X_{n_1}+\kappa n_1)}{x+\kappa n_1}\right)^2\right\}}\\
&\le \sqrt{E\left\{\left(\tfrac{\sqrt{x}(X_{n_2}-X_{n_1}+\kappa(n_2-n_1))}{x+\kappa n_2}\right)^2\right\}}+\sqrt{E\left\{\left(\tfrac{\kappa(n_2-n_1)\sqrt{x}(X_{n_1}+\kappa n_1)}{(x+\kappa n_2)(x+\kappa n_1)}\right)^2\right\}}\\
&= \tfrac{\sqrt{x}}{x+\kappa n_2}\sqrt{Var\{(X_{n_2}-X_{n_1})\}}+\tfrac{\kappa(n_2-n_1)\sqrt{x}}{(x+\kappa n_2)(x+\kappa n_1)}\sqrt{Var\{X_{n_1}\}}.
\end{aligned} \tag{3.22}$$

However, since $Var\{(X_{n_2} - X_{n_1})\} = Var\{X_{n_2-n_1}\}$ from the stationary increment property of X_n, $Var\{(X_{n_2}-X_{n_1})\}$ and $Var\{X_{n_1}\}$ are bounded by $GS(n_2-n_1)$ and GSn_1, respectively. Hence, from (3.22)

$$\begin{aligned} d^{(x)}(n_1,n_2) &\leq \frac{\sqrt{SGx(n_2-n_1)}}{x+\kappa n_2} + \frac{\kappa(n_2-n_1)\sqrt{SGxn_1}}{(x+\kappa n_1)(x+\kappa n_2)} \\ &\leq \left(\frac{\sqrt{SGx}}{x+\kappa n_2} + \frac{\kappa\sqrt{SGxn_1n_2}}{(x+\kappa n_1)(x+\kappa n_2)}\right)\sqrt{n_2-n_1} \\ &\leq \left(\sqrt{\frac{SG}{x}} + \frac{1}{4}\sqrt{\frac{SG}{x}}\right)\sqrt{n_2-n_1} \leq \sqrt{\frac{2SG}{x}}\sqrt{n_2-n_1} \end{aligned} \tag{3.23}$$

(from the fact that $\frac{\sqrt{x}}{x+\kappa n_2} \leq \frac{1}{\sqrt{x}}$ and $\frac{\sqrt{n}}{(x+\kappa n)} \leq \frac{1}{2\sqrt{x\kappa}}$).

This implies that if $|n_2 - n_1| \leq \frac{x}{2SG}\epsilon^2$, then $d^{(x)}(n_1,n_2) \leq \epsilon$. Consequently,

$$[n - \frac{x}{2SG}\epsilon^2, n + \frac{x}{2SG}\epsilon^2] \subset B_\epsilon^{(x)}(n). \tag{3.24}$$

Also, it can be easily shown that $Var\{Y_n^{(x)}\} \leq \epsilon^2$ for $n \geq \frac{SGx}{\epsilon^2\kappa^2}$. Since $Y_0^{(x)} = 0$, this implies that

$$[\frac{SGx}{\epsilon^2\kappa^2}, \infty) \subset B_\epsilon^{(x)}(0). \tag{3.25}$$

Now, let $k = \lceil \frac{x}{2SG}\epsilon^2 \rceil$, where $\lceil x \rceil$ denotes the smallest integer greater than or equal to x. Then, from (3.24) and (3.25), it follows that $d^{(x)}$-balls of radius of ϵ centered at ki ($i = 0, 1, \ldots, \lceil \frac{SGx}{k\epsilon^2\kappa^2} \rceil$) cover $\{0, 1, 2, \ldots\}$. Hence, for $\epsilon < \sqrt{\frac{SG}{4\kappa}}$, $N^{(x)}(\epsilon)$ is bounded by the following inequality.

$$N^{(x)}(\epsilon) \leq \left\lceil \frac{SGx}{k\epsilon^2\kappa^2} \right\rceil + 1 \leq \frac{2S^2G^2}{\kappa^2\epsilon^4} + 2. \tag{3.26}$$

From (3.21) and (3.26), $\bar{N}(\epsilon)$ defined by

$$\bar{N}(\epsilon) := \begin{cases} \frac{2S^2G^2}{\kappa^2\epsilon^4} + 2 & \text{if } \epsilon < \sqrt{\frac{SG}{4\kappa}}, \\ 1 & \text{otherwise,} \end{cases}$$

bounds $N^{(x)}(\epsilon)$ for all $x, \epsilon > 0$. Now, let $M := L\int_0^\infty \log^{\frac{1}{2}} \bar{N}(\epsilon)d\epsilon$ (it can be shown that the integral is finite). Then from Theorem 3

$$E\{\langle Y^{(x)}\rangle\} \leq M, \quad \text{for all } x > 0. \tag{3.27}$$

By applying Theorem 1 to $Y_n^{(x)}$ for $n \in [\frac{x}{\alpha\kappa}, \frac{\alpha x}{\kappa}]^c$, we get

$$\begin{aligned} P(\{\langle Y^{(x)}\rangle_{[\frac{x}{\alpha\kappa},\frac{\alpha x}{\kappa}]^c} > \sqrt{x}\}) &\leq 2e^{-\frac{\left(\sqrt{x}-E\left\{\langle Y^{(x)}\rangle_{[\frac{x}{\alpha\kappa},\frac{\alpha x}{\kappa}]^c}\right\}\right)^2}{2\langle\sigma_x^2\rangle_{[\frac{x}{\alpha\kappa},\frac{\alpha x}{\kappa}]^c}}} \\ &\leq 2e^{-\frac{\kappa(\sqrt{x}-E\{\langle Y^{(x)}\rangle\})^2(\alpha+1)}{S\sqrt{\alpha}}} \\ &\qquad \text{(from (3.20) and the fact that } \langle Y^{(x)}\rangle_{[\frac{x}{\alpha\kappa},\frac{\alpha x}{\kappa}]^c} \leq \langle Y^{(x)}\rangle) \\ &\leq 2e^{-\frac{\kappa(\sqrt{x}-M)^2(\alpha+1)}{S\sqrt{\alpha}}} \quad \text{(from (3.27))}, \end{aligned} \tag{3.28}$$

for x sufficiently large. Therefore,

$$\liminf_{x\to\infty} -\frac{1}{x}\log P(\{\langle Y^{(x)}\rangle_{[\frac{x}{\alpha\kappa},\frac{\alpha x}{\kappa}]^c} > \sqrt{x}\}) \geq \lim_{x\to\infty} \frac{\kappa(\sqrt{x}-M)^2(\alpha+1)}{Sx\sqrt{\alpha}} = \frac{\kappa(\alpha+1)}{S\sqrt{\alpha}}. \tag{3.29}$$

Additionally, it has been shown for very general Gaussian input processes [2, 19] that

$$\lim_{x\to\infty} -\frac{1}{x}\log P(\{\langle Y^{(x)}\rangle > \sqrt{x}\}) = \lim_{x\to\infty} -\frac{1}{x}\log P(\{\langle X\rangle > x\}) = \frac{2\kappa}{S}. \tag{3.30}$$

Since $\frac{\kappa(\alpha+1)}{S\sqrt{\alpha}} > \frac{2\kappa}{S}$ for all $a > 1$, (3.29) and (3.30) imply that

$$\lim_{x\to\infty} \frac{P(\{\langle Y^{(x)}\rangle_{[\frac{x}{\alpha\kappa},\frac{\alpha x}{\kappa}]^c} > \sqrt{x}\})}{P(\{\langle Y^{(x)}\rangle > \sqrt{x}\})} = 0,$$

and the theorem follows. *Q.E.D.*

So far we have considered the stochastic process X_n expressed by (2.3). Now, as a special case of such processes, consider a Gaussian random walk V_n with variance $Var\{V_n\} = a^2n$ and drift $E\{V_n\} = -bn$; that is,

$$V_n = aB_n - bn. \tag{3.31}$$

Here $B_n := \sum_{m=1}^{n} \chi_m$ denotes a standard Gaussian random walk, where $\{\chi_n : n = 1, 2, \ldots\}$ are centered *i.i.d.* Gaussian sequence with unit variance. This special case has received a lot of interest. An upper bound to the tail of its suprema distribution $P(\{\langle V\rangle > x\})$ is well known [25, page 236] and given by

$$P(\{\langle V\rangle > x\}) = P(\{aB_n > x + bn \quad \text{for some } n = 0,1,2,\ldots\}) \leq e^{-\frac{2bx}{a^2}}. \tag{3.32}$$

Using this result, we now derive an asymptotic upper bound to the tail probability $P(\{\langle X\rangle > x\})$.

Theorem 9 Under conditions (C1)-(C3), $\limsup_{x\to\infty} e^{\frac{2\kappa x}{S}} P(\{\langle X\rangle > x\}) \leq e^{-\frac{2\kappa^2 D}{S^2}}$. In other words, $e^{-\frac{2\kappa}{S}(x+\frac{\kappa D}{S})}$ asymptotically bounds $P(\{\langle X\rangle > x\})$.

Proof of *Theorem* 9 : Let $V_n := \sqrt{S}B_n - \kappa n$, and define a centered Gaussian process $Z_n^{(x)}$ $(n = 0, 1, \ldots)$ for each $x > 0$ by

$$Z_n^{(x)} := \frac{\sqrt{xg(n)}(V_n + \kappa n)}{x + \kappa n}.$$

From the definition, the autocovariance function $C_Z^{(x)}$ of $Z_n^{(x)}$ can be easily derived as

$$C_Z^{(x)}(n_1, n_2) = \frac{Sx\min\{n_1, n_2\}\sqrt{g(n_1)g(n_2)}}{(x+\kappa n_1)(x+\kappa n_2)} \tag{3.33}$$

From (3.7) and (3.33), we can see that the variance of $Z_n^{(x)}$ is equal to that of $Y_n^{(x)}$. Now, let $a > 1$. From Proposition 4(e), there exists an $n_o > 0$ such that for all $n \geq n_o$,

$$\frac{Var\{X_m\}}{m} \leq \frac{Var\{X_n\}}{n} \quad \text{for all } m < n. \tag{3.34}$$

If we assume x $\geq \alpha\kappa n_o$ and $n_2 > n_1 \geq \frac{x}{\alpha\kappa} \geq n_o$, then

$$\begin{aligned}
\frac{C_X(n_1,n_2)}{n_1} &= \frac{1}{2n_1}\left(Var\{X_{n_1}\} + Var\{X_{n_2}\} - Var\{X_{n_2-n_1}\}\right) \quad \text{(from Proposition 4(b))} \\
&= \frac{1}{2}\left(\frac{Var\{X_{n_1}\}}{n_1} + \frac{Var\{X_{n_2}\}}{n_2} + \frac{n_2-n_1}{n_1}\left(\frac{Var\{X_{n_2}\}}{n_2} - \frac{Var\{X_{n_2-n_1}\}}{n_2-n_1}\right)\right) \\
&\geq \frac{1}{2}\left(\frac{Var\{X_{n_1}\}}{n_1} + \frac{Var\{X_{n_2}\}}{n_2}\right) \quad \text{(from (3.34))} \\
&\geq \sqrt{\frac{Var\{X_{n_1}\}Var\{X_{n_2}\}}{n_1 n_2}} \quad \text{(since } \frac{Var\{X_n\}}{n} \geq 0\text{).}
\end{aligned}$$

This implies that

$$S\min\{n_1,n_2\}\sqrt{g(n_1)g(n_2)} = n_1\sqrt{\frac{Var\{X_{n_1}\}Var\{X_{n_2}\}}{n_1 n_2}} \leq C_X(n_1,n_2). \tag{3.35}$$

Therefore, from (3.5), (3.33), and (3.35), it follows that for $x \geq \alpha\kappa n_o$, $C_Y^{(x)}(n_1,n_2) \geq C_Z^{(x)}(n_1,n_2)$ for all $n_1, n_2 \in [\frac{x}{\alpha\kappa}, \frac{\alpha x}{\kappa}]$. Since we know $Var\{Y_n^{(x)}\} = Var\{Z_n^{(x)}\}$, we have $E\{(Y_{n_1}^{(x)} - Y_{n_2}^{(x)})^2\} \leq E\{(Z_{n_1}^{(x)} - Z_{n_2}^{(x)})^2\}$ for all $n_1, n_2 \in [\&, \frac{\alpha x}{\kappa}]$. Therefore, from Theorem 2,

$$P(\{\langle Y^{(x)}\rangle_{[\frac{x}{\alpha\kappa},\frac{\alpha x}{\kappa}]} > \sqrt{x}\}) \leq P(\{\langle Z^{(x)}\rangle_{[\frac{x}{\alpha\kappa},\frac{\alpha x}{\kappa}]} > \sqrt{x}\}) \quad \text{for all x} \geq \alpha\kappa n_o. \tag{3.36}$$

Now, we obtain an upper bound to $P(\{\langle Z^{(x)}\rangle_{[\frac{x}{\alpha\kappa},\frac{\alpha x}{\kappa}]} > \sqrt{x}\})$ as follows.

$$\begin{aligned}
P(\{\langle Z^{(x)}\rangle_{[\frac{x}{\alpha\kappa},\frac{\alpha x}{\kappa}]} > \sqrt{x}\}) &= P(\{Z_n^{(x)} > \sqrt{x} \text{ for any n} \in [\frac{x}{\alpha\kappa}, \frac{\alpha x}{\kappa}]\}) \\
&= P(\{\sqrt{Sg(n)}B_n > \text{x} + \kappa n \text{ for any n} \in [\frac{x}{\alpha\kappa}, \frac{\alpha x}{\kappa}]\}) \\
&\quad \text{(from the definition of } V_n \text{ and } Z_n^{(x)}\text{)} \\
&\leq P(\{\sqrt{Sg(\lceil\frac{\alpha x}{\kappa}\rceil)}B_n > \text{x} + \kappa n \text{ for any n} \in [\frac{x}{\alpha n}, \frac{\alpha x}{k}]\}) \\
&\quad \text{(since } g(n) \text{ is increasing on } [\frac{x}{\alpha\kappa}, \frac{\alpha x}{\kappa}] \text{ from (3.34))} \\
&\quad P(\{\sqrt{Sg(\lceil\frac{\alpha x}{\kappa}\rceil)}B_n > \text{x} + \kappa n \text{ for any n} \geq 0\}) \\
&\leq \mathrm{e}^{-\frac{2\kappa x}{Sg(\lceil\frac{\alpha x}{\kappa}\rceil)}} \quad \text{(from (3.32)).}
\end{aligned} \tag{3.37}$$

From (3.36) and (3.37), we have an asymptotic upper bound to $P(\{\langle Y^{(x)}\rangle_{[\frac{x}{\alpha\kappa},\frac{\alpha x}{\kappa}]} > \sqrt{x}\})$

$$P(\{\langle Y^{(x)}\rangle_{[\frac{x}{\alpha\kappa},\frac{\alpha x}{\kappa}]} > \sqrt{x}\}) \leq \mathrm{e}^{\;\frac{2\kappa x}{S}} \quad \text{for all x} \geq \alpha\kappa n_o. \tag{3.38}$$

On the other hand, from Proposition 4(d) and the fact that $g(n) \to 1$ as n $\to \infty$, we have

$$\begin{aligned}
\frac{2\kappa x}{S} - \frac{2\kappa x}{Sg(\lceil\frac{\alpha x}{\kappa}\rceil)} &= -\frac{2\kappa x\left(1 - \frac{Var\{X_{\lceil\frac{\alpha x}{\kappa}\rceil}\}}{S\lceil\frac{\alpha x}{\kappa}\rceil}\right)}{Sg(\lceil\frac{\alpha x}{\kappa}\rceil)} \quad \text{(from the definition of } g(t)\text{)} \\
&= -\frac{2\kappa x}{S^2\lceil\frac{\alpha x}{\kappa}\rceil}\frac{\lceil\frac{\alpha x}{\kappa}\rceil\left(S - \frac{Var\{X_{\lceil\frac{\alpha x}{\kappa}\rceil}\}}{\lceil\frac{\alpha x}{\kappa}\rceil}\right)}{g(\lceil\frac{\alpha x}{\kappa}\rceil)} \to -\frac{2\kappa^2 D}{\alpha S^2} \text{ as } x \to \infty.
\end{aligned} \tag{3.39}$$

Therefore, from Lemma 5, Theorem 8, (3.38) and (3.39), it follows that

$$\limsup_{x\to\infty} e^{\frac{2\kappa x}{S}} P(\{\langle X\rangle > x\}) = \limsup_{x\to\infty} e^{\frac{2\kappa x}{S}} P(\{\langle Y^{(x)}\rangle > \sqrt{x}\}) \leq e^{-\frac{2\kappa^2 D}{\alpha S^2}}$$

Since $\alpha > 1$ is arbitrary, finally we have limsup,,, $e^{\frac{2\kappa x}{S}} P(\{\langle X\rangle > x\}) \leq e^{-\frac{2\kappa^2 D}{S^2}}$. *Q.E.**D**.*

Theorem 9 gives us an exponential asymptotic upper bound ($e^{-\frac{2\kappa}{S}(x+\frac{\kappa D}{S})}$) to the tail probability $P(\{Q > x\}) = P(\{\langle X\rangle > x\})$. This also implies that Theorem 9 provides us with an upper bound $e^{-\frac{2\kappa^2 D}{S^2}}$ to the asymptotic constant C given in (1.1). However, in our previous research, it has been shown that a single exponential type of approximation (such as the asymptotic approximation and EB approximation) may not closely estimate the tail probability even for relatively large values of x. This is quite typical for traffic, that is correlated at multiple time scales, for which the tail probability converges to its asymptote slowly. Therefore, in spite of the theoretical significance of this asymptotic upper bound, it is expected to suffer from the same problem as other single exponential approximations for certain types of arrival traffic.

On the other hand, in our previous research [9, 10], the lower bound $\Psi\left(\frac{\sqrt{x}}{\langle\sigma_x\rangle}\right)$ has been found to be an accurate approximation to the tail probability and matches the tail probability curve even when it converges to its asymptote slowly. Remember that the lower bound $\Psi\left(\frac{\sqrt{x}}{\langle\sigma_x\rangle}\right)$ is a (standard Gaussian tail) function of $\frac{\sqrt{x}}{\langle\sigma_x\rangle}$, the maximum variance and the queue length. From the fact that the lower bound matches the shape of the tail probability quite well, we can infer that the maximum variance $\langle\sigma_x^2\rangle$, as a function of x, contains key information about the shape of the tail probability curve. Therefore, it would be an important result if we were to find an asymptotic upper bound in terms of the maximum variance (a:). In the following theorem we find such an asymptotic upper bound in terms of $\langle\sigma_x^2\rangle$ based on Theorem 9.

Theorem 10 Under conditions (C1) and (C2), $e^{-\frac{x}{2\langle\sigma_x^2\rangle}} \sim e^{-\frac{2\kappa}{S}(x+\frac{\kappa D}{S})}$ *as* $x \to \infty$. Therefore, with an additional condition (C3), $e^{-\frac{x}{2\langle\sigma_x^2\rangle}}$ *asymptotically* bounds $P(\{\langle X\rangle > x\})$.

Proof of Theorem *10* : From (3.6) and the definition of $\hat{n}_x$, we have (a;) $= \frac{x Var\{X_{\hat{n}_x}\}}{(x+\kappa\hat{n}_x)^2}$. Hence,

$$\frac{2\kappa x}{S} - \frac{x}{2\langle\sigma_x^2\rangle} = \frac{-4\kappa\frac{x}{\hat{n}_x}\hat{n}_x\left(S - \frac{Var\{X_{\hat{n}_x}\}}{\hat{n}_x}\right) - \frac{S\kappa^2}{\hat{n}_x}\left(\frac{x}{\kappa} - \hat{n}_x\right)^2}{2S\frac{Var\{X_{\hat{n}_x}\}}{\hat{n}_x}} \tag{3.40}$$

Since $\frac{x}{\hat{n}_x} \to \kappa$, $\frac{Var\{X_{\hat{n}_x}\}}{\hat{n}_x} \to S$, $(S - \frac{Var\{X_{\hat{n}_x}\}}{\hat{n}_x})\hat{n}_x \to D$, and $\frac{(\frac{x}{\kappa}-\hat{n}_x)^2}{\hat{n}_x} \to 0$ as $x \to \infty$ from Proposition 4 and Proposition 6, it follows from (3.40) that

$$\lim_{x\to\infty} \frac{2\kappa x}{S} - \frac{x}{2\langle\sigma_x^2\rangle} = -\frac{2\kappa^2 D}{S^2}.$$

Therefore, $\lim_{x\to\infty} e^{\frac{2\kappa x}{S}} e^{-\frac{x}{2\langle\sigma_x^2\rangle}} = e^{-\frac{2\kappa^2 D}{S^2}}$ *Q.E.D.*

As mentioned before, the lower bound $\Psi\left(\frac{\sqrt{x}}{\langle\sigma_x\rangle}\right)$ has been found to be an accurate approximation to the tail proability. However, from (3.17) and Theorem 10 it can easily be shown that

$$\Psi\left(\frac{\sqrt{x}}{\langle\sigma_x\rangle}\right) \sim \sqrt{\frac{S}{8\pi\kappa x}} e^{-\frac{2\kappa}{S}(x+\frac{\kappa D}{S})} \quad \text{as } x \to \infty, \tag{3.41}$$

which implies that the lower bound is not asymptotically exponential. Since the tail probability $P(\{Q > x\})$ is asymptotically exponential with great generality, the ratio of the lower bound to the tail probability asymptotically goes to 0; more precisely, there is a constant $c > 0$ such that $\frac{\Psi(\frac{\sqrt{x}}{\langle\sigma_x\rangle})}{P(\{Q>x\})} \sim \frac{c}{\sqrt{x}}$ as $x \to \infty$. Therefore, the lower bound may fail to accurately approximate the tail probability for very large values of x or very small tail probabilities.[†] Note, however that $e^{-\frac{x}{2\langle\sigma_x^2\rangle}}$ is in fact asymptotically exponential under conditions (C1) and (C2) (which are relatively weak absolute summability conditions). This fact, in conjunction with the observation that the maximum variance $\langle\sigma_x^2\rangle$ provides important information about the shape of the tail probability curve, suggests that $e^{-\frac{x}{2\langle\sigma_x^2\rangle}}$ will provide a good approximation to $P(\{Q > x\})$ even without requiring condition (C3). Our expectation will be experimentally validated in Chapter 4.

[†]Note that the term $\frac{c}{\sqrt{x}}$ vanishes much more slowly than the tail probability which vanishes exponentially. This is a reason that the lower bound usually approximates tail probabilities, which is not too small, fairly well.

4. Results and Discussion

In this chapter, we investigate the tightness of our asymptotic upper bounds via several numerical examples. Also, we illustrate the performance and applicability of our approximation $e^{-\frac{x}{2\langle\sigma_x^2\rangle}}$ even when condition (C3) does not hold. Since, in general, the exact tail probability $P(\{Q > x\})$ is not analytically obtainable we use simulation techniques to validate our results. In particular, we use the *Importance Sampling* simulation technique described in [7] to improve the reliability of the estimation. We have calculated 95% confidence intervals for each tail probability estimated via simulation by the method of batch mean [6]. However, to not unnecessarily clutter the figures, *we only show confidence intervals when, they are larger than* $\pm 20\%$ *of the estimated tail probability.*

For the importance sampling simulations, (pseudo) *regenerative cycles* [7] are defined to be the time period between successive time epochs. We define these time epochs to be the time at which the queue transitions from an empty state to a non-empty state. Generally, the accuracy of simulation via importance sampling improves as the number of regenerative cycles involved in the simulation increases [7]. Therefore, when $P(\{Q > 0\})$ is very small, even though this does not necessarily imply the rareness of the regenerative cycle, it is usually difficult to get enough number of regenerative cycles for the simulation. After extensive simulation studies, we found that reliable results even using importance sampling cannot usually be obtained (in a reasonable amount of time) when $P(\{Q > 0\})$ is less than 10^{-4}. Hence, for all experiments, we set the utilization $(\rho = \lambda/\mu)$ so that $P(\{Q > 0\})$ is greater than 10^{-4} (as shown in the numerical figures, we do, however, estimate significantly lower values of $P(\{Q > x\})$, for $x > 0$).

This chapter is composed of two parts: in Section 4.1, we test the performance of the asymptotic upper bounds derived in Theorems 9 and 10. We further apply the approximation $e^{-\frac{x}{2\langle\sigma_x^2\rangle}}$ in the case of Gaussian input processes that do not satisfy condition (C3). In Section 4.2 we apply this approximation to traffic source models for real-time applications such as voice and video, and illustrate its effectiveness for admission control.

4.1 Numerical Investigation for Gaussian Input Processes

Example 1 In this section we consider fluid queues fed by a Gaussian input process. By comparing the asymptotic upper bound to the exact tail probability estimated via simulation, we can investigate how tight our bounds are to the tail probability.

Conditions (C1) and (C2) are very weak conditions. Any (stationary) Gaussian process whose autocovariance function vanishes faster than $l^{-\epsilon}$ for any $\epsilon > 2$ (except, of course those processes for which $\sum_{l=-\infty}^{\infty} C_\xi(l) = 0$) satisfy these conditions. It is relatively more difficult to classify Gaussian processes that satisfy condition (C3). However any Gaussian process with a non-negative autocovariance function satisfies condition (C3). Therefore, in the first example we consider Gaussian input processes with nonnegative autocovariance functions that vanish exponentially.

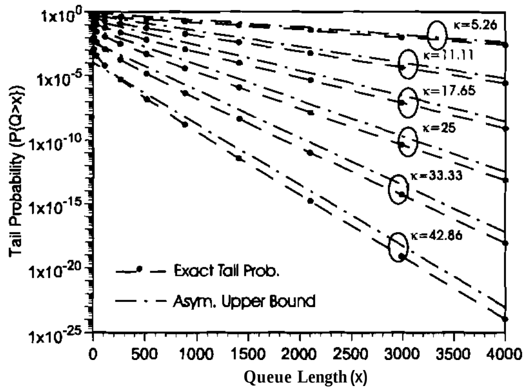

Figure 4.1: The exact tail probability and the asymptotic upper bound $e^{-\frac{2\kappa}{S}(x+\frac{\kappa D}{S})}$ for a Gaussian input process with autocovariance function $C_\xi(l) = 200 \times 0.95^{|l|}$.

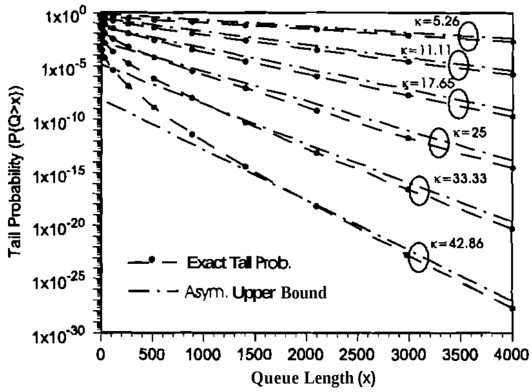

Figure 4.2: The exact tail probability and the asymptotic upper bound $e^{-\frac{2\kappa}{S}(x+\frac{\kappa D}{S})}$ for a Gaussian input process with autocovariance function $C_\xi(l) = 100 \times 0.9^{|l|} + 60 \times 0.98^{|l|}$.

In this example we will discuss Figures 4.1–4.6. For Figures 4.1–4.3 we focus on the asymptotic upper bound $e^{-\frac{2\kappa}{S}(x+\frac{\kappa D}{S})}$ obtained in Theorem 9. For the rest of the figures in this example, we focus on the asymptotic upper bound $e^{-\frac{x}{2(\sigma_x^2)}}$ obtained in Theorem 10.

We consider fluid queues fed by three different Gaussian input processes. The autocovariance functions of these Gaussian processes are given as $200 \times 0.95^{|l|}$, $100 \times 0.9^{|l|} + 60 \times 0.98^{|l|}$, and $104 \times 0.99^{|l|} + 64.14 \times 0.999^{|l|} + 31.86 \times 0.9999^{|l|}$. In Figure 4.1, we show the exact tail probability and asymptotic upper bound $e^{-\frac{2\kappa}{S}(x+\frac{\kappa D}{S})}$ for the Gaussian input with the autocovariance function $200 \times 0.95^{|l|}$ for six different values $(5.26, 11.11, 17.65, 25, 33.33, 42.86)$ of $\kappa = \mu - \bar{\lambda}$. As one can see in the figure, 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 $\frac{2\kappa}{S}$. Further note that the bound matches the simulation results quite well. This suggests that $e^{-\frac{2\kappa^2 D}{S^2}}$ is a good bound to the asymptotic constant. The tightness of the asymptotic upper bound is also demonstrated in Figure 4.2, which shows the same curves for the Gaussian input process with the autocovariance function $100 \times 0.9^{|l|} + 60 \times 0.98^{|l|}$ when $\kappa = 5.26, 11.11, 17.65, 25, 33.33, 42.86$. As in Figure 4.1, 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 Figure 4.2, 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 Figure 4.2 converges to its exponential asymptote slowly, while the tail probability in Figure 4.1 converges to its asymptote fairly fast and forms a nearly straight line. The autocovariance function of the Gaussian input used in the second experiment consists of two power terms with quite different decay rates. Therefore, the input is correlated at different time scales. A far more significant effect of this multiple time-scale correlation is demonstrated in Figure 4.3. In Figure 4.3, we show the exact tail probability, the asymptotic upper

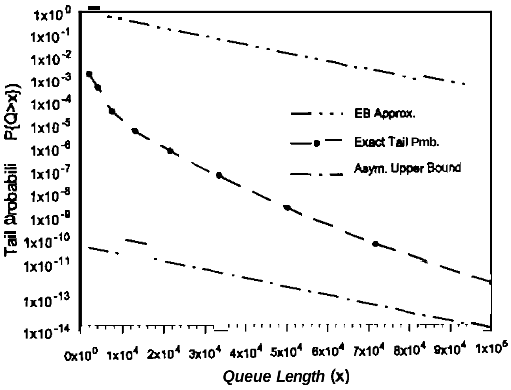

Figure 4.3: The exact tail probability, the asymptotic upper bound $e^{-\frac{2\kappa}{S}(x+\frac{\kappa D}{S})}$, and the EB approximation for a Gaussian input process with autocovariance function $C_\xi(l) = 104 \times 0.99^{|l|} + 64.14 \text{ x } 0.999^{|l|} + 31.86 \text{ x } 0.9999^{|l|}$ when $\kappa = 33.33$.

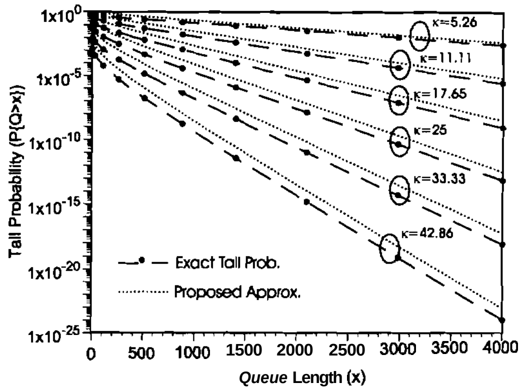

Figure 4.4: The exact tail probability and the asymptotic upper bound $e^{-\frac{x}{2\langle\sigma_x^2\rangle}}$ for a Gaussian input process with autocovariance function $C_\xi(l) = 200 \times 0.95^{|l|}$.

bound, and the EB approximation for the Gaussian input with autocovariance function $104 \text{ x } 0.99^{|l|} + 64.14 \times 0.999^{|l|} + 31.860.9999^{|l|}$ for $\kappa = 33.33$. Note that the autocovariance function is composed of three weighted power terms with very different decay rates. Consequently, the tail probability converges to its asymptote very slowly. Note that the true asymptote will lie below the asymptotic upper bound shown in Figure 4.3. Hence, any single exponential approximation including the asymptotic approximation, the EB approximation, and the asymptotic upper bound $e^{-\frac{2\kappa}{S}(x+\frac{\kappa D}{S})}$, provides a poor estimate of the tail probability even for values of x as large as 100,000.

Through a number of numerical examples (other than those just shown), we have found that the upper bound $e^{-\frac{2\kappa^2 D}{S^2}}$ to the asymptotic constant is usually very close to the asymptotic constant. Therefore, the asymptotic upper bound given by Theorem 9 approximates the tail probability for sufficiently large queue lengths. However, as shown above, single exponential approximations are fundamentally limited and may fail to accurately estimate the tail probability for small or even fairly large values of queue lengths.

On the bright side, since we now know that the asymptotic upper bound is accurate for large enough x, we expect that the bound $e^{-\frac{x}{2\langle\sigma_x^2\rangle}}$ given by Theorem 10 should also be accurate for large values of queue lengths (since they are asymptotically similar). Further, as discussed in Section 3.3, since the bound is expressed in terms of the maximum variance $\langle\sigma_x^2\rangle$, we can also expect it to match the shape of the tail probability curve (as in the case of the lower bound $\Psi\left(\frac{\sqrt{x}}{\langle\sigma_x\rangle}\right)$). To demonstrate the performance of this bound, in Figures 4.4, 4.5, and 4.6, we redo the experiments of Figures 4.1, 4.2, and 4.3, respectively. As expected, the bound $e^{-\frac{x}{2\langle\sigma_x^2\rangle}}$ is accurate over the entire range of queue lengths shown in all three figures. Also, the analysis matches the shape of the actual tail probability curve.

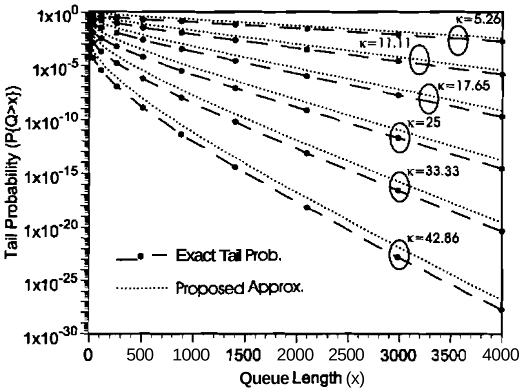

Figure 4.5: The exact tail probability and the asymptotic upper bound $e^{-\frac{x}{2\langle\sigma_x^2\rangle}}$ for a Gaussian input process with autocovariance function $C_\xi(l) = 100 \times 0.9^{|l|} + 60 \times 0.98^{|l|}$.

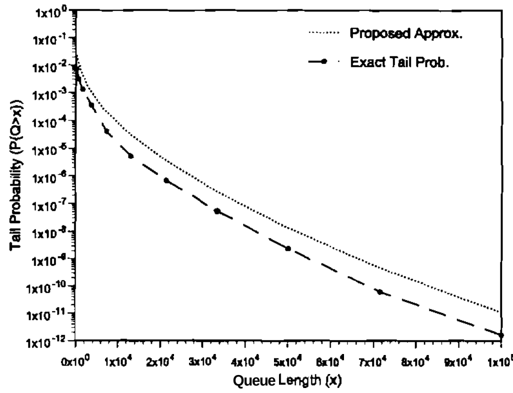

Figure 4.6: The exact tail probability and the asymptotic upper bound $e^{-\frac{x}{2\langle\sigma_x^2\rangle}}$ for a Gaussian input process with autocovariance function $C_\xi(l) = 104 \times 0.99^{|l|} + 64.14 \times 0.999^{|l|} + 31.86 \times 0.9999^{|l|}$ when $\kappa = 33.33$.

Example 2 Now, consider fluid queues fed by two different Gaussian processes whose autocovariance functions are given by $10 \times 0.95^{|l|} \cos\frac{\pi l}{18}$ and $10 \times 0.9^{|l|} \cos\frac{\pi l}{12} + 0.1 \times 0.99^{|l|}$. Neither of these autocovariance functions satisfy condition (C3). For this case we cannot guarantee that $e^{-\frac{x}{2\langle\sigma_x^2\rangle}}$ is an asymptotic upper bound. However, as mentioned earlier, we have found via extensive numerical experiments, that the lower bound $\Psi\left(\frac{\sqrt{x}}{\langle\sigma_x\rangle}\right)$ which is also a function of the maximum variance $\langle\sigma_x^2\rangle$, is accurate for Gaussian inputs that do not satisfy condition (C3) (see [10] for example). Therefore, we expect the analytical expression $e^{-\frac{x}{2\langle\sigma_x^2\rangle}}$ to accurately approximate the tail probability, which is in fact demonstrated in Figures 4.7 and 4.8.

In the following section, we use our analytical result $e^{-\frac{x}{2\langle\sigma_x^2\rangle}}$ to approximate the tail probability for more general (non-Gaussian) traffic source models.

4.2 Numerical Investigation for Voice and Video Traffic

In this section, we illustrate the performance of our analytical approximation $e^{-\frac{x}{2\langle\sigma_x^2\rangle}}$ by applying it to an ATM multiplexer serving voice and video traffic. Throughout this section we assume that the amount of fluid and time is measured in an abstract unit of "cell" and "slot," respectively. In our setting, a cell corresponds to a 53-bytes (48-bytes payload) ATM cell and a slot to a 10 msec interval.

Example 3 In this example, we study the queueing behavior of multiplexed voice source models. Each voice source is modeled by an On-Off *Discrete-time Markov Modulated Fluid* (DMMF) process. The state transition matrix and the input rate vector of the DMMF model are given as follows.

$$\text{State Transition Matrix}: \begin{bmatrix} 0.9833 & 0.01677 \\ 0.025 & 0.975 \end{bmatrix}$$

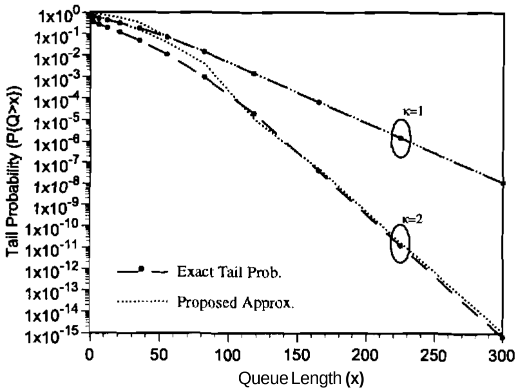

Figure 4.7: The exact tail probability and our analytical approximation ($e^{-\frac{x}{2(\sigma_x^2)}}$) for a Gaussian input process with autocovariance function $C_\xi(l) = 10 \times 0.95^{|l|} \cos\frac{\pi l}{18}$ and $\kappa = 1, 2$.

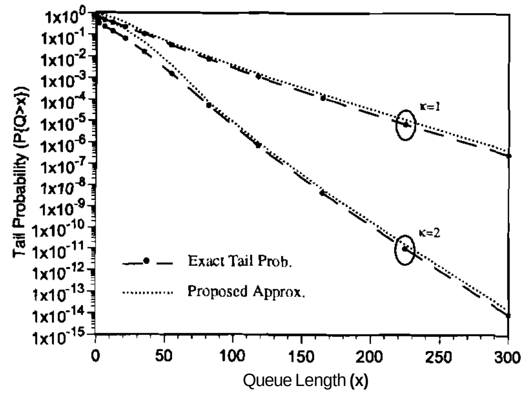

Figure 4.8: The exact tail probability and the analytical approximation ($e^{-\frac{x}{2(\sigma_x^2)}}$) for a Gaussian input process with autocovariance function $C_\xi(l) = 10 \times 0.9^{|l|} \cos\frac{\pi l}{12} + 0.1 \times 0.99^{-|l|}$ and $\kappa = 1, 2$.

$$\text{Input Rate Vector :} \quad \begin{bmatrix} 0 \\ 0.85 \end{bmatrix}$$

The DMMF voice traffic source model is obtained by discretizing the *Continuous-time Markov Modulated Fluid* model used in [28]. The service rate of the fluid server is set to 14672 cells/slot, which roughly corresponds to the capacity of an *OC-12* link (622.08 Mbps). The exact tail probability and our analytical approximation for 42500 and 42800 multiplexed DMMF sources are shown in Figure 4.9. Remember that the proposed technique uses only the first two moments of the aggregate traffic input to approximate the tail probability. As one can see in the figure, the approximation accurately matches the tail probability over the entire range of queue lengths shown in the figure.

Example 4 The exact tail probability and the proposed approximation for 105 and 107 simple JPEG-encoded video traffic source models are shown in Figure 4.10. This traffic source model is the superposition of a *i.i.d.* Gaussian process and three DMMF processes with very different state transition rates designed to capture the multiple time-scale correlation observed in the JPEG-encoded movie "Star Wars." The state transition matrices and the input rate vectors of three DMMF's and the mean and the variance of the *i.i.d.* Gaussian process are given as follows.

$$\text{State Transition Matrices :} \quad \begin{bmatrix} 0.99 & 0.01 \\ 0.01 & 0.99 \end{bmatrix} \quad \begin{bmatrix} 0.999 & 0.001 \\ 0.001 & 0.999 \end{bmatrix} \quad \begin{bmatrix} 0.9999 & 0.0001 \\ 0.0001 & 0.9999 \end{bmatrix}$$

$$\text{Input Rate Vectors :} \quad \begin{bmatrix} 0 \\ 45.516 \end{bmatrix} \quad \begin{bmatrix} 0 \\ 31.86 \end{bmatrix} \quad \begin{bmatrix} 0 \\ 18.204 \end{bmatrix}$$

Mean of *i.i.d.* Gaussian : 82.42

Variance of *i.i.d.* Gaussian : 8.6336

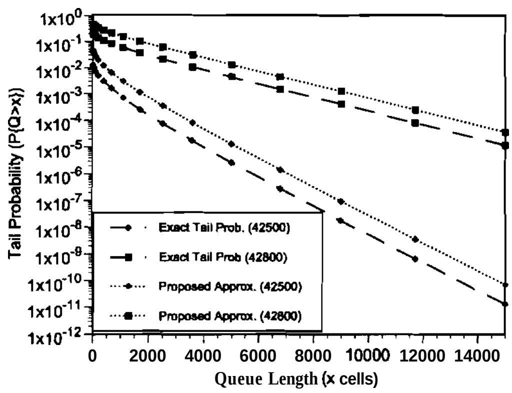

Figure 4.9: The exact tail probability and our analytical approximation for a fluid queue serving 42500 and 42800 voice traffic source models (DMMF). The service rate of the fluid server is set to 14672 cells/slot.

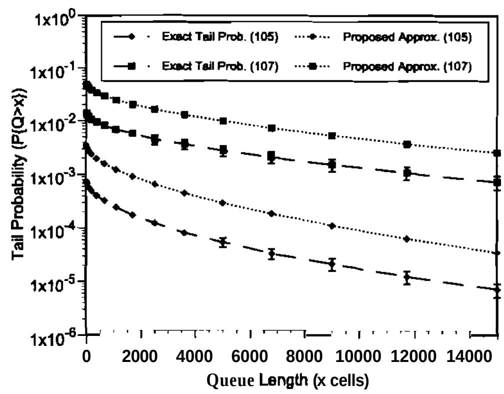

Figure 4.10: The exact tail probability and our analytical approximation for a fluid queue serving 105 and 107 JPEG-encoded video traffic source models. The service rate of the fluid server is set to 14672 cells/slot.

Again, we set the service rate of the fluid server to 14672 cells/slot.

As in the case of Example 4.9, the analysis matches the simulations quite closely, even though the number of traffic sources being multiplexed is significantly smaller. Also note that the approximation follows the shape of the exact tail probability quite well.

In general, the stochastic characteristics of a video traffic source change with the type of video application which the source represents. For instance, the video traffic source that mainly transmits movies is very likely to have different characteristics from that of the video source that does news programs. Further, the video coding schemes employed to reduce the required bandwidth can also significantly affect the stochastic characteristics of the video traffic generated. Therefore, modeling such diverse video traffic sources may not be an easy and efficient way of characterizing these sources. From this viewpoint, the traffic characterization based only on the first two moments (mean and autocovariance) have some advantage over the characterization based on explicit stochastic modeling, since the mean and autocovariance of a traffic source can be directly measured from the source. In most of the numerical examples provided so far, it has been illustrated that the first two moments contain very important information about the queueing behavior of the source and can be used to approximate the steady state queue length distribution accurately. However, it should be noted that in the previous examples, the first two moments of traffic sources have been analytically obtained from the source models. Hence, the question to ask is whether the moments measured directly from a source can be used to accurately capture its queueing behavior. In the next example, we demonstrate that from the mean and autocovariance measured from a real video source, the queue length distribution can be accurately computed.

Example 5 In this example, we use the frame size trace of the JPEG-encoded movie "Star Wars" to

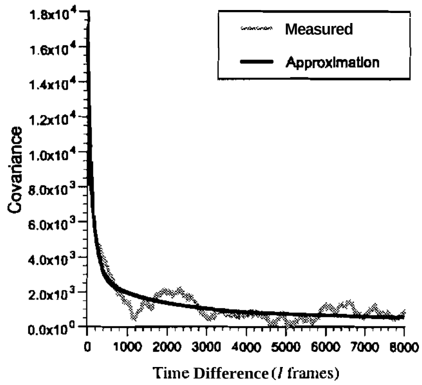

Figure 4.11: The autocovariance function measured from JPEG-encoded movie "Star Wars" and its approximation with the weighted sum of 4 exponential functions.

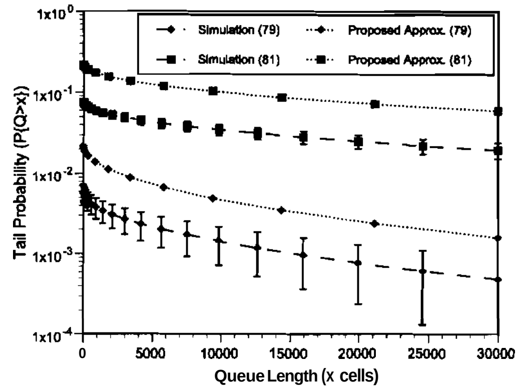

Figure 4.12: The simulation result and our analytical approximation for a multiplexer serving 79 and 81 JPEG-encoded movie "Star Wars" through an *OC-12* output link.

simulate real video sources, and experimentally obtain the tail probability $P(\{Q > x\})$ for these sources. Also, we use the mean and autocovariance function measured directly from the frame size trace to compute the approximation $e^{-\frac{x}{2(\sigma_x^2)}}$ and compare it to the tail probability obtained through simulation. In Figure 4.11, we show the autocovariance function measured directly from the trace. As one can see from the figure, the autocovariance function has quite an irregular shape for the time difference I larger than 1000, and hence cannot easily be expressed by a simple function of the time difference I. Therefore, using the least square method, we approximate the autocovariance function by the sum of 4 exponential terms which have very different decay rates, as shown in Figure 4.11. Using this approximated autocovariance function, we then compute the approximation $e^{-\frac{x}{2(\sigma_x^2)}}$ for the tail probability $P(\{Q > x\})$ for 79 and 81 of these sources, and compare them to the simulation results in Figure 4.12. Because importance sampling based simulation cannot be applied here (since we are using a real trace of JPEG-encoded video), the 95% confidence intervals displayed in the figure are quite large (especially for probabilities less than 10^{-3}. Nevertheless, as in the previous examples (where a stochastic model for JPEG-encoded video traffic is used for simulation, and the autocovariance function analytically obtained from the model is used to compute the approximation), the approximation follows the simulation results closely.

An important application of our analytical technique is for call admission control. We assume that a new call is admitted to an ATM multiplexer with buffer size B if the resulting tail probability $P(\{Q > x = B\}) \leq \varphi$, i.e. φ is the maximum tolerable tail probability for a call to be admitted.

Example 6 Consider an ATM multiplexer with an *OC-12* link and 20000 cell buffers, serving only JPEG-encoded video calls. We model this ATM multiplexer by a fluid queue fed by the video traffic source model used in Example 4. Again, we assume a 10 msec time-slot size and ATM cell size; therefore the server of the fluid queue can serve at most 14672 cell in each time slot. In Figure 4.13, we show the exact tail

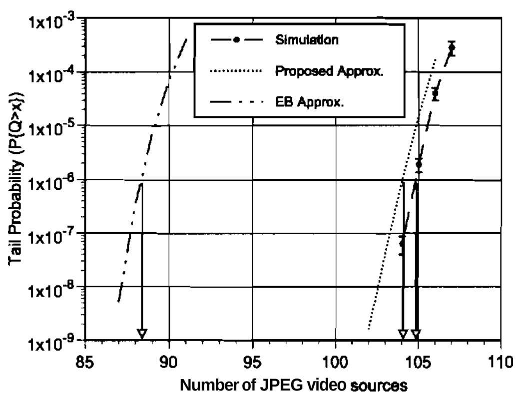

Figure 4.13: The exact tail probability, the EB approximation and our analytical approximation of the tail probability $P(\{Q > x))$ at x = 20000 cells versus the number of JPEG-encoded video traffic source models being multiplexed. The service rate (14672 cells/slot) of the fluid queue corresponds to the capacity of an *OC-12* link.

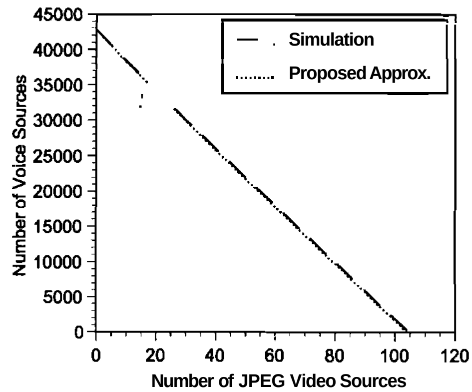

Figure 4.14: Admissible combinations of voice and JPEG-encoded video calls for an *OC-12* link with 20000 cell buffers, computed by simulation and our analytical approximation. The maximum tolerable tail probability (φ) is set to 10^{-6}.

probability, the EB approximation, and our analytical approximation at x = 20000 for different numbers of inputs. As one can see in the figure, assuming $\varphi = 10^{-6}$, our analytical approximation and the EB approximation estimate the maximum admissible number of video calls as 104 and 83, respectively. On the other hand, from the simulation results the maximum number of admissible JPEG-encoded video calls turns out to be 104. Therefore, the proposed approximation exactly estimates the maximum admissible number of calls (even though it slightly overestimates the tail probability), while the EB approximation underestimates the number by more than 15%.

So far, we have considered ATM multiplexers serving a large number of homogeneous inputs. In the next experiment, we demonstrate the applicability of the proposed technique for heterogeneous sources by determining the admissible region for voice and JPEG-video calls at an ATM multiplexer.

Example 7 In Figure 4.14, we show the admissible region for voice and JPEG-encoded video calls computed by simulation and our analytical approximation.

The maximum tolerable tail probability φ and the buffer size B are set to 10^{-6} and 20000 cells as in the previous example. Again, we assume an *OC-12* link for the transmission link of the ATM multiplexer. As one can see in the figure, the admissible regions computed by simulation and the proposed technique are so close that it is difficult to distinguish the boundaries of them. In fact, the proposed technique underestimates the maximum number of calls by less than 1% in terms of utilization.

5. Conclusion

In this report we have developed asymptotic upper bounds and approximations to the steady state tail probability $P(\{Q > x\})$ at an ATM multiplexer serving a large number of input processes. We model the ATM 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. After modeling the aggregate input process by a Gaussian process, we derived an exponential asymptotic upper bound $e^{-\frac{2\kappa}{S}(x+\frac{\kappa D}{S})}$ to the tail probability $P(\{Q > x\})$. This enabled us to find a good bound to the asymptotic constant. Further, we have derived another asymptotic upper bound $e^{-\frac{x}{2\langle\sigma_x^2\rangle}}$ in terms of the maximum variance $\langle\sigma_x^2\rangle$. Through extensive numerical experiments, we have found that $e^{-\frac{x}{2\langle\sigma_x^2\rangle}}$ accurately approximates the tail probability for very general types of traffic over a wide range of queue lengths x. We also used this analytical technique to accurately predict the tail probability for voice and JPEG-encoded types of video sources. Further, we were able to use our analytical technique for very efficient admission control.

In this report we provided results for 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 [11]. Although we have currently focusing on the analysis of ATM multiplexers, for future work we plan to concentrate on the analysis of intree-network statistical multiplexers.

LIST of REFERENCES

[1] J. Abate, G. L. Choudhury, and W. Whitt, "Exponential approximations for tail probabilities in queues, I: waiting times," Operations Research, 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. Comm., vol. 42, pp. 3150–3160, Dec. 1994.

[3] R. J. Adler, An Introduction to Continuity, Extrema, and Related Topics for General Gaussian Processes. Hayward, CA: Institute of Mathematical Statistics, 1990.

[4] S. M. Berman and N. Kono, "The Maximum of a Gaussian Process with Nonconstant Variance: A Sharp Bound for The Distribution Tail," Ann. Prob., vol. 17, no. 2, pp. 632–650, 1989.

[5] P. Billingsley, Probability and Measure. New York: John Wiley & Sons, 1979.

[6] P. Bratley, B. L. Fox, and L. E. Schrage, A Guide to Simulation. New York: Springer-Verlag, second ed., 1987.

[7] C.-S. Chang, P. Heidelberger, S. Juneja, and P. Shahabuddin, "Effective bandwidth and fast simulation of ATM intree networks," Performance Evaluation, vol. 20, pp. 45–65, 1994.

[8] C.-S. Chang and J. Thomas, "Effective Bandwidth in High-Speed Digital Networks," IEEE Journal On Selected Areas In Communications, vol. 13, pp. 1091–1114, August 1995.

[9] J. Choe and N. B. Shroff, "A Central Limit Theorem Based Approach to Analyze Queue Behavior in ATM Networks," in Proceedings of the 15th International *Teletraffic* Congress, 1997.

[10] J. Choe and N. B. Shroff, "A New Method to Determine the Queue Length Distribution at an ATM Multiplexer," in Proc. IEEE INFOCOM 97, pp. 550–557, 1997.

[11] J. Choe and N. B. Shroff, "A Novel approach for Determining the Queue Length Distribution at an ATM Multiplexer," tech. rep., Purdue University, W. Lafayette, IN 47907, 1997. Preprint.

[12] G. Choudhary, D. Lucantoni, and W. Whitt, "Squeezing the most out of ATM," IEEE Transactions on Communications, vol. 44, pp. 203–217, February 1996.

[13] A. Elwalid, D. Heyman, T. Lakshman, D. Mitra, and A. Weiss, "Fundamental bounds and approximations for ATM multiplexers with applications to video teleconferencing," IEEE Journal On Selected Areas In Communications, vol. 13, pp. 1004–1016, August 1995.

[14] A. Elwalid, D. Mitra, and R. Wentworth, "A New Approach for Allocating Buffers and Bandwidth to Heterogeneous Regulated Traffic in an ATM Node," IEEE Journal On Selected Areas In *Communications*, vol. 13, pp. 1115–1127, August 1995.

[15] A. I. Elwalid, "Markov Modulated Rate Processes for Modeling, Analysis and Control of Communication Networks," PhD thesis, Graduate School of Arts and Sciences, Columbia University, 1991.

[16] A. I. Elwalid and D. Mitra, "Effective Bandwidth of General Markovian Traffic Sources and Admission Control of High Speed Networks," *IEEE/ACM* Transactions on Networking, vol. 1, pp. 329–343, June 1993.

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

[18] I. I. Gikhman and A. V. Skorokhod, Introduction to the Theory of Random Processes. Philadelphia: W. B. Saunders, 1969.

[19] P. W. Glynn and W. Whitt, "Logarithmic asymptotics for steady-state tail probabilities in a single-server queue," Applied Probability, 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 Journal On Selected Areas In Communications, vol. 9, pp. 968–991, September 1991.

[21] F. Kelly, "Effective Bandwidth in Multiclass Queues," Queueing Systems, vol. 9, pp. 5–16, 1991.

[22] G. Kesidis, J. Walrand, and C.-S. Chang, "Effective Bandwidths for Multiclass Markov Fluids and other ATM sources," *IEEE/ACM* Transactions on Networking, vol. 1, pp. 424–428, August 1993.

[23] R. M. Loynes, "The Stability of a Queue with Non-independent Inter-arrival and Service Times," Proc. Cambridge Philos *Soc.*, vol. 58, pp. 497–520, 1962.

[24] M. Neuts, Matrix-geometric Solutions in Stochastic Models: an Algorithmic Approach. Johns Hopkins University Press, 1981.

[25] S. M. Ross, Stochastic Processes. New York: John Wiley & Son, 1983.

[26] G. Samorodnitsky, "Probability tails of Gaussian Extrema," Stochastic Process. Appl., vol. 38, no. 1, pp. 55–84, 1991.

[27] M. Schwartz, Broadband Integrated Networks. Upper Saddle River, New Jersey: Prentice Hall, 1996.

[28] N. Shroff and M. Schwartz, "Improved Loss Calculations at an ATM Multiplexer," in Proceedings of IEEE INFOCOM, vol. 2, pp. 561–568, March 1996.

[29] L. Takacs, Combinatorial Methods in the Theory of Stochastic Processes. New York: John Wiley, 1967.

[30] M. Talagrand, "Small tails for the supremum of a Gaussian process," Ann. *Inst.* Henri Poincare, vol. 24, pp. 307–315, 1988.