2-1-1998

A Central Limit Theorem Based Approach for Analyzing Queue Behavior 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

Follow this and additional works at: <http://docs.lib.purdue.edu/ecetr>

Choe, Jinwoo and Shroff, Ness B., "A Central Limit Theorem Based Approach for Analyzing Queue Behavior in High-Speed Networks" (1998). *ECE Technical Reports*. Paper 50.
<http://docs.lib.purdue.edu/ecetr/50>

This document has been made available through Purdue e-Pubs, a service of the Purdue University Libraries. Please contact epubs@purdue.edu for additional information.

A CENTRAL LIMIT THEOREM BASED APPROACH FOR ANALYZING QUEUE BEHAVIOR IN HIGH-SPEED NETWORKS

JINWOO CHOE
NESS B. SHROFF

TR-ECE 98-2
FEBRUARY 1998

SCHOOL OF ELECTRICAL
AND COMPUTER ENGINEERING
PURDUE UNIVERSITY
WEST LAFAYETTE, INDIANA 47907-1285

**A Central Limit Theorem Based Approach
for Analyzing Queue Behavior 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

	Page
Introduction	1
Fluid Queueing Model	4
Bounds for $\mathbb{P}(\{Q > x\})$	7
Lower Bound based on the Maximum Variance $\langle \sigma_x^2 \rangle$	7
Asymptotic Upper Bound	10
Interpretation of Time-Scaling $Y_n^{(x)}$	10
Single-Exponential Based Asymptotic Upper Bound	12
Numerical Examples and Discussions	13
Asymptotic Upper Bound Based on the Maximum Variance	18
Numerical Examples and Discussion	20
Applications for General Input Processes	22
General Gaussian Process	22
Applications to Voice and Video Traffic	24
Conclusion	30
Results from Extreme Value Theory	31
Preliminaries	32
Proofs for Main Results	36

LIST OF FIGURES

1.1	Statistical Multiplexing: K different network applications dynamically share a network link and a buffer.	
2.1	A fluid queue with an infinite buffer and a server. λ_t is the instantaneous rate of fluid fed into the system at time t , μ is the service rate, and Q_t is the amount of fluid in the queue at time t	
3.1	(a) $\mathbb{P}(\{Y_n^{(x)} > \sqrt{x}\})$ versus n . $\hat{n}_x$ turns out to be 1000. (b) Correlation factor between $Y_n^{(x)}$ and $Y_{\hat{n}_x}^{(x)}$ versus n . In this example, γ_n , the net input process is an <i>i.i.d.</i> Gaussian process with $\mathbb{E}\{\gamma_n\} = -0.1$ and $\text{Var}\{\gamma_n\} = 1$, and x is set to 100.	8
3.2	The exact tail probability and the lower bound for a Gaussian input process with autocovariance function $C_\lambda(l) = 200 \times 0.95^{ l }$	14
3.3	The exact tail probability and the lower bound for a Gaussian input process with autocovariance function $C_\lambda(l) = 100 \times 0.9^{ l } + 60 \times 0.98^{ l }$	14
3.4	The exact tail probability and the asymptotic upper bound for a Gaussian input process with autocovariance function $C_\lambda(l) = 200 \times 0.95^{ l }$	15
3.5	The exact tail probability and the asymptotic upper bound for a Gaussian input process with autocovariance function $C_\lambda(l) = 100 \times 0.9^{ l } + 60 \times 0.98^{ l }$	15
3.6	The exact tail probability, the EB approximation, the lower bound, and the asymptotic 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 $\kappa = 33.33$	16
3.7	The autocovariance functions ($C_1(l) = 25.641 \times 0.95^{ l }$ and $C_2(l) = 2.5063 \times 0.995^{ l }$) of two Gaussian input processes having the same sum $S = 1000$	17
3.8	The exact tail probability and the asymptotic upper bound for two Gaussian input processes with autocovariance functions $C_\lambda(l) = 25.641 \times 0.95^{ l }$ and $C_\lambda(l) = 2.5063 \times 0.995^{ l }$, κ is set to 5.	17
4.1	The difference $\log e^{-\frac{x}{2(\sigma_x^2)}} - \log \Psi\left(\sqrt{\frac{x}{(\sigma_x^2)}}\right)$ versus the MVA upper bound $\log e^{-\frac{x}{2(\sigma_x^2)}}$	19
4.2	The exact tail probability and the MVA upper bound for a Gaussian input process with autocovariance function $C_\lambda(l) = 200 \times 0.95^{ l }$	20
4.3	The exact tail probability and the MVA upper bound for a Gaussian input process with autocovariance function $C_\lambda(l) = 100 \times 0.9^{ l } + 60 \times 0.98^{ l }$	20
4.4	The exact tail probability, the lower bound, and the MVA upper bound for a Gaussian input process with autocovariance function $C_\lambda(l) = 104 \times 0.99^{ l } + 64.14 \times 0.999^{ l } + 31.86 \times 0.9999^{ l }$	21
5.1	The exact tail probability, the lower bound, and the MVA upper bound for a Gaussian input process with autocovariance function $C_\lambda(l) = 10 \times 0.95^{ l } \cos \frac{\pi l}{18}$ and $\kappa = 1, 2$	23
5.2	The exact tail probability, the lower bound, and the MVA upper bound for a Gaussian input process with autocovariance function $C_\lambda(l) = 10 \times 0.9^{ l } \cos \frac{\pi l}{12} + 0.1 \times 0.99^{ l }$ and $\kappa = 1, 2$	23

5.3	The exact tail probability, the lower bound and the MVA upper bound for a multiplexer serving 42500 and 42800 voice traffic sources. The output link capacity is set to 622.08 Mbps (<i>OC-12</i> line).	25
5.4	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 Mbps (<i>OC-3</i> line).	25
5.5	The autocovariance function measured from JPEG-encoded movie "Star Wars" and its approximation with the weighted sum of 3 exponential functions.	26
5.6	Simulation results, the lower bound, and the MVA upper bound for a multiplexer serving 79 and 81 JPEG-encoded movie "Star Wars" through an <i>OC-12</i> output link.	26
5.7	Admissible combinations of voice and JPEG-encoded video calls for an <i>OC-12</i> link with 20000 cell buffers, computed by simulation, the lower bound, and the MVA upper bound. The maximum tolerable tail probability (φ) is set to 10^{-6} .	27

Abstract

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

In this report we study $\mathbb{P}(\{Q > x\})$, the tail of the steady state queue length distribution at a high-speed multiplexer. The tail distribution $\mathbb{P}(\{Q > x\})$ is a fundamental measure of network congestion and thus important for the efficient design and control of these networks. In particular, we focus on the case when the aggregate traffic to the multiplexer can be characterized by a stationary Gaussian process. In our approach, a multiplexer is modeled by a fluid queue serving a large number of input processes. We propose a lower bound and two asymptotic upper bounds for $\mathbb{P}(\{Q > x\})$, and provide several numerical examples to illustrate the tightness of these bounds. We also use these bounds to study important properties of the tail probability. Further, we apply these bounds for a large number of 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.

1. 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 extremely important in such networks, since it increases network efficiency by allowing a large number of applications to share network resources, as shown in Figure 1.1. However, when these resources (e.g. buffer space and link capacity) 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 report is $\mathbb{P}(\{Q > x\})$, the tail of the steady state buffer occupancy (queue length) distribution at a multiplexer.

To analyze the performance of statistical multiplexing and to estimate the possibility of excessive network congestion that multiplexing may cause, a large number of queueing models have been studied. In particular, the rich theory of Markov processes has been found to be very useful for studying queueing behavior when arrival processes are “bursty” (correlated in time), as is typical in ATM networks. This is because many types of bursty network traffic can be modeled as Batch Markov Arrival Processes (BMAP) or as Markov Modulated Fluid (MMF) processes, and the resulting queueing models can be exactly analyzed [8, 22, 28, 39]. However, since a large number of heterogeneous network applications are expected to be multiplexed – e.g. commercial ATM switches already support 622 Mbps link speeds and gigabit-per-second switches are expected to appear soon – the exact analysis of the corresponding queueing system becomes increasingly difficult. For example, when Markovian queueing models are used to analyze queue behavior for a large number of multiplexed sources, one quickly runs into classical computational infeasibility problems due to the large dimension of the system [28, 46]. To address this problem, in this report, we will develop analytical techniques to determine $\mathbb{P}(\{Q > x\})$ for infinite buffer fluid queues serving a large number of arrival processes. Our approach is based on *Extreme Value Theory*, and will result in the development of a lower bound and asymptotic upper bounds for the tail probability. An important facet of our bounds is that they can be expressed in a simple elegant form that is easily computable, and thus they have both theoretical and practical value. We will study asymptotic properties of these bounds and will validate their accuracy via a thorough experimental study. Also, using these lower and itsymptotic upper bounds, we will study various aspects of the behavior of the tail probability. Before we describe the details of our approach, we first overview related work on $\mathbb{P}(\{Q > x\})$ and relate it to our own contribution in this report.

In the literature, the behavior of $\mathbb{P}(\{Q > x\})$ has been studied via various approaches, and a number of approximation techniques have been developed. The theoretical results which motivate appropriate approximations for $\mathbb{P}(\{Q > x\})$ can largely be classified into the following three categories.

1. Inequality category ($\lesseqgtr$)
2. Similarity category ($\sim$)
3. Log-similarity category ($\overset{\log}{\sim}$)

The first category ($\lesseqgtr$ -category) comprises all kinds of bounds for $\mathbb{P}(\{Q > x\})$. Once an upper or lower bound for the tail probability $\mathbb{P}(\{Q > x\})$ is found, it can be used to approximate the tail probability if it is tight. When both tight lower and upper bounds are available, they can provide us with a narrow range of values that encapsulate $\mathbb{P}(\{Q > x\})$. Therefore, the results in the first category are very useful when an admission control type of decision has to be made using the tail probability. However, in general,

Figure 1.1: Statistical Multiplexing: K different network applications dynamically share a network link and a buffer.

tight bounds for $\mathbb{P}(\{Q > x\})$ are difficult to obtain. Further, since the theoretical approach to derive a bound greatly depends on the class of queues being considered, a bound for a specific class of queueing systems cannot easily be extended to other classes of queues. For this reason, in this category there exist a relatively small number of results, each of which is obtained for specific classes of queueing models: for example, see [49] for continuous-time fluid queues with input processes having density and [23] for queues with Markovian arrival processes. In this report we will provide a lower bound and two (asymptotic) upper bounds for the tail probability for queueing systems serving a large number of input processes.

The second category ($\sim$ -category) includes all the asymptotic properties of $\mathbb{P}(\{Q > x\})$ which can be expressed by the similarity relation. Two functions $f(x)$ and $g(x)$ are often defined to be asymptotically similar ($f(x) \sim g(x)$) if $\lim_{x \rightarrow \infty} g(x)/f(x) = 1$. If we know that $\mathbb{P}(\{Q > x\})$ is asymptotically similar to some function $q(x)$, the function $q(x)$ can then be used to approximate the tail probability for large values of x . The advantage of this kind of approximation is that its (logarithmic) error $\log q(x) - \log \mathbb{P}(\{Q > x\})$ is guaranteed to vanish as $x \rightarrow \infty$, and hence, bounded over all values of x (of course, as long as $\mathbb{P}(\{Q > x\}), q(x) > 0$). One of the most important results in this category is the exponential similarity of the tail: for a very general class of queueing models, it has been shown that the tail probability is asymptotically exponential (for example, see [1, 2, 4, 9, 28, 48]), i.e.,

$$\mathbb{P}(\{Q > x\}) \sim Ce^{-\eta x}. \quad (1.1)$$

Here η is a positive constant called the asymptotic decay rate, and C is a positive constant called the asymptotic constant. As a result, the asymptote $Ce^{-\eta x}$ may be used 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 when a large number of arrival processes are multiplexed. However, an exact solution for C can only be determined for a limited class of queueing systems. Furthermore, even for this limited class of queueing systems, it is usually computationally problematic 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)

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

This approximation is the well known *Effective Bandwidth* (EB) approximation, which has been suggested for use in admission control [15, 27, 31, 34, 35]. 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 [19, 46]. Therefore, there is renewed interest in the asymptotic approximation, and methods have been

developed, to approximate the asymptotic constant C for special cases [4, 25, 26]. In this report we will develop a tight upper bound for the asymptotic constant C for a fairly large class of queueing systems fed by Gaussian input processes.

The third category is characterized by the log-similarity relation. Two functions $f(x)$ and $g(x)$ are defined to be asymptotically log-similar ($f(x) \stackrel{\log}{\sim} g(x)$) if $\log f(x)$ and $\log g(x)$ are asymptotically similar. The results based on Large Deviations techniques (see [21] for more about large deviations techniques) inherently belong to the third category. Since these large deviations techniques have been developed on very general mathematical settings, their applicability is remarkable, and asymptotic properties of the tail probability can be derived in the form of the log-similarity for diverse queueing systems. For instance, in [30], it has been shown that the relation

$$\mathbb{P}(\{Q > x\}) \stackrel{\log}{\sim} e^{-\eta x}, \quad (1.3)$$

holds for some constant η for several different queueing systems. Note that (1.3) provides another form of theoretical support for the EB approximation; albeit much weaker than that provided by (1.1). In [24], log-similarity has been extended to other classes of queueing systems such as queues fed by self-similar inputs. In [12], the asymptotic analysis of statistical multiplexing gain has been addressed by sending the size of a queueing system to infinity (instead of sending the queue length to infinity, which has been the usual direction in taking the limit). However, the great generality of the results in this category come at a cost: they are usually not as informative as the results that belong to the other categories. For instance, (1.3) does not imply (1.1), while (1.1) does (1.3). In fact, it is not difficult to see that there are an infinite number of functions such as $e^{-\eta x + \sqrt{x}}$ and $x^{10} e^{-\eta x}$, which are significantly different from $e^{-\eta x}$ and can replace $e^{-\eta x}$ in (1.3) to result in another valid log-similar relation. Due to the intrinsically poor “resolution” of the log-similarity relation, proposed approximations for $\mathbb{P}(\{Q > x\})$ can only be weakly supported by large deviations theory, and must be validated by extensive experimentation. For example, the lower bound that we introduce in this report has been shown to have a log-similar relationship to the exact tail probability $\mathbb{P}(\{Q > x\})$, but to validate its accuracy we need to perform an extensive and systematic simulation study. However, in turn, large deviation results have been found to be very useful in the development of Importance Sampling based simulation techniques (e.g. see [14, 44] and the references therein). In this report, we apply these simulation techniques to effectively capture rare events and significantly improve the reliability of our numerical studies.

The report is organized as follows. In Chapter 2, we briefly introduce the fluid queueing model of a high-speed multiplexer and provide some useful definitions. In Chapter 3, we introduce a simple lower bound and our first asymptotic upper bound for the tail probability. In Chapter 4, we develop our second asymptotic upper bound for the tail probability. In Chapter 5, we further weaken some of the assumptions made in earlier chapters, and use our bounds as approximations to the tail probability for more general fluid queues. Finally in Chapter 6, we bring the report to its conclusion.

2. Fluid Queueing Model

We model a high-speed statistical multiplexer by an infinite buffer fluid queue shown in Figure 2.1. The fluid queue consists of 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 , is closely related to 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 ($T = \{-\infty, \dots, -1, 0, 1, \dots, \infty\}$). In this report, we only consider discrete-time fluid queues, although equivalent results can also be obtained for continuous-time. Interested readers can find the corresponding results for the continuous-time case in [18].

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:

$$Q_n = (Q_{n-1} + \gamma_n)^+, \quad (2.1)$$

where $\gamma_n := \lambda_n - \mu$ is the net amount of fluid input at time n and $(x)^+ := \max\{0, x\}$. In [38], 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 (2.1) 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 under the same assumption that the distribution of $\sup_{n < 0} \sum_{m=n}^{-1} \gamma_m$ coincides with the steady state queue length distribution. Therefore, if we define a stochastic process X_n

$$X_n := \sum_{m=1}^n \gamma_{-m}, \quad (2.2)$$

then the supremum distribution of X_n is, in fact, the steady state queue distribution. In other words,

$$\mathbb{P}(\{Q > x\}) = \mathbb{P}\left(\left\{\sup_{n \geq 0} X_n > x\right\}\right). \quad (2.3)$$

This relation, which originally comes from [38, 51], has played a key role in obtaining a number of important results on the steady state queue length (or waiting time) distribution.

Throughout this report, we focus on the cases for which the aggregate arrival process can be effectively characterized by a stationary Gaussian process. Such queues have recently received some attention [4, 5, 16, 17, 36] because of two main reasons. Firstly, stationary Gaussian processes have several appealing properties. For example, stationary Gaussian processes are closed under superposition (assuming independence between superposed processes), and any stationary Gaussian process can completely be specified by its mean and autocovariance. Therefore, unlike the case of Markovian arrival 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. The other reason for Gaussian modeling, is that 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 precisely modeled by a Gaussian process, by appealing to the Central Limit Theorem, the multiplexer serving a large number of these application can be modeled and analyzed as a fluid queue with a stationary Gaussian input process.

Figure 2.1: A fluid queue with an infinite buffer and a server. λ_t is the instantaneous rate of fluid fed into the system at time t , μ is the service rate, and Q_t is the amount of fluid in the queue at time t .

Important Notations and Definitions

We now set the stage for our study of $\mathbb{P}(\{Q > x\})$, the tail of the steady state queue length distribution. Let $C_\gamma(l)$ denote the autocovariance function of the net input process $\gamma_n = \lambda_n - \mu$ (since the service rate of the fluid server is fixed to a constant μ in our case, $C_\gamma(l)$ is the same as $C_\lambda(l)$, the autocovariance function of the input process). We further define two important parameters S and D , that will be used extensively in our analysis.

$$S := \sum_{l=-\infty}^{\infty} C_\gamma(l) \quad \text{and} \quad D := 2 \sum_{l=0}^{\infty} l C_\gamma(l).$$

As motivated by the earlier discussion, we assume that γ_n , the net input process, is characterized by a stationary Gaussian process. Then, it is easy to see from the definition of X_n in (2.2), that it also is 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

$$\begin{aligned} \mathbb{E}\{X_n\} &= -\kappa n, \quad \text{and} \\ C_X(n_1, n_2) &:= \mathbb{E}\{(X_{n_1} + \kappa n_1)(X_{n_2} + \kappa n_2)\} = \sum_{l_1=1}^{n_1} \sum_{l_2=1}^{n_2} C_\gamma(l_2 - l_1). \end{aligned}$$

By a change of variables $m = m_2 - m_1$, the variance of X_n can be expressed as a weighted sum of $C_\gamma(l)$, i.e.,

$$\text{Var}\{X_n\} = \sum_{l_1=1}^n \sum_{l_2=1}^n C_\gamma(l_2 - l_1) = n C_\gamma(0) + 2 \sum_{l=1}^{n-1} (n-l) C_\gamma(l). \quad (2.4)$$

For notational simplicity, for each $x > 0$ we define a new stochastic process $Y_n^{(x)}$ as

$$Y_n^{(x)} := \frac{\sqrt{x}(X_n + \kappa n)}{x + \kappa n} \quad (2.5)$$

It then directly follows that

$$\text{for any } x > 0 \text{ and any } n \in \{0, 1, 2, \dots\}, \quad X_n > x \quad \text{if and only if} \quad Y_n^{(x)} > \sqrt{x}. \quad (2.6)$$

Therefore, we have

$$\mathbb{P}(\{Q > x\}) = \mathbb{P}\left(\left\{\sup_{n \geq 0} X_n > x\right\}\right) = \mathbb{P}\left(\left\{\sup_{n \geq 0} Y_n^{(x)} > \sqrt{x}\right\}\right). \quad (2.7)$$

Note that for each x , $Y_n^{(x)}$ is a centered Gaussian process, and its autocovariance function $C_{Y^{(x)}}$, in terms of C_X , is given by

$$C_{Y^{(x)}}(n_1, n_2) := \mathbb{E}\{Y_{n_1}^{(x)} Y_{n_2}^{(x)}\} = \frac{x C_X(n_1, n_2)}{(x + \kappa n_1)(x + \kappa n_2)}. \quad (2.8)$$

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

$$\sigma_{x,n}^2 = \frac{x \text{Var}\{X_n\}}{(x + \kappa n)^2} = \frac{x \left(n C_\gamma(0) + 2 \sum_{l=1}^{n-1} (n-l) C_\gamma(l) \right)}{(x + \kappa n)^2}. \quad (2.9)$$

For notational simplicity, we 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 . For example, $\langle \sigma_x^2 \rangle$ represents the supremum of $\sigma_{x,n}^2 = \text{Var}\{Y_n^{(x)}\}$ over $n \in \{0, 1, 2, \dots\}$ (the index omitted in $\langle \cdot \rangle$), and $\langle Y^{(x)} \rangle_{[a,b]}$ represents the supremum of $Y_n^{(x)}$ over $n \in [a, b]$. Also, unless explicitly mentioned otherwise, we assume that our net input process γ_n is a stationary Gaussian process.

We now list three useful conditions on $C_\gamma(l)$. The different theoretical results that we will derive in this report; will depend on one or more of these conditions.

- (C1) $C_\gamma(l)$ is absolutely summable and $\sum_{l=-\infty}^{\infty} C_\gamma(l) > 0$.
- (C2) $l C_\gamma(l)$ is absolutely summable.
- (C3) $\sum_{l=1}^m l C_\gamma(l) + \sum_{l=m+1}^{\infty} m C_\gamma(l) > 0$ for all $m = 1, 2, \dots$ and $\sum_{l=1}^{\infty} l C_\gamma(l) > 0$.

3. Bounds for $\mathbb{P}(\{Q > x\})$

In this chapter, we introduce two bounds for $\mathbb{P}(\{Q > x\})$ in the case of fluid queues driven by stationary Gaussian net input processes, and investigate their tightness through numerical examples. Also, we discuss the advantages and disadvantages of using these bounds as approximations to the tail probability.

3.1 Lower Bound based on the Maximum Variance $\langle \sigma_x^2 \rangle$

For a general (including non-Gaussian) stationary ergodic net input process γ_n , it can be shown that

$$\mathbb{P}(\{X_n > x\}) \rightarrow 0, \quad \text{as } n \rightarrow \infty. *$$

Therefore, there must exist a finite value of $n = \hat{n}_x$ at which the function $\mathbb{P}(\{X_n > x\})$ attains its maximum, i.e., $\sup_{n \geq 0} \mathbb{P}(\{X_n > x\})^\dagger$. From (2.3), we get the following simple lower bound for $\mathbb{P}(\{Q > x\})$.

$$\mathbb{P}(\{Q > x\}) = \mathbb{P}(\{X > x\}) \geq \sup_{n \geq 0} \mathbb{P}(\{X_n > x\}) = \mathbb{P}(\{X_{\hat{n}_x} > x\}). \quad (3.1)$$

At first 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 $\{0, 1, 2, \dots\}$ made of infinite elements. However, the lower bound is expected to be tight under certain circumstances (for example, see the heuristic explanation provided in [24]), and has been used to study the steady state queue behavior of different queueing systems [5, 12, 24, 41]. For example, it can be shown that for certain types of arrival processes the lower bound given by $\mathbb{P}(\{X_{\hat{n}_x} > x\})$ satisfies the large deviation type of asymptotic log-similarity

$$\mathbb{P}(\{X_{\hat{n}_x} > x\}) \stackrel{\log}{\sim} e^{-\frac{2\kappa x}{\langle \sigma_x^2 \rangle}} \stackrel{\log}{\sim} \mathbb{P}(\{Q > x\}). \quad (3.2)$$

The above relation can be easily obtained for general classes of input processes from large the deviation work in the literature (e.g., by a minor modification of the proof in [30, Theorem 2]). An explicit proof of (3.2) for Gaussian input processes can also be found in [16].

However, as mentioned in the introduction, log-similarity provides only weak support, and approximations based on it may yield significant errors. To provide further support in using the lower bound as a good approximation to the tail probability, we focus on the case when the net input process γ_n is stationary and Gaussian. When the input process is Gaussian, the time instant $\hat{n}_x$ at which $\mathbb{P}(\{X_n > x\})$ achieves its maximum value is also the time instant at which the process $Y_n^{(x)}$ (from (2.5)) attains its maximum variance. Further, from (2.6) and (3.1), we can rewrite (3.1) in terms of $Y_n^{(x)}$ as

$$\mathbb{P}(\{Q > x\}) \geq \mathbb{P}(\{X_{\hat{n}_x} > x\}) = \mathbb{P}(\{Y_{\hat{n}_x} > \sqrt{x}\}) = \Psi\left(\sqrt{\frac{x}{\langle \sigma_x^2 \rangle}}\right), \quad (3.3)$$

where $\langle \sigma_x^2 \rangle := \sup_{n \geq 0} \text{Var}\{Y_n^{(x)}\}$ and $\Psi(z) := \frac{1}{\sqrt{2\pi}} \int_z^\infty e^{-\frac{y^2}{2}} dy$ is the tail function of the standard Gaussian distribution. Remember that $Y_n^{(x)}$ is a zero-mean Gaussian process. This is important since in the Extreme Value *Theory* for Gaussian processes the maximum variance of a centered Gaussian process ζ_t (with non-constant variance) has been frequently emphasized as a very important factor in studying the supremum distribution of ζ_t [6, 7, 11, 45, 52] (e.g., see Theorem A.1 in Appendix A). The local behavior of ζ_t

*Note that since γ_n is ergodic, this implies that $\frac{1}{n} X_n = \frac{1}{n} \sum_{m=1}^n \gamma_{-m} \rightarrow -\kappa < 0$ as $n \rightarrow \infty$ almost surely. Therefore, $\mathbb{P}(\{X_n > x\}) \rightarrow 0$ as $n \rightarrow \infty$.

[†] $\mathbb{P}(\{X_n > x\})$ may attain its maximum at several different indices. In this case, $\hat{n}_x$ can be arbitrarily chosen among these indices, and our discussion is valid for any choice of $\hat{n}_x$.

Figure 3.1: (a) $\mathbb{P}(\{Y_n^{(x)} > \sqrt{x}\})$ versus n . $\hat{n}_x$ turns out to be 1000. (b) Correlation factor between $Y_n^{(x)}$ and $Y_{\hat{n}_x}^{(x)}$ versus n . In this example, γ_n , the net input process is an *i.i.d.* Gaussian process with $\mathbb{E}\{\gamma_n\} = -0.1$ and $\text{Var}\{\gamma_n\} = 1$, and x is set to 100.

around the index t_o where the maximum variance is achieved, **has** been found to essentially determine the supremum distribution of ζ_t (see [7, Section 5.3–5.4]). Therefore, one may expect that $\mathbb{P}(\{\zeta_{t_o} > x\})$ and $\mathbb{P}(\{\langle \zeta \rangle > x\})$ are not very different from each other, and in fact, the former turns out to be a fairly accurate estimate of the latter, for moderately large values of x [7, page 5]. These general observations made in the Extreme Value Theory literature suggest that our lower bound given by $\mathbb{P}(\{Y_{\hat{n}_x} > \sqrt{x}\})$ should accurately approximate $\mathbb{P}(\{Q > x\}) = \mathbb{P}(\{\langle Y^{(x)} \rangle > \sqrt{x}\})$. Also note that since $\mathbb{P}(\{Y_{\hat{n}_x} > \sqrt{x}\})$ can be calculated by evaluating the tail of the standard Gaussian distribution, the lower bound based approximation is computationally very simple. We now provide the following simple example to further illustrate the accuracy of the lower bound approximation.

Example 1 Let γ_n be an *i.i.d.* Gaussian process with $\mathbb{E}\{\gamma_n\} = -0.1$ and $\text{Var}\{\gamma_n\} = 1$, and set $x = 100$. One possible way of constructing γ_n is to define it as $\gamma_n := B_n - B_{n-1} - 0.1$, where B_t is the standard Brownian motion (Wiener) process. Therefore, the corresponding $Y_n^{(x)}$ and its autocovariance function can be expressed as

$$Y_n^{(x)} = \frac{\sqrt{x}B_n}{x + \kappa n}, \quad \text{and}$$

$$C_{Y^{(x)}}(n_1, n_2) = \frac{x \min\{n_1, n_2\}}{(100 + 0.1n_1)(100 + 0.1n_2)}, \quad \text{respectively.}$$

In Figure 3.1(a), we plot $\mathbb{P}(\{Y_n^{(x)} > \sqrt{x}\})$ over the interval (0,3000]. As one can see in the figure, the graph forms a sharp peak and attains its maximum at $\hat{n}_x = 1000$. In Figure 3.1(b), we show the correlation coefficient between $Y_n^{(x)}$ and $Y_{\hat{n}_x}^{(x)}$ by plotting $\frac{\text{Cov}\{Y_n^{(x)}, Y_{\hat{n}_x}^{(x)}\}}{\sigma_{x, \hat{n}_x} \sigma_{x, n}}$. In this figure, it can be observed that $Y_n^{(x)}$ is strongly correlated to $Y_{\hat{n}_x}^{(x)}$ for values of n close to $\hat{n}_x = 1000$. Since $\mathbb{P}(\{Y_n^{(x)} > \sqrt{x}\})$ is very small when n is far from $\hat{n}_x$, and since $Y_n^{(x)}$ is strongly correlated to $Y_{\hat{n}_x}^{(x)}$ for n close to $\hat{n}_x$, the probability $\mathbb{P}(\{\langle Y^{(x)} \rangle > \sqrt{x} \text{ and } Y_{\hat{n}_x}^{(x)} \leq \sqrt{x}\})$ should be small, and therefore, $\mathbb{P}(\{\langle Y^{(x)} \rangle > \sqrt{x}\})$ will be dominated by $\mathbb{P}(\{Y_{\hat{n}_x}^{(x)} > \sqrt{x}\})$. Even though Example 1 is for the case when γ_n is an *i.i.d.* Gaussian

process, similar observations can also be made in the limit (as $x \rightarrow \infty$), when γ_n belongs to more general classes of Gaussian processes (as will be discussed in Section 3.2).

We will now derive an asymptotic result that demonstrates the importance of $\hat{n}_x$, the value of n corresponding to the maximum variance of a stochastic process. We first restate Proposition B.2 derived in the appendix, that shows that the time index at which $\sigma_{x,n}^2$ (or equivalently $\mathbb{P}(\{X_n > x\})$) attains its maximum, is asymptotically a linear function of x .

Proposition B.2 *Let A_x be the index at which $\sigma_{x,n}^2$ attains its maximum (a). Then, under condition (C1),*

$$\hat{n}_x \sim \frac{x}{\kappa} \quad \text{as } x \rightarrow \infty. \quad (3.4)$$

Further, under conditions (C1) and (C2), and for all $\epsilon > 0$,

$$\lim_{x \rightarrow \infty} \frac{\hat{n}_x - \frac{x}{\kappa}}{x^\epsilon} = 0.$$

Proof of Proposition B.2 : See Appendix B.

Now, when the input process γ_n satisfies condition (C1), we introduce the following theoretical result which illustrates the importance of the local behavior of $Y_n^{(x)}$ around $\hat{n}_x$ in estimating $\mathbb{P}(\{X\} > x)$. This theorem will also be very important in deriving an asymptotic upper bound to $\mathbb{P}(\{Q > x\})$.

Theorem 3.1 *Under condition (C1), for any $a > 1$,*

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

Proof of Theorem 3.1 : See Appendix C.

From (3.41), note that for any $a > 1$, the interval $[\frac{\hat{n}_x}{\alpha}, \alpha \hat{n}_x]$ (and hence $\hat{n}_x$ itself) will eventually be contained in $[\frac{x}{2\alpha\kappa}, \frac{2\alpha x}{\kappa}]$ as x increases. Therefore, Theorem 3.1 implies that for any $a > 1$,

$$\mathbb{P}(\{\langle Y^{(x)} \rangle_{[\frac{\hat{n}_x}{\alpha}, \alpha \hat{n}_x]} > \sqrt{x}\} | \{Y^{(x)}\} > \sqrt{x}) = \frac{\mathbb{P}(\{\langle Y^{(x)} \rangle_{[\frac{\hat{n}_x}{\alpha}, \alpha \hat{n}_x]} > \sqrt{x}\})}{\mathbb{P}(\{Y^{(x)}\} > \sqrt{x})} \rightarrow 1, \quad \text{as } x \rightarrow \infty. \quad (3.5)$$

In other words, as x increases, $\mathbb{P}(\{Q > x\}) = \mathbb{P}(\{Y^{(x)}\} > \sqrt{x})$ is essentially determined on a relatively small interval around the maximum variance index A_x . Also, (3.5) can be interpreted as a rigorous verification of the qualitative statement "rare events take place only in the most probable way [24, 41]."

Note that $\mathbb{P}(\{\langle Y^{(x)} \rangle_{[\frac{\hat{n}_x}{\alpha}, \alpha \hat{n}_x]} > \sqrt{x}\})$ with $a = 1$ corresponds to the lower bound $\Psi\left(\sqrt{\frac{x}{(\sigma_x^2)}}$. Since (3.5) holds for any arbitrary a greater than 1, it suggests that even if the lower bound were to asymptotically diverge from the exact tail probability, it would do so very slowly. In fact, although we know from (3.2), that the lower bound is log-similar, the (logarithmic) difference between the lower bound and the tail probability generally diverges. Even so, as will be shown in Chapter 4, the rate of divergence is relatively slow, and the bound accurately approximates the exact tail probability over a large range of queue lengths.

3.2 Asymptotic Upper Bound

In this section, we will derive an asymptotic upper bound for $\mathbb{P}(\{Q > x\})$. We say that $f(x)$ asymptotically bounds $g(x)$ from above if $\limsup_{x \rightarrow \infty} g(x)/f(x) \leq 1$. It should be noted here that Simonian [49] 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 (for example, check the bound obtained for the special case of the *Ornstein-Uhlenbeck* input processes in the paper), thus limiting its practical value. Moreover, the asymptotic tightness of this upper bound has only been shown for the Ornstein-Uhlenbeck process, and for more general processes we do not even know if the bound is asymptotically log-similar to the tail probability.

In contrast, the asymptotic upper bound for $\mathbb{P}(\{Q > x\})$ that we derive in this report is in a simple exponential form which can easily be obtained from the mean and autocovariance of the net input Gaussian process. Even though it is not a global upper bound, but an asymptotic upper bound, it is of both theoretical and practical importance, as will be discussed later. We will use this bound, in conjunction with the lower bound, to develop a good approximation for the tail probability.

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

3.2.1 Interpretation of Time-Scaling $Y_n^{(x)}$

Consider a continuous-time stochastic process $\tilde{Y}_t^{(x)}$ defined for each $x > 0$ as

$$\tilde{Y}_t^{(x)} := Y_{\lfloor \frac{xt}{\kappa} \rfloor}^{(x)}, \quad (3.6)$$

where $\lfloor z \rfloor$ denotes the largest integer that is smaller than or equal to z . The stochastic process $\tilde{Y}_t^{(x)}$ is simply an interpolated (by zero-order holding) and scaled (in time) version of $Y_n^{(x)}$, enforced to attain its maximum variance around $t = 1$, as $x \rightarrow \infty$. From the definition of $\tilde{Y}_t^{(x)}$, the following equations can easily be verified.

$$\langle \tilde{Y}^{(x)} \rangle = \langle Y^{(x)} \rangle, \quad (3.7)$$

$$\mathbb{P}(\{Q > x\}) = \mathbb{P}(\{\tilde{Y}^{(x)} > \sqrt{x}\}), \quad (3.8)$$

$$\sup_{t \geq 0} \text{Var}\{\tilde{Y}_t^{(x)}\} = \langle \sigma_x^2 \rangle, \quad \text{and} \quad (3.9)$$

$$\lim_{x \rightarrow \infty} C_{\tilde{Y}^{(x)}}(t_1, t_2) = \lim_{x \rightarrow \infty} C_{Y^{(x)}}(\lfloor \frac{xt_1}{\kappa} \rfloor, \lfloor \frac{xt_2}{\kappa} \rfloor) = \frac{S \min\{t_1, t_2\}}{\kappa(1+t_1)(1+t_2)} \quad (3.10)$$

(from Proposition B.1(c)).

Since $\tilde{Y}_t^{(x)}$ is a centered Gaussian process for each $x > 0$, (3.10) implies that $\tilde{Y}_t^{(x)}$ converges in distribution to a centered Gaussian process U_t having autocovariance function $C_U(t_1, t_2) = \frac{S \min\{t_1, t_2\}}{\kappa(t_1+1)(t_2+1)}$, as $x \rightarrow \infty$. One way of constructing the process U_t is to define it in terms of the standard Brownian motion process B_t as

$$U_t := \frac{\sqrt{S} B_t}{\sqrt{\kappa(1+t)}}.$$

Since $\tilde{Y}_t^{(x)}$ and U_t are continuous time processes, we briefly move our attention to continuous-time fluid queues. For continuous-time fluid queues, continuous-time stochastic processes $\dot{X}_t$, $\dot{Y}_t^{(x)}$, and $\dot{\tilde{Y}}_t^{(x)}$,

can be defined in an analogous way to their discrete-time counterparts:

$$\begin{aligned}\dot{X}_t &:= \Gamma_0 - \Gamma_{-t}, \\ \dot{Y}_t^{(x)} &:= \frac{\sqrt{x}(\dot{X}_t + \kappa t)}{x + \kappa t}, \quad \text{and} \\ \dot{\tilde{Y}}_t^{(x)} &:= \dot{Y}_{\frac{x}{\kappa}t}^{(x)}\end{aligned}$$

Here, Γ_t is a stochastic process with stationary increments and negative drift such that $\Gamma_t - \Gamma_s$ ($s \leq t$) represents the net input (the input rate minus the service rate) during the interval $(s, t]$ and $\kappa := -\frac{\mathbb{E}\{\Gamma_t - \Gamma_s\}}{t-s}$. Further, as mentioned in Chapter 2, the results (including (2.7) and (3.8)) obtained for discrete-time fluid queues can also be derived for continuous-time fluid queues under these definitions [18]. Note that if Γ_t is a Gaussian process with stationary and independent increments such that $\frac{\text{Var}\{\Gamma_t - \Gamma_s\}}{t-s} = S$ and $\kappa = -\frac{\mathbb{E}\{\Gamma_t - \Gamma_s\}}{t-s} = \kappa$, then $\tilde{Y}_t^{(x)}$ is identically distributed to U_t for every $x > 0$. Therefore, from (3.8), the queue length distribution of the corresponding continuous-time fluid queue is given by $\mathbb{P}(\{\langle U \rangle > \sqrt{x}\})$.

Roughly speaking, the continuous fluid queue driven by a Gaussian process Γ_t with stationary and independent increments, corresponds to the discrete-time fluid queue with an *i.i.d.* Gaussian net input γ_n . Hence, the convergence (in distribution) of $\tilde{Y}_t^{(x)}$ to U_t indicates that as x increases, $\tilde{Y}_t^{(x)}$ (or $Y_n^{(x)}$) behaves as if the net input process is an *i.i.d.* Gaussian process.[‡] This phenomenon can be intuitively interpreted as follows. From (3.4), $\hat{n}_x$, the time at which X_n is most likely to be larger than 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 almost invisible on the time scale of $\hat{n}_x$, and $Y_n^{(x)}$ behaves as if the input is *i.i.d.* Gaussian (with the same value of S as $Y_n^{(x)}$). 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, obviously although χ_n is not correlated, ζ_n is a correlated process. However, if we compare 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. Further, for such large values of n , these two partial sums will exhibit very similar stochastic behavior.

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

$$\mathbb{P}(\{Q > x\}) = \mathbb{P}(\{\tilde{Y}^{(x)} > \sqrt{x}\}) \approx \mathbb{P}(\{\langle U \rangle > \sqrt{x}\}). \quad (3.11)$$

The first equality of (3.11) is from (3.8) and the second step is from the fact that $\tilde{Y}_t^{(x)}$ converges to U_t in distribution. This approximation is intriguing because $\mathbb{P}(\{\langle U \rangle > \sqrt{x}\})$ can be computed in a simple form, i.e.,

$$\mathbb{P}(\{\langle U \rangle > \sqrt{x}\}) = \mathbb{P}\left(\left\{B_t > \sqrt{\frac{\kappa x}{S}}(t+1) \quad \text{for some } t \geq 0\right\}\right) = e^{-\frac{2\kappa x}{S}} \quad (\text{e.g. see [43, page 199]}).$$

In other words, this approach, in fact, results in the famous EB approximation. Therefore, to go beyond the EB approximation and obtain some information about the asymptotic constant in (1.1), 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 distribution of $\tilde{Y}_t^{(x)}$ converges to its limiting distribution.

[‡]Therefore, for sufficiently large x , the plots of $\mathbb{P}(\{Y_n^{(x)} > \sqrt{x}\})$ and the correlation coefficient between $Y_{\hat{n}_x}$ and $Y_n^{(x)}$ should look very similar to the two figures shown in Example 1.

3.2.2 Single-Exponential Based Asymptotic Upper Bound

Let B_t be the standard Brownian motion process and define a centered Gaussian process $Z_n^{(x)}$ ($n = 0, 1, \dots$) for each $x > 0$ by $Z_n^{(x)} := \frac{\sqrt{xSg(n)}B_n}{\sqrt{1+\kappa n}}$, where $g(n)$ is a function defined by (b.1) in Appendix B, such that $\lim_{n \rightarrow \infty} g(n) = 1$. Further, as in (3.6), we define $\tilde{Z}_t^{(x)} := Z_{\lfloor \frac{t}{\kappa} \rfloor}^{(x)}$. Now, it can easily be shown that $\tilde{Z}_t^{(x)}$ also converges to U_t in distribution. Further, as we will show in the proof of (the following) Theorem 3.2, the processes $Y_n^{(x)}$ and $Z_n^{(x)}$ (and hence $\tilde{Y}_t^{(x)}$ and $\tilde{Z}_t^{(x)}$) have the same variance. Therefore, by considering $\tilde{Z}_t^{(x)}$ we can capture how the variance of $\tilde{Y}_t^{(x)}$ converges to its limiting variance. This enables us to obtain an upper bound to the asymptotic constant which takes into account statistical multiplexing. More specifically, under conditions (C1) and (C3), it can be shown (see Appendix C) by using Slepian's inequality and Theorem 3.1, that $\mathbb{P}(\{\langle Z^{(x)} \rangle > x\})$ asymptotically bounds $\mathbb{P}(\{\langle Y^{(x)} \rangle > x\})$ from above. All of the above arguments can be made rigorous and lead to the following key theorem.

Theorem 3.2 Under conditions (C1)–(C3), $\limsup_{x \rightarrow \infty} e^{\frac{2\kappa x}{S}} \mathbb{P}(\{\langle X \rangle > x\}) \leq e^{-\frac{2\kappa^2 D}{S^2}}$. In other words, $e^{-\frac{2\kappa}{S}(x + \frac{\kappa D}{S})}$ asymptotically bounds $\mathbb{P}(\{\langle X \rangle > x\})$.

Proof of Theorem 3.2 : See Appendix C.

Theorem 3.2 gives us an exponential asymptotic upper bound ($e^{-\frac{2\kappa}{S}(x + \frac{\kappa D}{S})}$) to the tail probability $\mathbb{P}(\{\langle Q \rangle > x\}) = \mathbb{P}(\{\langle X \rangle > x\})$. Further, since it has been shown under condition (C1) that (1.1) holds for stationary Gaussian input processes with $\eta = \frac{2\kappa}{S}$ [4], Theorem 3.2 also provides us with an upper bound $e^{-\frac{2\kappa^2 D}{S^2}}$ to the asymptotic constant C given in (1.1). Note that the asymptotic upper bound exploits the advantage of statistical multiplexing in the sense that the bound for the asymptotic constant decreases exponentially when more sources are multiplexed. For instance, consider a fluid queueing system with an infinite buffer, a server having rate μ , and a stationary Gaussian input A_n . Then, the bound for the asymptotic constant of the corresponding tail probability is $e^{-\frac{2\kappa^2 D}{S^2}}$ where $S = \sum_{l=-\infty}^{\infty} C_\gamma(l)$, $D = 2 \sum_{l=1}^{\infty} l C_\gamma(l)$, and $\kappa = -\mathbb{E}\{\gamma_0\} = \mu - \mathbb{E}\{\lambda_0\}$. If we now increase the service rate by a factor of $M > 1$, and at the same time also increase the input rate by M (which corresponds to multiplexing M i.i.d. Gaussian sources), then the resulting bound for the asymptotic constant is $e^{-\frac{2\kappa^2 D}{S^2} M}$. Note that the bound decreases exponentially as M increases. Since the above properties hold for our upper bound to the asymptotic constant, it implies the following: 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. Here, it should be noted that this result coincides with the observation made on the asymptotic constant based on experimental studies [19].

The form of the upper bound to the asymptotic constant gives us more insight into the queueing behavior for stationary Gaussian sources. It is well known that S , in conjunction with κ , determines the asymptotic decay rate η given in (1.1) [4, 30], and that the generalized version of the index of dispersion for counts can be expressed in terms of S [5]. Therefore S can be thought of as a measure of the total “burstiness” of the input process, which is invariant to filtering or finite time-shifting of the arrival process. For example, let $a_n \in [0, 1]$ be a sequence that sums to 1, and consider a linear smoothing system which delays a portion of the input at time n by $m \geq 0$. Then, the output process $\tilde{\lambda}_n$ can be expressed as a convolution of a_n and the input process λ_n , i.e., $\tilde{\lambda}_n = \sum_{m=0}^{\infty} a_m \lambda_{n-m}$. From this relation, the autocovariance function of $\tilde{\lambda}_n$ can be computed as $C_{\tilde{\lambda}}(l) = \sum_{m_1=0}^{\infty} \sum_{m_2=0}^{\infty} a_{m_1} a_{m_2} C_\lambda(l + m_1 - m_2)$. Hence, we have

$$\sum_{l=-\infty}^{\infty} C_{\tilde{\lambda}}(l) = \sum_{m_1=0}^{\infty} a_{m_1} \sum_{m_2=0}^{\infty} a_{m_2} \sum_{l=-\infty}^{\infty} C_\lambda(l) = \sum_{l=-\infty}^{\infty} C_\lambda(l).$$

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, it is not difficult to check that $\sum_{l=0}^{\infty} lC_{\lambda}(l)$ may be different from $\sum_{l=0}^{\infty} lC_{\lambda'}(l)$, i.e., 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 non-negative 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 , and 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), 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 .

3.3 Numerical Examples and Discussions

In this section, we experimentally investigate the tightness of the lower bound and asymptotic upper bound and discuss their properties as approximations to the tail probability. Since, in general, the exact tail probability $\mathbb{P}(\{Q > x\})$ is not analytically obtainable, throughout this report, we use simulation techniques to validate our theoretical results. In particular, we use the Importance Sampling simulation technique described in [14] 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 [13]. 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 [14] are defined to be the time period between successive time epochs. We define these 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. Therefore, when $\mathbb{P}(\{Q > 0\})$ is very small, even though this does not necessarily imply the rareness of the regenerative cycle, it is usually difficult to get a sufficient 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 $\mathbb{P}(\{Q > 0\})$ is less than 10^{-4} . Hence, for all experiments, we set the utilization ($\rho = \mathbb{E}\{\lambda_0\}/\mu$) so that $\mathbb{P}(\{Q > 0\})$ is greater than 10^{-4} (as shown in the numerical figures, we do, however, estimate significantly lower values of $\mathbb{P}(\{Q > x\})$, for $x > 0$).

Example 2 In this example we consider fluid queues fed by two different Gaussian input processes. The autocovariance functions of these Gaussian processes are given as $C_{\lambda}(l) = 200 \times 0.95^{|l|}$ and $C_{\lambda}(l) = 100 \times 0.9^{|l|} + 60 \times 0.98^{|l|}$. Note that all the covariance functions are non-negative and vanish exponentially as l increases, so that they satisfy condition (C1).

In Figures 3.2 and 3.3, we show the exact tail probability and the lower 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 (5.26, 11.11, 17.65, 25, 33.33, 42.86) of $\kappa = \mu - \mathbb{E}\{\lambda_0\}$. As one can see in both figures, the lower bound matches the simulation results quite well. Also, note that, as expected from (3.2), the limiting slope of the lower bound approaches the limiting slope of the simulation curve. This, coupled with the fact that the lower bound closely matches the simulation results (over the range of values of x that

Figure 3.2: The exact tail probability and the lower bound for a Gaussian input process with autocovariance function $C_\lambda(l) = 200 \times 0.95^{|l|}$.

Figure 3.3: The exact tail probability and the lower bound for a Gaussian input process with autocovariance function $C_\lambda(l) = 100 \times 0.9^{|l|} + 60 \times 0.98^{|l|}$.

are shown in the figures), suggests that the lower bound should accurately approximate tail probabilities for even larger values of x . We have conducted an extensive experimental study which confirms that the lower bound does match the shape of the tail probability curve, and is accurate over a wide range of queue lengths [16, 17]. As will soon turn out, this is an important feature of the **maximum** variance based lower bound which cannot be achieved by single exponential (in terms of the queue length x) types of approximations.

Example 3 It is easy to check that the two autocovariance functions used in the previous example satisfy conditions (C1) – (C3). Therefore, from Theorem 3.2, an exponential **asymptotic** upper bound for the tail probability can be computed for these two Gaussian sources. In this example, we compute the asymptotic upper bound for the tail probability using exactly the same settings as in Example 2, and investigate its tightness.

In Figure 3.4, we show the exact tail probability and asymptotic upper bound for the Gaussian input process with autocovariance function $C_\lambda(l) = 200 \times 0.95^{|l|}$. 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 $-\frac{2\kappa}{S}$. 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 $e^{-\frac{2\kappa^2 D}{S^2 D}}$ is a tight bound to the asymptotic constant. The tightness of the asymptotic upper bound is also demonstrated in Figure 3.5, which shows the same curves for the Gaussian input process with the autocovariance function $C_\lambda(l) = 100 \times 0.9^{|l|} + 60 \times 0.98^{|l|}$. As in Figure 3.4, 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 3.5, 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 3.5 converges to its exponential asymptote **slowly**, while the tail probability in Figure 3.4 converges to its asymptote fairly fast, and forms a nearly straight line. The reason for this is that the autocovariance function of the Gaussian input used in Figure 3.5 consists of

Figure 3.4: The exact tail probability and the asymptotic upper bound for a Gaussian input process with autocovariance function $C_\lambda(l) = 200 \times 0.95^{|l|}$.

Figure 3.5: The exact tail probability and the asymptotic upper bound for a Gaussian input process with autocovariance function $C_\lambda(l) = 100 \times 0.9^{|l|} + 60 \times 0.98^{|l|}$.

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 will be demonstrated.

Example 4 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 this source is correlated at very different time scales. In Figure 3.6, the lower bound, the asymptotic upper bound, the EB approximation and simulation results are shown at $\kappa = 33.33$. As in the previous numerical results, the lower bound matches the whole simulation curve quite well from very small values of x to values of x as large as 10^5 . However, 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 (1.1) of this tail probability, 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 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 tight 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 [19].

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 [32]. 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

Figure 3.6: The exact tail probability, the EB approximation, the lower bound, and the asymptotic 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 $\kappa = 33.33$.

very likely to be correlated at multiple time scales. Therefore, it is important, as in the case with the lower bound, to be able to analyze the queue behavior for such traffic. In Chapter 4, we will develop a new asymptotic upper bound based on the maximum variance $\langle \sigma_x^2 \rangle$ which will be useful even for traffic correlated at multiple time-scales.

Example 5 In this example, we show that the asymptotic constant and the statistical multiplexing gain can be very different even for stationary Gaussian input processes having the same autocovariance sum S . In Figure 3.7, we plot 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 $S = 1000$. Even though these functions have the same values of S , as one can see in the figure, $C_2(l)$ is spread over a wider range of l than $C_1(l)$. Therefore, $C_2(l)$ has a significantly larger value of D than $C_1(l)$ (19487.16 for $C_1(l)$ versus 199501.48 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 Figure 3.8, 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 $\kappa = 5$. As in the previous examples, the upper bound for the asymptotic constant turns out to be tight, 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 number of sources in the system) increases as fast as $e^{\frac{2\kappa^2 D}{S^2} M}$. Therefore, as the number of the input processes and the link capacity are proportionally increased, the (logarithmic) difference between the asymptotic constants for these two Gaussian input processes will also increase very fast.

The above example is also 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

Figure 3.7: The autocovariance functions ($C_1(l) = 25.641 \times 0.95^{|l|}$ and $C_2(l) = 2.5063 \times 0.995^{|l|}$) of two Gaussian input processes having the same sum $S = 1000$.

Figure 3.8: The exact tail probability and the asymptotic upper bound for two Gaussian input processes with autocovariance functions $C_\lambda(l) = 25.641 \times 0.95^{|l|}$ and $C_\lambda(l) = 2.5063 \times 0.995^{|l|}$. κ is set to 5.

coefficients a_m ($m = 0, 1, \dots$). Therefore, this example illustrates that smoothing some types of network traffic which are correlated over a relatively short time scale, can significantly reduce network congestion. On the other hand, we also can expect that 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 been observed (e.g. [47]).

4. Asymptotic Upper Bound Based on the Maximum Variance

In the previous chapter, we developed two bounds for the tail probability $\mathbb{P}(\{Q > x\})$.

- Asymptotic Upper Bound: $e^{-\frac{2\kappa}{S}(x + \frac{\kappa D}{S})}$

- Lower Bound: $\Psi\left(\sqrt{\frac{x}{\langle\sigma_x^2\rangle}}\right)$

The asymptotic upper bound is in a very simple and elegant single exponential form, and has been derived using important results from Extreme Value Theory. 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 it differs from the exact tail only by a finite multiplicative constant. In fact, through empirical observations we have found that the leading term $e^{-\frac{2\kappa^2 D}{S^2}}$ provides a tight upper bound to the asymptotic constant in (1.1) which accounts for statistical multiplexing gain. However, in spite of the simplicity and theoretical value of the asymptotic upper bound, as we have discussed earlier, it may not accurately estimate the tail probability $\mathbb{P}(\{Q > x\})$ when the input traffic is correlated at different time-scales.

In contrast, the lower bound that we have developed is based on the maximum variance $\langle\sigma_x^2\rangle$ and was found to match the shape of the tail probability curve, and was hence accurate even for multiple time-scale correlated traffic. Thus, in this chapter we develop another asymptotic upper bound (under the conditions (C1)–(C3)) for the tail probability which has all the nice properties of the lower bound and the asymptotic upper bound derived in the previous chapter.

Remember that the lower bound is a simple (standard Gaussian tail distribution) function of $\sqrt{\frac{x}{\langle\sigma_x^2\rangle}}$. From Theorem 3.1, and the fact that the lower bound matches the shape of the tail probability curve, we can infer that the term $\frac{x}{\langle\sigma_x^2\rangle}$, as a function of x , contains key information about the shape of the tail probability curve. Our idea is to find a function $q(z)$ which resembles $\Psi(z)$ such that $q\left(\sqrt{\frac{x}{\langle\sigma_x^2\rangle}}\right)$ is similar to the asymptotic upper bound $e^{-\frac{2\kappa}{S}(x + \frac{\kappa D}{S})}$. In this way, $q\left(\sqrt{\frac{x}{\langle\sigma_x^2\rangle}}\right)$ would asymptotically bound the exact tail probability from above, and also closely track the shape of the tail probability curve. In other words, by finding such a function $q(z)$, we hope to develop a new asymptotic upper bound for the tail probability which is not only asymptotically tight, but also accurately approximates $\mathbb{P}(\{Q > x\})$ for any value of x . In the following proposition, which is based on Theorem 3.2, we find such an asymptotic upper bound.

Proposition 4.1 *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 $x \rightarrow \infty$. Therefore, with an additional condition (C3), $e^{-\frac{x}{2\langle\sigma_x^2\rangle}}$ asymptotically bounds $\mathbb{P}(\{X > x\})$.*

Proof of :Proposition4.1 : See Appendix C.

To avoid confusion with the asymptotic upper bound derived in Section 3.2, we name this new asymptotic upper bound $e^{-\frac{x}{2\langle\sigma_x^2\rangle}}$, the *Maximum Variance Asymptotic* (MVA) upper bound. Note that the MVA upper bound, as a function of $z = \sqrt{\frac{x}{\langle\sigma_x^2\rangle}}$, can be written as $q(z) = e^{-\frac{z^2}{2}}$. Further, from the following well known bound for $\Psi(z)$ [29], i.e.,

$$\frac{1 - z^{-2}}{\sqrt{2\pi}} z^{-1} e^{-\frac{z^2}{2}} \leq \Psi(z) \leq \frac{1}{\sqrt{2\pi}} z^{-1} e^{-\frac{z^2}{2}} \quad \text{for all } z > 0, \quad (4.1)$$

Figure 4.1: The difference $\log e^{-\frac{x}{2\langle\sigma_x^2\rangle}} - \log \Psi\left(\sqrt{\frac{x}{\langle\sigma_x^2\rangle}}\right)$ versus the MVA upper bound $\log e^{-\frac{x}{2\langle\sigma_x^2\rangle}}$.

we have

$$\Psi(z) \sim \frac{e^{-\frac{z^2}{2}}}{\sqrt{2\pi z}} = \frac{q(z)}{\sqrt{2\pi z}}. \quad (4.2)$$

Since the inequalities in (4.1) are very tight, even for moderately small values of z , the ratio $\Psi(z)/\frac{e^{-\frac{z^2}{2}}}{\sqrt{2\pi z}}$ converges to 1 very rapidly. Therefore, the major difference between $\Psi(z)$ and $e^{-\frac{z^2}{2}}$ is the multiplicative term $\frac{1}{\sqrt{2\pi z}}$ in the right-hand side of (4.2). However, this term is very slowly varying compared to the remaining part $e^{-\frac{z^2}{2}}$. Therefore, the shape of the MVA upper bound curve should almost be the same as that of the lower bound. Further, since the MVA upper bound is asymptotically similar to the asymptotic upper bound derived in the previous chapter, we expect that it should be an accurate approximation for any queue lengths x . Also, note that the MVA upper bound is being obtained by lifting the lower bound in such a way that it becomes an asymptotic upper bound. Hence, unlike the asymptotic upper bound in Section 3.2, we expect that the MVA upper bound will bound the tail probability even for very small values of queue lengths; a prediction that will be verified through simulations.

Now, a direct result of Proposition 4.1 is that under conditions (C1)–(C2),

$$\Psi\left(\sqrt{\frac{x}{\langle\sigma_x^2\rangle}}\right) \sim \sqrt{\frac{\langle\sigma_x^2\rangle}{2\pi x}} e^{-\frac{x}{2\langle\sigma_x^2\rangle}} \sim \sqrt{\frac{S}{8\pi\kappa x}} e^{-\frac{2\kappa}{S}(x + \frac{\kappa D}{S})}. \quad (4.3)$$

Note that the second similarity is from Propositions B.3 and 4.1. From (4.3), 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 $\sqrt{\frac{S}{8\pi\kappa x}}$ is slowly decreasing compared to the remaining term $e^{-\frac{2\kappa}{S}(x + \frac{\kappa D}{S})}$, as $x \rightarrow \infty$. For this reason, the deviation of the lower bound from the tail probability was basically unrecognizable in Figures 3.2, 3.3, and 3.6. In fact in all our tested sequences the eventual divergence of the lower bound is not observed, even for probabilities as small as 10^{-20} . Perhaps the following observation will shed further light on this issue.

An interesting observation is that the (logarithmic) difference $\log e^{-\frac{x}{2\langle\sigma_x^2\rangle}} - \log \Psi\left(\sqrt{\frac{x}{\langle\sigma_x^2\rangle}}\right)$ between the

Figure 4.2: The exact tail probability and the MVA upper bound for a Gaussian input process with autocovariance function $C_\lambda(l) = 200 \times 0.95^{|l|}$.

Figure 4.3: The exact tail probability and the MVA upper bound for a Gaussian input process with autocovariance function $C_\lambda(l) = 100 \times 0.9^{|l|} + 60 \times 0.98^{|l|}$.

MVA upper bound and the lower bound is actually a function of $\sqrt{\frac{x}{\langle \sigma_x^2 \rangle}}$, that can be closely approximated by $\frac{1}{2} \log \frac{2\pi x}{\langle \sigma_x^2 \rangle}$. 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 Figure 4.1. As one can see 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} . This also suggests that the MVA upper bound and lower bound provide a narrow envelope that encapsulates the exact tail probability over a wide range of queue lengths. Figure 4.1 indicates that this envelope will be quite tight even at probabilities as small as 10^{-20} .

4.1 Numerical Examples and Discussion

In this section, we investigate the tightness of the MVA upper bound by applying it to exactly the same situations as in Examples 2, 3, and 4. In all of these examples we will observe that the MVA upper bound accurately tracks the tail probability over a wide range of queue lengths.

Example 6 In Figures 4.2 and 4.3, we show the exact tail probability and the MVA upper bound corresponding to the same setting as in Figures 3.2 and 3.3, respectively. By comparing these two figures with Figures 3.2 and 3.3, one can see that, as expected, the shape of the MVA upper bound curve closely resembles that of the lower bound. Further, also as expected, the MVA upper bound, bounds the tail probability not only for large values of queue lengths (as did our first asymptotic upper bound in Figures 3.4 and 3.5) but for the entire range of queue lengths.

Example 7 In Figure 4.4, the exact tail probability, the lower bound, and the MVA upper bound for a Gaussian input process with autocovariance function $C_\lambda(l) = 104 \times 0.99^{|l|} + 64.14 \times 0.999^{|l|} + 31.86 \times 0.9999^{|l|}$ are displayed. As in Example 4, κ is set to 33.33. Note that the lower bound and the MVA upper bound tightly 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 have indicated that: (1) the tail probability almost never escapes from the envelope constructed by the bounds, as long as conditions (C1)–(C3) are satisfied; and

Figure 4.4: The exact tail probability, the lower bound, and the MVA upper bound for a Gaussian input process with autocovariance function $C_\lambda(l) = 104 \times 0.99^{|l|} + 64.14 \times 0.999^{|l|} + 31.86 \times 0.9999^{|l|}$.

(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 around an order of magnitude.

5. Applications for General Input Processes

The numerical results provided in Chapters 3 and 4 were for stationary Gaussian input processes. Further, both the asymptotic upper bounds developed in the previous chapters were derived under three conditions (C1)–(C3). In this chapter, 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 not Gaussian.

5.1 General Gaussian Process

The relation (3.1) is very generally true. Hence, the lower bound in (3.3) given by $\Psi\left(\sqrt{\frac{x}{\langle\sigma_x^2\rangle}}\right)$ is valid long as the input process is stationary Gaussian. On the other hand, both the asymptotic upper bounds developed in Section 3.2 and Chapter 4, require conditions (C1)–(C3). Hence, in order to identify the class of stationary Gaussian processes for which the asymptotic upper bounds are valid, it is important to know what kind of stationary Gaussian processes satisfy these conditions.

The 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 self-similar behavior and those that do not [5] (see also [37, 40, 41] for the definition and properties of self-similar processes). Also, (C1) is a sufficient condition for the ergodicity of a stationary Gaussian process [53], and therefore, under this condition the tail probability satisfies (1.1) with $\eta = \frac{2\kappa}{S}$, and some finite constant C [4].

Condition (C2) is on the absolute summability of a weighted autocovariance function of the input process. It is easy to see that (C2) is somewhat more restrictive than (C1), and that this condition is 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. However, it should be noted that some types of network applications (such as MPEG video) generate network traffic in a fairly periodic fashion, which may result in a large enough negative component of the autocovariance function to violate condition (C3). Thus, in the following example, we first investigate the performance of the lower bound and the MVA upper bound for input processes that do not satisfy condition (C3).

Example 8 In Figures 5.1 and 5.2, we show the exact tail probability, the lower bound, and the MVA upper bound for two Gaussian input processes whose autocovariance functions are $10 \times 0.95^{|l|} \cos \frac{\pi l}{18}$ and $10 \times 0.9^{|l|} \cos \frac{\pi l}{12} + 0.1 \times 0.99^{|l|}$, respectively. One can easily check that these autocovariance functions do not satisfy the 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 $\langle\sigma_x^2\rangle$. Further, in both figures, the MVA upper bound seems to be asymptotically close to the tail probability. This suggests that the bound $e^{-\frac{2\kappa^2 D}{s^2}}$ to the asymptotic constant C in (1.1) 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 $e^{-\frac{2\kappa^2 D}{s^2}}$ has important properties that the asymptotic constant is known to have such as: (1) if the input process is *i.i.d.* Gaussian, then

Figure 5.1: The exact tail probability, the lower bound, and the MVA upper bound for a Gaussian input process with autocovariance function $C_\lambda(l) = 10 \times 0.95^{|l|} \cos \frac{\pi l}{18}$ and $\kappa = 1, 2$.

Figure 5.2: The exact tail probability, the lower bound, and the MVA upper bound for a Gaussian input process with autocovariance function $C_\lambda(l) = 10 \times 0.9^{|l|} \cos \frac{\pi l}{12} + 0.1 \times 0.991^{|l|}$ and $\kappa = 1, 2$.

$D = 0$ and the asymptotic upper bound simply becomes $e^{-\frac{2\kappa x}{S}}$ which is a well-known bound for the level crossing probability of a random walk (see [43, page 236]); (2) Also, D can have a negative value, only when the input process shows periodic behavior, i.e., the amount of input at time n reduces the expected amount of input time $n+1$, for some 1 . If D takes on a negative value, then $e^{-\frac{2\kappa x D}{S^2}}$ is greater than 1 , and will increase exponentially with the size of the system (as explained in Section 9.2). This suggests 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 [19, 46].

As mentioned above, the input process shows self-similar behavior when condition (C1) is violated. In this case, the tail probability may not even be asymptotically exponential [40], and hence one cannot obtain an asymptotic upper bound in a single exponential form. 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 the MVA upper bound. In fact, in [40, 41], an equivalent approximation to the MVA upper bound has been computed and used to approximate the tail probability for a special class of Gaussian processes called **Fractal Brownian** motion that belong to the class of self-similar input processes. In these papers, the tail probability was approximated by the lower bound given in (3.1), but the lower bound itself was evaluated through yet another approximation $\Psi(z) \approx e^{-\frac{z^2}{2}}$, instead of the exact standard Gaussian tail function $\Psi(z)$. As a consequence, the approximation used in these papers does not correspond to the real lower bound in (3.3) but actually corresponds to our MVA upper bound. Nevertheless, the experimental result in [41] shows that the MVA upper bound (used only as an approximation) can approximate the tail probability reasonably well even when condition (C1) is violated. Since the lower bound is closely related to the MVA upper bound as shown in Figure 4.1, it too can be useful in analyzing Fractal Brownian motion processes.

In the following section, we weaken the Gaussian assumption on the input process itself, 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.

5.2 Applications to Voice and Video Traffic

As mentioned in Chapter 2, 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 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, hundreds or even thousands of network applications are expected to share an ATM link; an **OC-3** (155.52 Mbps) line can accommodate over 6800 voice calls (assuming 16 Kbps mean bit-rate) and an **OC-12** line over 300 MPEG video calls (assuming 1.5 Mbps 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 to characterize the aggregate input process by a Gaussian process [16, 17, 361]. Through empirical evidence we have found that a few hundred sources are generally sufficient for the Gaussian approximation to be quite good.

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 through several numerical examples involving voice and video traffic. It should be emphasized that since we have weakened the Gaussian assumption, our theoretical results cannot strictly be thought of as bounds, but 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 approximation is reasonably good, our analytical approximations do behave like real bounds over the tail probabilities of interest.

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 an multiplexer being served by an **OC-3** (155.52 Mbps) or **OC-12** (622 Mbps) line.

Voice Traffic Sources:

Example 9 The typical behavior of efficiently encoded voice traffic is that it alternate; between "active" and "inactive" states [20, 33]. Hence, Markov modulated On-Off processes have frequently been used to model voice traffic [20, 50]. For our experiment, we assume a 10 msec slot size and use a discrete-time On-Off MMF process as a voice traffic source model whose state transition matrix and rate vector are given as follows.

$$\begin{aligned} \text{State Transition Matrix : } & \begin{bmatrix} 0.9833 & 0.0167 \\ 0.025 & 0.975 \end{bmatrix} \\ \text{Input Rate Vector : } & \begin{bmatrix} 0 \text{ cells/slot} \\ 0.85 \text{ cells/slot} \end{bmatrix} \end{aligned}$$

This voice traffic source model is obtained by discretizing the continuous-time MMF voice traffic source model used in [46]. In Figure 5.3, we show the exact tail, the lower bound and the MVA upper bound for 42500 and 42800 voice sources served by an **OC-12** (622.08 Mbps) line. As one can see in the figure, the simulation results are tightly bounded between the lower bound and the MVA upper bound.

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

Figure 5.3: The exact tail probability, the lower bound and the MVA upper bound for a multiplexer serving 42500 and 42800 voice traffic sources. The output link capacity is set to 622.08 Mbps (OC-12 line).

Figure 5.4: 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 Mbps (OC-3 line).

is likely to have different characteristics from that of a video source that transmit;; 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, the detailed modeling of such diverse video traffic sources may not be an easy and efficient way of characterizing these sources. From this viewpoint, traffic characterization based only on the first two moments (mean and autocovariance) has advantages 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 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.

Example 10 In this example, we use real MPEG video (frame size) traces generated by Rose [42]. To simulate MPEG-encoded video traffic, 16 different MPEG coded traces of 40000 frames are concatenated into one trace of 640000 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/sec, the total length (640000 frames) of the concatenated frame size trace corresponds to more than 7 hours of play time. Since the trace is quite 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 msec 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 msec or equivalently 4 slots). In Figure 5.4, 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, consequently, 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

Figure 5.5: The autocovariance function measured from JPEG-encoded movie "Star Wars" and its approximation with the weighted sum of 3 exponential functions.

Figure 5.6: 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.

exact tail probability within an order of magnitude. It is important to note that because of the structure of MPEG coding scheme, the traffic shows strong periodic behavior, and its autocovariance function may have a significant negative portion. This in turn may result in situations when condition (C3) is violated. However, as is illustrated in this example, even in those cases, our maximum variance based bounds can be used to accurately approximate the tail probability.

Example 11 In this example, we use a frame size trace of the JPEG-encoded movie "Star Wars" to simulate real video sources, and experimentally obtain the tail probability $\mathbb{P}(\{Q > x\})$ for these 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, and use the model to obtain our bounds and another set of simulation results. In Figure 5.5, we show the autocovariance function measured directly from the trace and its approximation. As one can see from the figure, the autocovariance function measured from the frame size trace has quite an irregular shape. Further, the autocovariance function takes on large positive values at very large values of l , the time difference. This implies that the traffic is correlated over a long time. In fact, many types of video traffic have been found to be heavily correlated over multiple time-scale; or even thought to exhibit self-similar behavior over a certain time-period [3, 10, 32]. To capture this multiple time-scale correlation of the frame size trace, we can model the JPEG video traffic source as the superposition of 3 two-state MMF processes with very different mean state sojourn times, as specified below.

$$\begin{aligned}
 &\text{State Transition Matrices :} \\
 &\begin{bmatrix} 0.999138 & 0.000862 \\ 0.000862 & 0.999138 \end{bmatrix} \quad \begin{bmatrix} 0.9999138 & 0.0000862 \\ 0.0000862 & 0.9999138 \end{bmatrix} \quad \begin{bmatrix} 0.99999138 & 0.00000862 \\ 0.00000862 & 0.99999138 \end{bmatrix} \\
 &\text{Input Rate Vectors :} \\
 &\begin{bmatrix} 99.5296 \text{ cells/slot} \\ 151.8123 \text{ cells/slot} \end{bmatrix} \quad \begin{bmatrix} 0 \text{ cells/slot} \\ 22.3987 \text{ cells/slot} \end{bmatrix} \quad \begin{bmatrix} 0 \text{ cells/slot} \\ 15.4486 \text{ cells/slot} \end{bmatrix}
 \end{aligned}$$

More precisely, this source model is obtained by matching the autocovariance function measured from the frame size trace using the *Least Square* method. The approximated autocovariance function is compared

Figure 5.7: Admissible combinations of voice and JPEG-encoded video calls for an *OC-12* link with 20000 cell buffers, computed by simulation, the lower bound, and the MVA upper bound. The maximum tolerable tail probability (φ) is set to 10^{-6} .

to the measured autocovariance function in Figure 5.5. The purpose of designing a model for the JPEG traffic model is to demonstrate that the queueing 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 Figure 5.6, we show simulation results, the lower bound, and the 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 msec. Since the frame size trace is from video sequences captured at 30 frames/sec, each frame size should be read out over 4 slots. As in the previous example, we assume that a frame is uniformly transmitted over 4 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 close to each other and encompassed within the lower and MVA upper bounds.

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.

Example 12 In Figure 5.7, 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 20000 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 9, we use a JPEG video traffic model that is somewhat different from the model used in Example 11. The reason is that the

traffic source models used in Example 11 results in the generation of too small a number of regenerative cycles in a reasonable amount of time, to effectively employ the importance sampling simulation technique. Instead we use a more generic model that captures the multiple-time scale correlation observed in JPEG video traces. Specifically, the JPEG video traffic source model used in this example is a superposition of a *i.i.d.* Gaussian process and 3 two-state MMF processes. The state transition matrices and the input rate vectors of these MMF processes and the mean and the variance of the *i.i.d.* Gaussian process are given as follows.

$$\begin{array}{ccc}
& \text{State Transition Matrices :} & \\
\begin{bmatrix} 0.99 & 0.01 \\ 0.01 & 0.99 \end{bmatrix} & \begin{bmatrix} 0.999 & 0.001 \\ 0.001 & 0.999 \end{bmatrix} & \begin{bmatrix} 0.9999 & 0.0001 \\ 0.0001 & 0.9999 \end{bmatrix} \\
& \text{Input Rate Vectors :} & \\
\begin{bmatrix} 0 \text{ cells/slot} \\ 45.516 \text{ cells/slot} \end{bmatrix} & \begin{bmatrix} 0 \text{ cells/slot} \\ 31.86 \text{ cells/slot} \end{bmatrix} & \begin{bmatrix} 0 \text{ cells/slot} \\ 18.204 \text{ cells/slot} \end{bmatrix}
\end{array}$$

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

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

It is interesting to note that in Figure 5.7, the admissible regions 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. This example is quite typical of the accuracy of our maximum variance based bounds for admission control.

6. Conclusion

In this report we introduce a simple lower bound and derive two asymptotic upper bounds to analyze the tail of the steady state distribution $\mathbb{P}(\{Q > x\})$ in 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.

We first introduce a simple lower bound for Gaussian input processes, based on the maximum variance $\langle \sigma_x^2 \rangle$. We then provide an intuitive explanation and develop a theoretical result that emphasizes the importance of the maximum variance point in capturing the supremum distribution of Gaussian processes.

For a Gaussian input process satisfying fairly general conditions, we derive an exponential asymptotic upper bound $e^{-\frac{2x}{S}(x + \frac{xP}{S})}$ to the tail probability $\mathbb{P}(\{Q > x\})$ using key results in Extreme Value Theory. This asymptotic upper bound in turn provides a theoretical contribution to the Extreme Value literature. The asymptotic upper bound also results in a tight upper bound to the asymptotic constant.

Building upon our exponential asymptotic upper bound, we derive another asymptotic (MVA) upper bound $e^{-\frac{x}{2\langle \sigma_x^2 \rangle}}$, 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 report 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 [18]. 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.

A. Results from Extreme Value Theory

Here, we quote three results from [7], which are used at critical steps in proving our main results.

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

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

where $\langle \sigma^2 \rangle := \sup_{t \in T} \mathbb{E}\{\zeta_t^2\}$.

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

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

Theorem A.3 Let $\{\zeta_t : t \in T\}$ be a centered Gaussian process and define a pseudo-metric d on T as $d(t_1, t_2) := \sqrt{\mathbb{E}\{(\zeta_{t_1} - \zeta_{t_2})^2\}}$ (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 K such that

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

B. Preliminaries

In this appendix, we provide several propositions which will be used to prove Theorem 3.1 and Theorem 3.2.

Even though the following proposition includes results that may be well known (e.g. (b)), we provide complete proofs to save the readers' inconvenience in searching for the proper references.

Proposition B.1

- (a) For $n > 1$, $\frac{\text{Var}\{X_n\}}{n} - \frac{\text{Var}\{X_{n-1}\}}{n-1} = \frac{2}{n(n-1)} \sum_{l=1}^{n-1} l C_\gamma(l)$.
 (b) $C_X(n_1, n_2) = \frac{1}{2} (\text{Var}\{X_{n_1}\} + \text{Var}\{X_{n_2}\} - \text{Var}\{X_{|n_1-n_2|}\})$.
 (c) Under condition (C1), for any two non-negative sequences k_i and l_i such that $k_i, l_i \rightarrow \infty$ and $\frac{k_i}{l_i} \rightarrow \alpha \geq 1$ as $i \rightarrow \infty$,

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

In particular, $\lim_{n \rightarrow \infty} \frac{\text{Var}\{X_n\}}{n} = S$.

- (d) Let $\tilde{D} := \sum_{l=1}^{\infty} l |C_\gamma(l)|$. Then, under conditions (C1) and (C2), $\left| \frac{\text{Var}\{X_{n_1}\}}{n_1} - \frac{\text{Var}\{X_{n_2}\}}{n_2} \right| \leq \frac{\tilde{D} |n_1 - n_2|}{n_1 n_2}$ for all $n_1, n_2 > 0$, and $\lim_{n \rightarrow \infty} n \left(S - \frac{\text{Var}\{X_n\}}{n} \right) = D$.
 (e) Under conditions (C1)–(C3), $\frac{\text{Var}\{X_n\}}{n} < S$ and there exists an n_0 such that for all $n \geq n_0$, $\frac{\text{Var}\{X_n\}}{n} = \sup_{0 < m \leq n} \frac{\text{Var}\{X_m\}}{m}$.

Proof of Proposition B.1 : (a) From (2.4), for $n > 1$

$$\frac{\text{Var}\{X_n\}}{n} - \frac{\text{Var}\{X_{n-1}\}}{n-1} = 2 \sum_{l=1}^{n-1} \left(1 - \frac{l}{n}\right) C_\gamma(l) - 2 \sum_{l=1}^{n-2} \left(1 - \frac{l}{n-1}\right) C_\gamma(l) = \frac{2}{n(n-1)} \sum_{l=1}^{n-1} l C_\gamma(l).$$

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

$$\begin{aligned} 2C_X(n_1, n_2) &= \sum_{l_1=1}^{n_1} \sum_{l_2=1}^{n_2} C_\gamma(l_2 - l_1) + \sum_{l_1=1}^{n_2} \sum_{l_2=1}^{n_1} C_\gamma(l_2 - l_1) \\ &= \sum_{l_1=1}^{n_2} \sum_{l_2=1}^{n_2} C_\gamma(l_2 - l_1) - \sum_{l_1=n_1+1}^{n_2} \sum_{l_2=1}^{n_2} C_\gamma(l_2 - l_1) \\ &\quad + \sum_{l_1=1}^{n_1} \sum_{l_2=1}^{n_1} C_\gamma(l_2 - l_1) + \sum_{l_1=n_1+1}^{n_2} \sum_{l_2=1}^{n_1} C_\gamma(l_2 - l_1) \\ &= \sum_{l_1=1}^{n_2} \sum_{l_2=1}^{n_2} C_\gamma(l_2 - l_1) + \sum_{l_1=1}^{n_1} \sum_{l_2=1}^{n_1} C_\gamma(l_2 - l_1) - \\ &\quad \sum_{l_1=1}^{n_2-n_1} \sum_{l_2=1}^{n_2-n_1} C_\gamma(l_2 - l_1) \\ &= \text{Var}\{X_{n_2}\} + \text{Var}\{X_{n_1}\} - \text{Var}\{X_{n_2-n_1}\}. \end{aligned}$$

- (c) From the symmetry of the autocovariance function, it suffices to show that $\lim_{i \rightarrow \infty} \frac{C_X(k_i, l_i)}{l_i} = S$. Let $h_i(m)$ be defined as

$$h_i(m) = \begin{cases} \left(1 + \frac{m}{\min\{k_i, l_i\}}\right) C_\gamma(m) & \text{if } -\min\{k_i, l_i\} < m < 0, \\ C_\gamma(m) & \text{if } 0 \leq m \leq |k_i - l_i|, \\ \left(1 - \frac{m - |k_i - l_i|}{\min\{k_i, l_i\}}\right) C_\gamma(m) & \text{if } |k_i - l_i| < m < \max\{k_i, l_i\}, \\ 0 & \text{otherwise.} \end{cases}$$

Then we get

$$\frac{C_\gamma(k_i, l_i)}{\min\{k_i, l_i\}} = \frac{1}{\min\{k_i, l_i\}} \sum_{m_1=1}^{k_i} \sum_{m_2=1}^{l_i} C_\gamma(m_1 - m_2) = \sum_{m=-\infty}^{\infty} h_i(m).$$

However, since $\lim_{i \rightarrow \infty} h_i(m) = C_\gamma(m)$ and $|h_i(m)| < |C_\gamma(m)|$, it follows from the Dominated Convergence *Theorem* (DCT) that

$$\lim_{i \rightarrow \infty} \frac{C_\gamma(k_i, l_i)}{l_i} = \lim_{i \rightarrow \infty} \frac{C_\gamma(k_i, l_i)}{\min\{k_i, l_i\}} \lim_{i \rightarrow \infty} \frac{\min\{k_i, l_i\}}{l_i} = \lim_{i \rightarrow \infty} \sum_{m=-\infty}^{\infty} h_i(m) = S.$$

(d) W.L.O.G. assume $n_2 > n_1 > 0$. From (2.4), we have

$$\begin{aligned} \frac{\text{Var}\{X_{n_2}\}}{n_2} - \frac{\text{Var}\{X_{n_1}\}}{n_1} &= 2 \left(\sum_{l=1}^{n_2-1} \left(1 - \frac{l}{n_2}\right) C_\gamma(l) - \sum_{l=1}^{n_1-1} \left(1 - \frac{l}{n_1}\right) C_\gamma(l) \right) \\ &= \frac{2(n_2 - n_1)}{n_1 n_2} \left(\sum_{l=1}^{n_1-1} l C_\gamma(l) + \sum_{l=n_1}^{n_2-1} \frac{n_1(n_2 - l)}{n_2 - n_1} C_\gamma(l) \right). \end{aligned}$$

Since $0 \leq \frac{n_1(n_2 - l)}{n_2 - n_1} \leq l$ for $l = n_1, n_1 + 1, \dots, n_2 - 1$, it follows that

$$\begin{aligned} \left| \frac{\text{Var}\{X_{n_2}\}}{n_2} - \frac{\text{Var}\{X_{n_1}\}}{n_1} \right| &\leq \frac{2(n_2 - n_1)}{n_1 n_2} \left(\sum_{l=1}^{n_1-1} l |C_\gamma(l)| + \sum_{l=n_1}^{n_2-1} \frac{n_1(n_2 - l)}{n_2 - n_1} |C_\gamma(l)| \right) \\ &\leq \frac{2(n_2 - n_1)}{n_1 n_2} \sum_{l=1}^{n_2-1} l |C_\gamma(l)| \leq \frac{\tilde{D}(n_2 - n_1)}{n_1 n_2}. \end{aligned}$$

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

$$h_n(m) := \begin{cases} m C_\gamma(m) & \text{if } m = 0, 1, \dots, n, \\ n C_\gamma(m) & \text{otherwise.} \end{cases}$$

Then, from (2.4) and the definition of S ,

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

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

(e) From (2.4) and the definition of S ,

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

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

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

Also, since $\frac{\text{Var}\{X_n\}}{n}$ is increasing for $n \geq n_1$, $\frac{\text{Var}\{X_m\}}{m} \leq \frac{\text{Var}\{X_n\}}{n}$ for $m \in (n_1, n)$. Therefore, for all $n \geq n_0$, $\frac{\text{Var}\{X_n\}}{n} = \sup_{0 < m \leq n} \frac{\text{Var}\{X_m\}}{m}$. Q.E.D.

Proposition B.2 Let A_n be the index at which $\sigma_{x,n}^2$ attains its maximum $\langle \sigma_x^2 \rangle$. Then, under condition (C1), $\hat{n}_x \sim \frac{x}{\kappa}$ as $x \rightarrow \infty$. Further, under conditions (C1) and (C2), $\lim_{x \rightarrow \infty} \frac{\hat{n}_x - \frac{x}{\kappa}}{x^\epsilon} = 0$ for all $\epsilon > 0$.

Proof of Proposition B.2 : For notational simplicity, we define a function $g(n)$ for $n = 0, 1, 2, \dots$ as

$$g(n) := \begin{cases} 0 & \text{if } n = 0, \\ \frac{x}{\kappa n} & \text{otherwise.} \end{cases} \quad (\text{b.1})$$

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

$$\sigma_{x,n}^2 = \frac{S x n}{(x + \kappa n)^2} g(n). \quad (\text{b.2})$$

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

$$\frac{S x \hat{n}_x g(\hat{n}_x)}{(x + \kappa \hat{n}_x)^2} = \sigma_{x, \hat{n}_x}^2 \leq \sigma_{x, n_x}^2 = \frac{S x n_x g(n_x)}{(x + \kappa n_x)^2} \leq \frac{S x \hat{n}_x G}{(x + \kappa \hat{n}_x)^2}. \quad (\text{b.3})$$

By solving (b.3) for A_n , we have

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

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

Since $\frac{S x n}{(x + \kappa n)^2}$ attains its maximum at $n = n_x$, we know from (b.3) that $g(\hat{n}_x) \leq g(n_x)$, and that the following relation should hold.

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

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

$$\lim_{x \rightarrow \infty} \frac{\kappa \hat{n}_x}{x} = 1. \quad (\text{b.5})$$

Thus, we have proven the first part of the proposition. Now, assume $C_\gamma(l)$ satisfies conditions (C1) and (C2). From Proposition B.1(d), note that

$$|g(n_1) - g(n_2)| \leq \frac{\tilde{D}|n_2 - n_1|}{S n_1 n_2}. \quad (\text{b.6})$$

From (b.4), it follows that

$$\left| \hat{n}_x - \frac{x}{\kappa} \right| \leq \frac{2x}{\kappa} \left(\left| \frac{g(\hat{n}_x)(\frac{x}{\kappa} + \hat{n}_x)^2}{4\hat{n}_x \frac{x}{\kappa} g(\hat{n}_x)} - 1 \right| + \sqrt{\frac{g(\hat{n}_x)(\frac{x}{\kappa} + \hat{n}_x)^2}{4\hat{n}_x \frac{x}{\kappa} g(\hat{n}_x)} \left| \frac{g(\hat{n}_x)(\frac{x}{\kappa} + \hat{n}_x)^2}{4\hat{n}_x \frac{x}{\kappa} g(\hat{n}_x)} - 1 \right|} \right). \quad (\text{b.7})$$

On the other hand,

$$\left| \frac{g(\hat{n}_x)(\frac{x}{\kappa} + \hat{n}_x)^2}{4\hat{n}_x \frac{x}{\kappa} g(\hat{n}_x)} - 1 \right| \leq \frac{1}{g(\hat{n}_x)} \left(\frac{(\frac{x}{\kappa} + \hat{n}_x)^2}{4\hat{n}_x \frac{x}{\kappa}} |g(\hat{n}_x) - g(\hat{n}_x)| + g(\hat{n}_x) \frac{(\frac{x}{\kappa} - \hat{n}_x)^2}{4\hat{n}_x \frac{x}{\kappa}} \right). \quad (\text{b.8})$$

Since $\frac{(\frac{x}{\kappa} + \hat{n}_x)^2}{4\hat{n}_x \frac{x}{\kappa}}$, $\frac{\kappa \hat{n}_x}{x}$ and $g(\hat{n}_x)$ approach 1 as $x \rightarrow \infty$ and since $|\frac{x}{\kappa} - \hat{n}_x| \leq 1$, it follows from (b.8) that for sufficiently large x ,

$$\left| \frac{g(\hat{n}_x)(\frac{x}{\kappa} + \hat{n}_x)^2}{4\hat{n}_x \frac{x}{\kappa} g(\hat{n}_x)} - 1 \right| \leq 2|g(\hat{n}_x) - g(\hat{n}_x)| + \frac{\kappa^2}{x^2}. \quad (\text{b.9})$$

Therefore, from (b.7) and (b.9), for sufficiently large x , we have

$$\begin{aligned} \left| \hat{n}_x - \frac{x}{\kappa} \right| &\leq \frac{4|g(\hat{n}_1) - g(\hat{n}_x)|}{\frac{\kappa}{x}} + \frac{2\kappa}{x} + \sqrt{16 \frac{|g(\hat{n}_x) - g(\hat{n}_x)|}{\frac{\kappa^2}{x^2}}} + 8 \\ &\quad (\text{from the fact that } \frac{g(\hat{n}_x)(\frac{x}{\kappa} + \hat{n}_x)^2}{4\hat{n}_x \frac{x}{\kappa} g(\hat{n}_x)} \rightarrow 1 \text{ as } x \rightarrow \infty) \\ &\leq \frac{4\tilde{D}|\hat{n}_x - \hat{n}_x|}{S \frac{\kappa}{x} \hat{n}_x \hat{n}_x} + \frac{2\kappa}{x} + \sqrt{16 \frac{\tilde{D}|\hat{n}_x - \hat{n}_x|}{S \hat{n}_x \hat{n}_x \frac{\kappa}{x^2}}} + 8 \quad (\text{from (b.6)}) \\ &\leq 1 + \frac{2\kappa}{x} + \sqrt{32 \frac{\tilde{D}|\hat{n}_x - \hat{n}_x|}{S}} + 8 \\ &\quad (\text{since } \frac{|\hat{n}_x - \hat{n}_x|}{\hat{n}_x} \rightarrow 0 \text{ and } \frac{\hat{n}_x \kappa}{x}, \frac{\hat{n}_x \kappa}{x} \rightarrow 1 \text{ as } x \rightarrow \infty). \end{aligned} \quad (\text{b.10})$$

Now, assume that $\lim_{x \rightarrow \infty} \frac{\hat{n}_x - \frac{x}{\kappa}}{x^\epsilon} = 0$ for some $\epsilon > 0$ (from (b.5), we already know that this holds for any $\epsilon > 1$). Then, since $|\hat{n}_x - \hat{n}_x| \leq (A - \frac{x}{\kappa}) + 1$, from (b.10) we have

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

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

Proposition B.3 Under condition (C1), $\lim_{x \rightarrow \infty} \langle \sigma_x^2 \rangle = \frac{S}{4\kappa}$.

Proof of Proposition B.3 : From (2.9), we have $\langle \sigma_x^2 \rangle = \frac{x \text{Var}\{X_{\hat{n}_x}\}}{(x + \kappa \hat{n}_x)^2} - \frac{1}{\kappa} \frac{\text{Var}\{X_{\hat{n}_x}\} \kappa \hat{n}_x}{x (1 + \frac{\kappa \hat{n}_x}{x})^2}$. However, we know that $\frac{\text{Var}\{X_{\hat{n}_x}\}}{\hat{n}_x} \rightarrow S$ (Proposition B.1(c)) and $\frac{\kappa \hat{n}_x}{x} \rightarrow 1$ (Proposition B.2), as $x \rightarrow \infty$. Thus, $\lim_{x \rightarrow \infty} \langle \sigma_x^2 \rangle = \frac{S}{4\kappa}$. *Q.E.D.*

C. Proofs for Main Results

Proof of Theorem 3.1 : To prove the theorem, it suffices to show that

$$\lim_{x \rightarrow \infty} \frac{\mathbb{P}(\{Y^{(x)}\}_{[\frac{x}{\alpha\kappa}, \frac{\alpha x}{\kappa}]^c} > \sqrt{x})}{\mathbb{P}(\{Y^{(x)}\} > \sqrt{x})} = 0$$

for all $\alpha > 1$, where $\mathbf{A}^c$ denotes the complementary set of $\mathbf{A}$.

Let $\alpha > 1$. Since $g(n) \rightarrow 1$ ($g(n)$ is defined in (b.1)) as $n \rightarrow \infty$, there exists an n_o such that $g(n) \leq \frac{\alpha+1}{2\sqrt{\alpha}}$ for all $n \geq n_o$. Now, let $G := \sup_{n \geq 0} g(n)$, then there exists an $x_o > \alpha\kappa n_o$ such that $\frac{Sx n_o G}{(x + \kappa n_o)^2} \leq \frac{S\sqrt{\alpha}}{2\kappa(\alpha+1)}$ for all $x \geq x_o$. Since $\frac{Sx n G}{(x + \kappa n)^2}$ is an increasing function for $n \leq \frac{x}{\kappa}$, this (in conjunction with (b.2)) implies that,

$$\sigma_{x,n}^2 \leq \frac{Sx n G}{(x + \kappa n)^2} \leq \frac{S\sqrt{\alpha}}{2\kappa(\alpha+1)} \quad \text{for all } x \geq x_o \text{ and } n \leq n_o. \quad (\text{c.1})$$

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

$$\sigma_{x,n}^2 = \frac{Sx n g(n)}{(x + \kappa n)^2} \leq \frac{S\sqrt{\alpha}}{2\kappa(\alpha+1)} \quad \text{for } x \geq x_o \text{ and } n \in (n_o, \frac{x}{\alpha\kappa}) \cup (\frac{\alpha x}{\kappa}, \infty). \quad (\text{c.2})$$

Now from (c.1) and (c.2), it follows that

$$\langle \sigma_x^2 \rangle_{[\frac{x}{\alpha\kappa}, \frac{\alpha x}{\kappa}]^c} \leq \frac{S\sqrt{\alpha}}{2\kappa(\alpha+1)} \quad \text{for all } x \geq x_o. \quad (\text{c.3})$$

We now define a pseudo-metric $d^{(x)}$ on $\{0, 1, 2, \dots\}$ as $d^{(x)}(n_1, n_2) := \sqrt{\mathbb{E}\{(Y_{n_2}^{(x)} - Y_{n_1}^{(x)})^2\}}$. Also, let $B_\epsilon^{(x)}(n) := \{m : d^{(x)}(n, m) \leq \epsilon\}$ 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 $\{0, 1, 2, \dots\}$. Since $\text{Var}\{Y_n^{(x)}\} \leq \frac{SGx n}{(x + \kappa n)^2} \leq \frac{SG}{4\kappa}$ and since $Y_0^{(x)} = 0$, $B_\epsilon^{(x)}(0)$ covers $\{0, 1, 2, \dots\}$ when $\epsilon \geq \sqrt{\frac{SG}{4\kappa}}$. Therefore, for all $x > 0$,

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

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

$$\begin{aligned} d^{(x)}(n_1, n_2) &= \sqrt{\mathbb{E}\left\{\left(\frac{\sqrt{x}(X_{n_2} + \kappa n_2)}{x + \kappa n_2} - \frac{\sqrt{x}(X_{n_1} + \kappa n_1)}{x + \kappa n_1}\right)^2\right\}} \\ &= \sqrt{\mathbb{E}\left\{\left(\frac{\sqrt{x}(X_{n_2} + \kappa n_2)}{x + \kappa n_2} - \frac{\sqrt{x}(X_{n_1} + \kappa n_1)}{x + \kappa n_2} + \frac{\sqrt{x}(X_{n_1} + \kappa n_1)}{x + \kappa n_2} - \frac{\sqrt{x}(X_{n_1} + \kappa n_1)}{x + \kappa n_1}\right)^2\right\}} \\ &\leq \sqrt{\mathbb{E}\left\{\left(\frac{\sqrt{x}(X_{n_2} - X_{n_1} + \kappa(n_2 - n_1))}{x + \kappa n_2}\right)^2\right\}} + \sqrt{\mathbb{E}\left\{\left(\frac{\kappa(n_2 - n_1)\sqrt{x}(X_{n_1} + \kappa n_1)}{(x + \kappa n_2)(x + \kappa n_1)}\right)^2\right\}} \\ &= \frac{\sqrt{x}}{x + \kappa n_2} \sqrt{\text{Var}\{(X_{n_2} - X_{n_1})\}} + \frac{\kappa(n_2 - n_1)\sqrt{x}}{(x + \kappa n_2)(x + \kappa n_1)} \sqrt{\text{Var}\{X_{n_1}\}}. \end{aligned} \quad (\text{c.5})$$

However, since $\text{Var}\{(X_{n_2} - X_{n_1})\} = \text{Var}\{X_{n_2 - n_1}\}$ from the stationary increment property of X_n , $\text{Var}\{(X_{n_2} - X_{n_1})\}$ and $\text{Var}\{X_{n_1}\}$ are bounded by $GS(n_2 - n_1)$ and GSn_1 , respectively. Hence, from (c.5)

$$d^{(x)}(n_1, n_2) \leq \frac{\sqrt{SGx(n_2 - n_1)}}{x + \kappa n_2} + \frac{\kappa(n_2 - n_1)\sqrt{SGxn_1}}{(x + \kappa n_1)(x + \kappa n_2)}$$

$$\begin{aligned}
&\leq \left(\frac{\sqrt{SGx}}{x + \kappa n_2} + \frac{\kappa \sqrt{SGx n_1 n_2}}{(x + \kappa n_1)(x + \kappa n_2)} \right) \sqrt{n_2 - n_1} \\
&\leq \left(\sqrt{\frac{SG}{x}} + \frac{1}{4} \sqrt{\frac{SG}{x}} \right) \sqrt{n_2 - n_1} \leq \sqrt{\frac{2SG}{x}} \sqrt{n_2 - n_1} \\
&\quad \text{(from the fact that } \frac{\sqrt{x}}{x + \kappa n_2} \leq \frac{1}{\sqrt{x}} \text{ and } \frac{\sqrt{n}}{(x + \kappa n)} \leq \frac{1}{2\sqrt{x\kappa}} \text{)}.
\end{aligned}$$

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

$$[n - \frac{x}{2SG}\epsilon^2, n + \frac{x}{2SG}\epsilon^2] \subset B_\epsilon^{(x)}(n). \quad (\text{c.6})$$

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

$$[\frac{SGx}{\epsilon^2 \kappa^2}, \infty) \subset B_\epsilon^{(x)}(0). \quad (\text{c.7})$$

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 (c.6) and (c.7), it follows that $\lceil \frac{SGx}{k\epsilon^2 \kappa^2} \rceil + 1$ $d^{(x)}$ -balls of radius of ϵ centered at ki ($i = 0, 1, \dots, \lceil \frac{SGx}{k\epsilon^2 \kappa^2} \rceil$) cover $\{0, 1, 2, \dots\}$. Hence, for $\epsilon < \sqrt{\frac{SG}{4\kappa}}$, $N^{(x)}(\epsilon)$ is bounded by the following inequality.

$$N^{(x)}(\epsilon) \leq \left\lceil \frac{SGx}{k\epsilon^2 \kappa^2} \right\rceil + 1 \leq \frac{2S^2 G^2}{\kappa^2 \epsilon^4} + 2. \quad (\text{c.8})$$

From (c.4) and (c.8), $\bar{N}(\epsilon)$ defined by

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

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

$$\mathbb{E}\{\langle Y^{(x)} \rangle\} \leq \mathcal{M}, \quad \text{for all } x > 0. \quad (\text{c.9})$$

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

$$\begin{aligned}
\mathbb{P}(\{\langle Y^{(x)} \rangle_{[\frac{x}{\alpha\kappa}, \frac{\alpha x}{\kappa}]^c} > \sqrt{x}\}) &\leq 2e^{-\frac{(\sqrt{x} - \mathbb{E}\{\langle Y^{(x)} \rangle_{[\frac{x}{\alpha\kappa}, \frac{\alpha x}{\kappa}]^c}\})^2}{2(\sigma^2)_{[\frac{x}{\alpha\kappa}, \frac{\alpha x}{\kappa}]^c}}} \\
&\leq 2e^{-\frac{\kappa(\sqrt{x} - \mathbb{E}\{\langle Y^{(x)} \rangle_{[\frac{x}{\alpha\kappa}, \frac{\alpha x}{\kappa}]^c}\})^2(\alpha+1)}{S\sqrt{\alpha}}} \\
&\quad \text{(from (c.3) and the fact that } \langle Y^{(x)} \rangle_{[\frac{x}{\alpha\kappa}, \frac{\alpha x}{\kappa}]^c} \leq \langle Y^{(x)} \rangle) \\
&\leq 2e^{-\frac{\kappa(\sqrt{x} - \mathcal{M})^2(\alpha+1)}{S\sqrt{\alpha}}} \quad \text{(from (c.9)),} \quad (\text{c.10})
\end{aligned}$$

for x sufficiently large. Therefore,

$$\limsup_{x \rightarrow \infty} \frac{1}{x} \log \mathbb{P}(\{\langle Y^{(x)} \rangle_{[\frac{x}{\alpha\kappa}, \frac{\alpha x}{\kappa}]^c} > \sqrt{x}\}) \leq \lim_{x \rightarrow \infty} -\frac{\kappa(\sqrt{x} - \mathcal{M})^2(\alpha+1)}{Sx\sqrt{\alpha}} = -\frac{\kappa(\alpha+1)}{S\sqrt{\alpha}}. \quad (\text{c.11})$$

Additionally, we know from [30] that

$$\lim_{x \rightarrow \infty} \frac{1}{x} \log \mathbb{P}(\{\langle Y^{(x)} \rangle > \sqrt{x}\}) = \lim_{x \rightarrow \infty} \frac{1}{x} \log \mathbb{P}(\{\langle X \rangle > x\}) = -\frac{2\kappa}{S}. \quad (\text{c.12})$$

Since $-\frac{\kappa(\alpha+1)}{S\sqrt{\alpha}} < -\frac{2\kappa}{S}$ for all $\alpha > 1$, (c.11) and (c.12) imply that

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

and the theorem follows. Q.E.D.

Proof of Theorem 3.2 : Let B_t be the standard Brownian motion process and define a centered Gaussian process $Z_n^{(x)}$ ($n = 0, 1, \dots$) for each $x > 0$ by

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

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

$$C_{Z^{(x)}}(n_1, n_2) = \frac{Sx \min\{n_1, n_2\} \sqrt{g(n_1)g(n_2)}}{(x + \kappa n_1)(x + \kappa n_2)} \quad (\text{c.13})$$

From (b.2) and (c.13), we can see that the variance of $Z_n^{(x)}$ is equal to that of $Y_n^{(x)}$. Now, let $\alpha > 1$. From Proposition B.1(e), there exists an $n_o > 0$ such that for all $n \geq n_o$,

$$\frac{\text{Var}\{X_m\}}{m} \leq \frac{\text{Var}\{X_n\}}{n} \quad \text{for all } m < n. \quad (\text{c.14})$$

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

$$\begin{aligned} \frac{C_X(n_1, n_2)}{n_1} &= \frac{1}{2n_1} (\text{Var}\{X_{n_1}\} + \text{Var}\{X_{n_2}\} - \text{Var}\{X_{n_2-n_1}\}) \quad (\text{from Proposition B.1(b)}) \\ &= \frac{1}{2} \left(\frac{\text{Var}\{X_{n_1}\}}{n_1} + \frac{\text{Var}\{X_{n_2}\}}{n_2} + \frac{n_2-n_1}{n_1} \left(\frac{\text{Var}\{X_{n_2}\}}{n_2} - \frac{\text{Var}\{X_{n_2-n_1}\}}{n_2-n_1} \right) \right) \\ &\geq \frac{1}{2} \left(\frac{\text{Var}\{X_{n_1}\}}{n_1} + \frac{\text{Var}\{X_{n_2}\}}{n_2} \right) \quad (\text{from (c.14)}) \\ &\geq \sqrt{\frac{\text{Var}\{X_{n_1}\}\text{Var}\{X_{n_2}\}}{n_1 n_2}} \quad (\text{since } \frac{\text{Var}\{X_n\}}{n} \geq 0). \end{aligned}$$

This implies that

$$S \min\{n_1, n_2\} \sqrt{g(n_1)g(n_2)} = n_1 \sqrt{\frac{\text{Var}\{X_{n_1}\}\text{Var}\{X_{n_2}\}}{n_1 n_2}} \leq C_X(n_1, n_2). \quad (\text{c.15})$$

Therefore, from (2.8), (c.13), and (c.15), 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 $\text{Var}\{Y_n^{(x)}\} = \text{Var}\{Z_n^{(x)}\}$, we have $\mathbb{E}\{(\langle Y^{(x)} \rangle - \langle Z^{(x)} \rangle)^2\} \leq \mathbb{E}\{(\langle Y^{(x)} \rangle - \langle Z^{(x)} \rangle)^2\}$ for all $n_1, n_2 \in [\frac{x}{\alpha\kappa}, \frac{\alpha x}{\kappa}]$. Therefore, from Theorem A.2,

$$\mathbb{P}(\{\langle Y^{(x)} \rangle_{[\frac{x}{\alpha\kappa}, \frac{\alpha x}{\kappa}]} > \sqrt{x}\}) \leq \mathbb{P}(\{\langle Z^{(x)} \rangle_{[\frac{x}{\alpha\kappa}, \frac{\alpha x}{\kappa}]} > \sqrt{x}\}) \quad \text{for all } x \geq \alpha\kappa n_o. \quad (\text{c.16})$$

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

$$\begin{aligned} \mathbb{P}(\{\langle Z^{(x)} \rangle_{[\frac{x}{\alpha\kappa}, \frac{\alpha x}{\kappa}]} > \sqrt{x}\}) &= \mathbb{P}(\{Z_n^{(x)} > \sqrt{x} \text{ for any } n \in [\frac{x}{\alpha\kappa}, \frac{\alpha x}{\kappa}]\}) \\ &= \mathbb{P}(\{\sqrt{Sg(n)}B_n > x + \kappa n \text{ for any } n \in [\frac{x}{\alpha\kappa}, \frac{\alpha x}{\kappa}]\}) \\ &\quad (\text{from the definition of } Z_n^{(x)}) \end{aligned}$$

$$\begin{aligned}
&\leq \mathbb{P}(\{\sqrt{Sg(\lceil \frac{\alpha x}{\kappa} \rceil)} B_n > x + \alpha n \text{ for any } n \in [\frac{x}{\alpha \kappa}, \frac{\alpha x}{\kappa}]\}) \\
&\quad (\text{since } g(n) \text{ is increasing on } [\frac{x}{\alpha \kappa}, \frac{\alpha x}{\kappa}] \text{ from (c.14)}) \\
&\leq \mathbb{P}(\{\sqrt{Sg(\lceil \frac{\alpha x}{\kappa} \rceil)} B_t > x + \kappa t \text{ for any } t \in [0, \infty)\}) \\
&= e^{-\frac{2\kappa x}{Sg(\lceil \frac{\alpha x}{\kappa} \rceil)}} \quad (\text{see [43, page 191]}). \tag{c.17}
\end{aligned}$$

From (c.16) and (c.17), we have an asymptotic upper bound to $\mathbb{P}(\{\langle Y^{(x)} \rangle_{[\frac{x}{\alpha \kappa}, \frac{\alpha x}{\kappa}]} > \sqrt{x}\})$

$$\mathbb{P}(\{\langle Y^{(x)} \rangle_{[\frac{x}{\alpha \kappa}, \frac{\alpha x}{\kappa}]} > \sqrt{x}\}) \leq e^{-\frac{2\kappa x}{Sg(\lceil \frac{\alpha x}{\kappa} \rceil)}} \quad \text{for all } x \geq \alpha \kappa n_o. \tag{c.18}$$

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

$$\begin{aligned}
\frac{2\kappa x}{S} - \frac{x}{Sg(\lceil \frac{\alpha x}{\kappa} \rceil)} &= -\frac{2\kappa x \left(1 - \frac{\text{Var}\{X_{\lceil \frac{\alpha x}{\kappa} \rceil}\}}{S^2 \lceil \frac{\alpha x}{\kappa} \rceil}\right)}{Sg(\lceil \frac{\alpha x}{\kappa} \rceil)} \quad (\text{from the definition of } g(t)) \\
&= -\frac{\lceil \frac{\alpha x}{\kappa} \rceil \left(S - \frac{\text{Var}\{X_{\lceil \frac{\alpha x}{\kappa} \rceil}\}}{\lceil \frac{\alpha x}{\kappa} \rceil}\right)}{S^2 \lceil \frac{\alpha x}{\kappa} \rceil g(\lceil \frac{\alpha x}{\kappa} \rceil)} \rightarrow -\frac{2\kappa^2 D}{\alpha S^2} \text{ as } x \rightarrow \infty. \tag{c.19}
\end{aligned}$$

Therefore, from (2.6), and from Theorem 3.1, (c.18) and (c.19), it follows that

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

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

Proof of Proposition 4.1 : From (2.9) and the definition of $\hat{n}_x$, we have $\langle \sigma_x^2 \rangle = \frac{x \text{Var}\{X_{\hat{n}_x}\}}{(x + \kappa \hat{n}_x)^2}$. Hence,

$$\frac{2\kappa x}{S} - \frac{x}{2\langle \sigma_x^2 \rangle} = \frac{-4\kappa \frac{x}{\hat{n}_x} \hat{n}_x \left(S - \frac{\text{Var}\{X_{\hat{n}_x}\}}{\hat{n}_x}\right) - \frac{S\kappa^2}{\hat{n}_x} \left(\frac{x}{\kappa} - \hat{n}_x\right)^2}{2S \frac{\text{Var}\{X_{\hat{n}_x}\}}{\hat{n}_x}} \tag{c.20}$$

Since $\frac{x}{\hat{n}_x} \rightarrow \kappa$, $\frac{\text{Var}\{X_{\hat{n}_x}\}}{\hat{n}_x} \rightarrow S$, $(S - \frac{\text{Var}\{X_{\hat{n}_x}\}}{\hat{n}_x})\hat{n}_x \rightarrow D$, and $\frac{(x/\kappa - \hat{n}_x)^2}{\hat{n}_x} \rightarrow 0$ as $x \rightarrow \infty$ from Proposition B.1 and Proposition B.2, it follows from (c.20) that

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

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

List of References

- [1] J. Abate, G. L. Choudhury, and W. Whitt, "Asymptotics for steady-state tail probabilities in structured Markov queueing models," *Comm. Statist. Stochastic Models*, vol. 10, no. 1, pp. 99–143, 1994.
- [2] 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.
- [3] R. G. Addie and M. Frater, "Loss forecasting by means of Gaussian models of video traffic," in *Proceedings of 8th ATRS*, (Melbourne, Australia), Dec. 1993.
- [4] R. G. Addie and M. Zukerman, "An Approximation for Performance Evaluation of Stationary Single Server Queues," *IEEE Transactions on Communications*, vol. 42, pp. 3150–3160, Dec. 1994.
- [5] R. G. Addie, M. Zukerman, and T. Neame, "Fractal Traffic: Measurements, Modeling, and Performance Evaluation," in *Proceedings of IEEE INFOCOM*, pp. 977–984, 1995.
- [6] R. Adler and G. Samorodnitsky, "Tail Behaviour for the Supremum of Gaussian Processes with Applications to Empirical Processes," *The Annals of Probability*, no. 4, pp. 1339–1351, 1987.
- [7] R. J. Adler, *An Introduction to Continuity, Extrema, and Related Topics for General Gaussian Processes*. Hayward, CA: Institute of Mathematical Statistics, 1990.
- [8] D. Anick, D. Mitra, and M. M. Sondhi, "Stochastic Theory of a Data Handling System with Multiple Sources," *Bell Technical Journal*, vol. 61, pp. 1871–1894, Oct. 1982.
- [9] S. Asmussen, "On Cycle Maxima, First Passage Problems and Extreme Value Theory for Queues," *Stochastic Models*, vol. 8, no. 3, pp. 163–179, 1992.
- [10] J. Beran, R. Sherman, M. S. Taqqu, and W. Willinger, "Long-range dependence in variable-bit-rate video traffic," *IEEE Transactions on Communications*, vol. 43, pp. 1566–1579, Feb.–April 1995.
- [11] S. M. Berman and N. Kono, "The Maximum of a Gaussian Process with Nonconstant Variance: A Sharp Bound for The Distribution Tail," *The Annals of Probability*, vol. 17, no. 2, pp. 632–650, 1989.
- [12] D. D. Botvich and N. G. Duffield, "Large deviations, the shape of the loss curve, and economies of scale in large multiplexers," *Queueing Systems*, vol. 20, pp. 293–320, 1995.
- [13] P. Bratley, B. L. Fox, and L. E. Schrage, *A Guide to Simulation*. New York: Springer-Verlag, second ed., 1987.
- [14] C.-S. Chang, "Stability, Queue Length, and Delay of Deterministic and Stochastic Queueing Networks," *IEEE Transactions on Automatic Control*, vol. 39, pp. 913–931, May 1994.
- [15] 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, Aug. 1995.
- [16] 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.
- [17] J. Choe and N. B. Shroff, "A New Method to Determine the Queue Length Distribution at an ATM Multiplexer," in *Proceedings of IEEE INFOCOM*, pp. 550–557, 1997.
- [18] J. Choe and N. B. Shroff, "On the supremum distribution of integrated stationary Gaussian processes with negative linear drift," *Advances in Applied Probability*, March 1999. to appear.

- [19] G. L. Choudhury, D. M. Lucantoni, and W. Whitt, "Squeezing the Most Out of ATM," *IEEE Transactions on Communications*, vol. 44, pp. 203–217, Feb. 1996.
- [20] J. N. Daigle and J. D. Langford, "Models for Analysis of Packet Voice Communication Systems," *IEEE Journal on Selected Areas in Communications*, vol. 4, pp. 847–855, Sep. 1986.
- [21] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications*. Boston: Jones and Bartlett, 1993.
- [22] L. Donatiello and R. Nelson, *Models and Techniques for Performance Evaluation: of Computer and Communication Systems*. Springer-Verlag, 1993.
- [23] N. G. Duffield, "Exponential bounds for queues with Markovian arrivals," *Queueing Systems*, vol. 17, pp. 413–430, 1994.
- [24] N. G. Duffield and N. O'Connell, "Large deviations and overflow probabilities for the general single server queue, with application," *Proc. Cambridge Philos. Soc.*, vol. 118, pp. 363–374, 1995.
- [25] 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, Aug. 1995.
- [26] 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, Aug. 1995.
- [27] 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.
- [28] 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.
- [29] W. Feller, *An Introduction to Probability Theory and its Applications I*. New York: John Wiley & Son, 1968.
- [30] P. W. Glynn and W. Whitt, "Logarithmic asymptotics for steady-state tail probabilities in a single-server queue," *Journal of Applied Probability*, pp. 131–155, 1994.
- [31] 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. 968–981, Sep. 1991.
- [32] P. R. Jelenkovic, A. A. Lazar, and N. Semret, "The Effect of Multiple Time Scales and Subexponentiality in MPEG Video Streams on Queueing Behavior," *IEEE Journal on Selected Areas in Communications*, vol. 15, August 1997. to appear.
- [33] Y. Jenq, "Approximations for Packetized Voice Traffic in Statistical Multiplexer," in *Proceedings of IEEE INFOCOM*, pp. 256–259, 1984.
- [34] F. Kelly, "Effective Bandwidth in Multiclass Queues," *Queueing Systems*, vol. 9, pp. 5–16, 1991.
- [35] G. Kesidis, J. Walrand, , and C.-S. Chang, "Effective Bandwidth for Multiclass Markov Fluid and other ATM Sources," *IEEE/ACM Transactions on Networking*, vol. 1, pp. 424–438, Aug. 1993.

- [36] E. Knightly, "H-BIND: A new approach to providing statistical performance guarantees to VBR traffic," in Proceedings of IEEE INFOCOM, pp. 1091–1099, 1996.
- [37] W. E. Leland, M. Taqqu, W. Willinger, and D. V. Wilson, "On the Self-similar Nature of Ethernet Traffic (Extended Version)," *IEEE/ACM Transactions on Networking*, vol. 2, pp. 1–15, Feb. 1994.
- [38] 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.
- [39] D. M. Lucantoni, "New Results on the Single Server Queue with A Batch Markovian Arrival Process," *Stochastic Models*, vol. 7, pp. 1–46, 1991.
- [40] I. Norros, "A storage model with self-similar input," *Queueing Systems*, vol. 16, pp. 387–396, 1994.
- [41] I. Norros, "On the Use of Fractal Brownian Motion in the Theory of Connectionless Networks," *IEEE Journal on Selected Areas in Communications*, vol. 13, pp. 953–962, Aug. 1995.
- [42] O. Rose, "Statistical properties of MPEG video traffic and their impact on traffic modeling in ATM systems," in Proceedings of the 20th Conference on Local Computer Networks, pp. 397–406, Minneapolis, MN, Oct. 1995.
- [43] S. M. Ross, *Stochastic Processes*. New York: John Wiley & Son, 1983.
- [44] J. S. Sadowsky, "Large Deviations Theory and Efficient Simulation of Excessive Backlogs in a GI/GI/m Queue," *IEEE Transactions on Automatic Control*, vol. 36, pp. 1383–1394, Dec. 1991.
- [45] G. Samorodnitsky, "Probability tails of Gaussian Extrema," *Stochastic Processes and their Applications*: vol. 38, no. 1, pp. 55–84, 1991.
- [46] N. Shroff and M. Schwartz, "Improved Loss Calculations at an ATM Multiplexer," in Proceedings of IEEE: INFOCOM, vol. 2, pp. 561–568, March 1996.
- [47] N. B. Shroff, "Traffic Modeling and Analysis in High Speed ATM Networks," PhD thesis, Graduate School of Arts and Sciences, Columbia University, New York, NY, 1995.
- [48] A. Simonian, "Stationary Analysis of a Fluid Queue with Input Rate Varying as an Ornstein-Uhlenbeck Process," *SIAM Journal on Applied Mathematics*, vol. 51, pp. 828–842, 1991.
- [49] A. Simonian, "Transient and Stationary Distributions for Fluid Queues and Input Processes with a Density," *SIAM Journal on Applied Mathematics*, vol. 51, pp. 1732–1739, 1991.
- [50] K. Sriram and W. Whitt, "Characterizing Superposition Arrival Processes in Packet Multiplexer for Voice and Data," *IEEE Journal on Selected Areas in Communications*, vol. 4, pp. 833–846, Sep. 1986.
- [51] L. Takacs, *Combinatorial Methods in the Theory of Stochastic Processes*. New York: John Wiley & Son, 1967.
- [52] M. Titlagrand, "Small tails for the supremum of a Gaussian process," *Ann. Inst. Henri Poincaré*, vol. 24, pp. 307–315, 1988.
- [53] E. Wong and B. Hajek, *Stochastic Processes in Engineering Systems*. New York: Springer-Verlag, 1985.