INVESTIGATION OF DELAY JITTER OF HETEROGENEOUS TRAFFIC IN BROADBAND NETWORKS

By

HOOI MIIN SOO

Bachelor of Science in Electrical Engineering
Oklahoma State University
Stillwater, Oklahoma
2000

Master of Science in Electrical Engineering
Oklahoma State University
Stillwater, Oklahoma
2003

Submitted to the Faculty of the
Graduate College of the
Oklahoma State University
in partial fulfillment of
the requirements for
the Degree of
DOCTOR OF PHILOSOPHY
in Electrical Engineering
Dec 2006

INVESTIGATION OF DELAY JITTER OF
HETEROGENEOUS TRAFFIC IN
BROADBAND NETWORKS

Thesis Approved:

Dr. Jong-Moon Chung
Thesis Advisor

Dr. Keith A. Teague

Dr. Charles F. Bunting

Dr. Weili Zhang

Dr. Venkatesh Sarangan

Dr. A. Gordon Emslie
Dean of the Graduate Collage

PREFACE

There is a great demand for wired and wireless network architectures to support a variety of applications with very high speed communication. Nowadays, the pace of our business and social lives has created a great demand for not only instant transferring and accessing various kinds of data by a click on the mouse (i.e. digitized voice and video, file transfers, e-commerce, transactions, e-mail, telnet, web browsing and multicast), but also for reliable and better quality of Internet service. As a result, quality of service (QoS) currently becomes an issue of concern for wide area network (WAN) telecommunications providers or backbone carriers. For instance, what is the required bandwidth for the local area network (LAN) or wide area network (WAN) to provide sufficient capacity for video transferring with desired minimum level jitter?

A critical challenge for both wired and wireless networking vendors and carrier companies is to be able to accurately estimate the quality of service (QoS) that will be provided based on the network architecture, router/switch topology, and protocol applied. The variation in QoS performance based on the priority assignment is of significant importance, due to the fact that the differentiated services (DiffServ) capable networks should be able to effectively control QoS through classifying traffic according to desired service level (prioritize the data flows) and marking the packets so that the routers can recognize the prioritized packet. We need a tool to investigate whether the packet prioritization could truly reduce the delay jitter or not.

Therefore, the main objective of this study is to provide a theoretical analysis of delay jitter, which investigates the issues of traffic jitter characteristic in the homogenous wireless communication and heterogeneous wired networks deploying different priority scheduling algorithms.

Although simulation can be a powerful tool, an engineer must understand the mathematical model underlying each program model to be applied correctly. In addition, one may not afford an extensive simulation, which requires an extensive data mining process. On the other hand, a mathematical formula is much easier to use compared to using a simulation model. Thus, the efficient way is to use a simulation tool in combination with a probably less realistic mathematical/analytic model.

Plan of this thesis

This thesis is organized as follows. In Chapter 1, an overview of this thesis is provided, which states why the proposed methodology is important and how it may contribute to designing and planning of internetworking system. Literature review on WAN networks and Internet traffic is given in Chapter 2, where the challenge of the heterogeneous traffic modeling compared to the homogeneous traffic modeling is discussed. The understanding of the WAN networks characteristics is needed to construct the analytic model. Chapter 3 provides mathematical derivations for the per-class jitter characteristic of a wireless network with automatic repeat request (ARQ) error control and head-of-line (HOL) priority scheduling. The derivations provide a direct method to analyze/evaluate the per-class jitter based on the DiffServ network and protocol parameters. In Chapter 4, the Gaussian traffic modeling combining with Maximum

Variance Approach to conduct the queue length analysis, which further apply in heterogeneous traffic jitter analysis of multiple priority queues..

The conclusion of this thesis is in Chapter 5, which summarizes the essential elements in order to provide guidelines for network designers trying to meet engineering specifications of performance deliverables. The references list is provided in Reference section.

The main contributions of this dissertation are as follow:

1. **Chapter 3**
   - Derive mathematical equations, which to be able to provide a mathematical tool to analyze the homogeneous traffic in wireless networks with adequate accuracy.
   - Propose a call admission control (CAC) scheme to improve the jitter performance.
2. **Chapter 4**
   - Derive mathematical equations, which to be able to provide mathematical tool to analyze the heterogeneous traffic in wired broadband networks with adequate accuracy. Simulation is done to show the accuracy of the mathematical model.
   - Propose an adaptive distortion-reducing peak output-rate enforcing (APORE) scheme to control/improve the jitter performance.

ACKNOWLEDGMENTS

First, I wish to express my heartfelt thanks to my major advisor, Dr. Jong-Moon Chung, for his encouragement, guidance, valuable discussions and advice throughout this work. I extend my sincere appreciation to my other committee members Dr. Weili Zhang, Dr. Charles F. Bunting and Dr. Venkatesh Sarangan, for their guidance. I extend my genuine gratitude to Dr. Keith A. Teague for serving as my dissertation defense chair.

I wish to express my special appreciation to my family, my parent, for their unlimited and continuing love, encouragement, support and sacrifice. Moreover, I am sincerely thankful to my husband, Mr. Lee who has been always wonderfully encouraging me through out this work. Without them, I would not have come this far.

Last but not the least, I would like to thank all that have assisting me in accomplishing this work.

TABLE OF CONTENTS

Chapter **Page**

Chapter 1 1

1.1 Overview 1

1.2 Delay Jitter Analysis Methodology 6

1.3 Queueing and Teletraffic Theory Applied in High-Speed Wide Area Networks (WANs) with Heterogeneous Type Traffic 8

1.4 Approximations for the Queue Length Distribution Gaussian Model for Broadband Traffic 15

1.4.1 Single-Asymptotic Upper Bound 16

1.4.2 Maximum Variance Asymptotic (MVA) 17

1.4.3 MVA Lower Bound 18

1.4.4 MVA Upper Bound 18

1.4.5 MVA general approach (not bound) 19

1.4.6 Long Range Dependent Gaussian model: The Discrete Gaussian Model 20

1.4.7 The Continuous Gaussian Model 21

1.4.8 Large two-stage models: Heavy Traffic 22

Chapter 2 23

2.1 Introduction: Analyzing traffic traces 23

2.2 Aggregated Variance Method 23

2.2.1 Description and Methodology 23

2.2.1 Simulation Results 26

2.2.3 Simulation Flow Chart 28

2.3 R/S Method 29

2.3.1 Description and Methodology 29

2.3.2 Simulation Results 30

2.3.3 Simulation Flow Chart 32

2.4 Estimating the Hurst Exponent using Wavelet Spectral Density 33

2.4.1 Description and Methodology 33

2.4.2 Simulation Results 35

2.4.3 Simulation Flow Chart 36

2.5 Observation and Discussion 37

Chapter 3 39

3.1 Abstract 39

3.2 Introduction 39

3.3 Analytical Model 40

3.3.1 Network and System Model Assumptions 40

3.3.2 Theoretical Formulations ... 42
3.3.3 Jitter Analysis with Non-Priority ARQ Traffic ... 45
3.3.4 Jitter Analysis with Priority ARQ Traffic ... 46
3.4 Discussions and Conclusions ... 49
Chapter 4 ... 51
4.1 Introduction ... 53
4.2 Discrete Time Fluid Queue Model and Extreme Value [26] ... 55
4.3 Jitter Analysis of Heterogeneous Traffic Networks ... 65
4.3.1 Simulation and Numerical Study ... 67
4.4 Jitter Analysis of Heterogeneous Traffic Networks with Differentiated Services ...72
4.4.1 Simulation and Numerical Study ... 78
Chapter 5 ... 81
Conclusions...81
Reference ...83

LIST OF FIGURES

Figure **Page**

Fig. 1.1 The actual Ethernet traffic trace in original time scale (time unit = 1ms). 13
Fig. 1.2 The Ethernet traffic trace is aggregated 10 folds (time unit = 0.01s). 13
Fig. 1.3 Original Ethernet traffic.. 14
Fig. 1.4 Aggregated 100 Ethernet traffic flows. 14
Fig. 2.1 The relationship between aggregated variance equation and the least squares line. 25
Fig 2.2 This figure shows that Ethernet traffic is very self-similar, which is indicated by a fairly large *H* value.. 27
Fig 2.3 This figure shows that Star Wars IV data stream is self-similar, and the level of self-similarity is indicated by the H value. 27
Fig 2.4 This figure shows that the data set is self-similar, and the level of self-similarity is indicated by the H value. Here, K = 10 and upto $\log_{10}$(d)=3.4 is used. 31
Fig 2.5 This figure shows that Star Wars IV data stream is self-similar, and the level of self-similarity is indicated by the H value. K=10 and up to $\log_{10}$(d)=3.4 is used..... 31
Fig 2.6 The figure shows that Ethernet traffic is self-similar, and the level of self-similarity is indicated by the H value. 35
Fig 2.7 The figure shows that Star Wars IV data stream is self-similar, and the level of self-similarity is indicated by the H value. 35
Fig 2.8 The comparison between the three methods is shown, where the level of self-similarity is indicated by the H value. The estimation range is between (0.82, 0.93) for Ethernet traffic for low m aggregated level, where all the three methods have proved that the Ethernet traffic is very self-similar. 38
Fig. 3.1 Comparison between wireless networks deploying the priority and non-priority ARQ scheme with Pe = 10%. The solid line represents the non-priority ARQ scheme, and the dotted line represents the priority ARQ scheme. 47
Fig. 3.2 Comparison between wireless networks non-priority ARQ scheme with Pe = 20%. 48
Fig. 3.3 Comparison between wireless networks deploying the priority and non-priority ARQ scheme with Pe = 1% and 10% for class 1 through class 5........................... 48
Fig. 4.1 Plot of variance of Z_m versus *m*. The variance reaches a maximum value and then begins to converge to 0. The parameters used are *q*=1000, μ=0.61, *N*=100, each individual arrival process with mean $\overline{X}_s$= 10 cells/sec, and auto-covariance $C_s(l) = 100e^{-0.04l}$ 64

Fig. 4.2 Comparison between the approximation equation and the simulation results of the jitter distribution for utilization of 50% and 90% with 95% CI, for the case of *T* = 10. 70
Fig. 4.3 Distribution of jitter for the comparison between utilization of 50% and 70%, and for the comparison between *T* = 10 and 20. 71
Fig. 4.4 Distribution of jitter for utilization of 60%, *T*=10, and various A_{min} values with 95% CI, and *T* = 10. 79

LIST OF FLOWCHARTS

Figure **Page**

Flowchart 1 The simulation algorithm used to calculate delay jitter at a multiplexer ……………………………………………………….... 70

Chapter 1

Introduction

1.1 Overview

Delay jitter is one of the most important parameters in quality of service (QoS) that indicates the level of variation of traffic services of real-time data transfer. In real-time wired/wireless data communication, delay jitter is defined as the difference in time delay of the arrival at the destination that each transmitted packet experiences commonly, which is also simply referred to as jitter.

Delay jitter has been successfully analyzed for the performance of asynchronous transfer mode (ATM) network applications, in order to accurately estimate the quality of service (QoS) in terms of jitter that will be provided based on the network architecture, router/switch topology, and protocol applied.

Several papers [2], [3], [4], and [5] have successfully analyzed the performance of the delay jitter in order to be more easily, consistently, and effectively implemented in various ATM network applications. In [2], the analysis of delay jitter was conducted on voice traffic using an exponential On/Off source in cellular wireless ATM network. In [3], the jitter bound for various classes of constant bit rate (CBR) traffic is analyzed, where the performance bound is used in the call/connection admission control (CAC) mechanism. The multiple access control topology applied in [3] is based on a non-preemptive priority polling scheme [4].

The papers of [6], [7], and [8] address delay jitter for more general type of wireless networks, rather than ATM. In [6], the time division multiple access (TDMA)/time division duplex (TDD) scheduling scheme is used. In [7], jitter is analyzed for mobile ad hoc networks (MANET) based on the ad hoc on-demand distance vector (AODV) routing protocol, where the authors propose a novel handover mechanism. Both [6] and [7] provide results on jitter performance analysis based on computer simulation for the wireless network models. In [8], the authors provide a jitter analysis on wireless networks involving automatic retransmission request (ARQ) error recovery, where the delay jitter is calculated using the window-length generating function and the numerical results are verified through simulation. In [9], the jitter of homogeneous traffic in reference to the different priority classes has been analyzed. The delay jitter research of [9] extends the results of [1] and [10] for DiffServ networks, but all three papers do not consider the channel error probability or the ARQ error control scheme, which makes the results difficult to apply to wireless networking applications. Therefore, Chapter 3 the mathematical derivations to analyze the delay jitter performance of homogenous wireless networks that apply ARQ error recovery with time constraints have been developed, which resulted in [11]. In [11], we attempt to provide a novel analysis of the per-class jitter performance of DiffServ networks based on wireless channels that experience packet errors, assuming a non-preemptive head-of-the-line (HOL) priority scheme. The derivations provide a direct method to analyze/evaluate the per-class jitter based on the DiffServ network, retransmission time constraints, and network packet error parameters

[12]. In addition, we evaluated the effects on the delay jitter in reference to the priority control scheme of the ARQ traffic for the two cases of: 1) the ARQ traffic has a priority over the original transmission traffic; and 2) the ARQ traffic has no priority over the original transmission traffic.

As the next step, this homogenous jitter study is extended to heterogeneous wired and wireless network jitter analysis by deriving a general approach, which is going to be covered in this dissertation. In general, WAN traffic is often characterized by a diverse mix of heterogeneous data streams (e.g., multimedia, digitized voice, Internet, video, teleconferencing and many newly evolved applications), the periodicity of the individual streams is different from each other. In other words, in the heterogeneous environment, the input traffic streams are a superposition of data packet streams from different data sources with different periods. To further investigate heterogeneous jitter characteristics, first, a study is conducted on wired networks, and then the focus will be extended to wireless networks, which requires a longer time period due to its complexity. The studies that have been conducted and completed (but not limited to) in this dissertation are as follows:

- Various possible traffic modeling techniques have been investigated
- Diverse statistical properties of heterogeneous wireless networks is evaluated.
- The traffic modeling combined with heterogeneous jitter analysis.

Traffic analysis can assist one in developing a good traffic model for analyzing the jitter characteristic of heterogeneous wireless networks, where the results could serve as a tool to aid the network engineers and developers in wireless and wired network planning and architecture structuring.

The models, such as Markov Modulated Poisson Process (MMPP) and M/M/1/B have been applied in wireless network traffic modeling. In wired or wireless networks, the Poisson arrival may not be an appropriate model since actual networks traffic is very bursty in nature due to its long-range dependency (LRD) and self-similar characteristics.

Numerous studies have found that aggregated traffic at a node (router, switch for wired networks case) or base station (for wireless networks case) exhibit LRD, and such traffic can be very bursty. Thus, many broadband traffic studies have evolved around Gaussian, and they show that Gaussian models indeed provide a good approximation to networks traffic if the aggregated traffic is sufficiently large [13],[14]. Hence, heterogeneous wireless network traffic processes can be modeled as a Gaussian process, where some of these models have been discussed in Chapter 4. In extension of the MVA approach, we can derive the MVA lower bound and upper bound to provide a boundary conditions on the traffic. Another popular approach in traffic modeling is by using Fractional Gaussian Noise (fGN), which is a stationary Gaussian process with LRD. Recent research [15][16][17] models the heterogeneous networks traffic in the wavelet domain, where self-similar traffic exhibits long-range-dependent (LRD) correlations and non-Gaussian marginal distribution [15]; it shows that the aggregated traffic can be decomposed into those two groups [15].

In conclusion, the following is the summary of the work conducted in this dissertation:

1. Investigate the characteristics of the wireless networks that applying retransmission request (ARQ) with homogeneous type traffic.

2. Investigate the issues of traffic jitter characteristic in heterogeneous wired communication networks deploying different scheduling algorithms.
3. The emphasis of this dissertation is to investigate the various possible traffic modeling techniques and evaluate the challenges in characterizing the diverse statistical properties of heterogeneous wireless networks.
4. Apply the Gaussian traffic modeling using MVA approach to conduct the queue length analysis, which will be further used in heterogeneous jitter analysis.
5. Analyze the difference between jitter probability of multiple priority queues and servers.

The contributions of this dissertation are as follow:

1. Evaluate the characteristics of the wireless networks that applying retransmission request (ARQ) with homogeneous type traffic.
2. Evaluate the characteristics of the diverse statistical properties of wired networks with heterogeneous type traffic.
3. Investigate various possible traffic modeling techniques.
4. Apply the Gaussian traffic modeling using the MVA approach to analyze the queue length, and further generalize the heterogeneous jitter analysis.
5. Analyze the jitter process and the jitter probability distribution of multi priority class traffic.
6. Propose a scheme to control and improve networks performance in term of jitter.

In order to do these, we have to first:

- Study the traffic characteristics
- Investigate the traffic modeling techniques

Then, we'll be able to analyze the jitter process, and propose a scheme to control it.

1.2 Delay Jitter Analysis Methodology

In real-time wired or wireless communications, each transmitted packet may experience a different time delay during arrival at the destination. This variation in delay is often referred to as delay jitter or simply jitter.

For instance, let d_n represent the packet transmission delay experienced by the *n*th packet from a tagged stream. Thus, the process of delay jitter between the *n*th and (*n*+1)th packet of a tagged stream with respect to data stream original period (inter-departure time between *n*th and (*n*+1)th packets transmitted from the source), *T*, is

$$\widetilde{J}_n = d_{n+1} - d_n, \qquad \forall n .$$

(1.1)

In the steady-state limit $\lim_{n=\infty} \overline{d}_{n+1} = \overline{d}_n$, therefore jitter is a zero mean random sequence. Another interpretation of jitter is that it represents the difference in waiting times experienced by successive packets. Hence, we can relate the transmission delay and the queue length (number of packets) ahead of the tagged packet (i.e., Q_n represents the number of customers ahead of the *n*th tagged packet). Let any customer's service time be equal to one slot and follows first-in-first-out (FIFO) order, $Q_n = d_n$ (in time slot units). Therefore, (1.1) implies that

$$\widetilde{J}_n = Q_{n+1} - Q_n, \quad n = 1, 2, \dots \ . \tag{1.2}$$

To prove this, let the variable Q_n denote the number of customers, regardless of their class affiliation, that are waiting for service just ahead of the arrival of *n*th packet of the reference user. Thus, the *n*th packet of the reference user begins to enter service at time

$$\tau_n = t_n + Q_n \ . \tag{1.3}$$

Hence,

$$J_n = \tau_{n+1} - \tau_n = (t_{n+1} - t_n) + (Q_{n+1} - Q_n) = T + \Delta Q(n) \, . \tag{1.4}$$

The normalized jitter with respect to the original data stream period (inter-departure time between *n*th and (*n*+1)th packets transmitted from the source), *T* can be expressed as::

$$\widetilde{J}_n = J_n - T = \Delta Q(n), \quad n = 1, 2, \ldots, \text{ where } \Delta Q(n) \overset{def}{=} Q_{n+1} - Q_n \ . \tag{1.5}$$

From (1.5), we can observe the study of jitter can be equivalently changed to the study of the queue size variation $\Delta Q(n)$ of *p*-customer arrivals, which is the methodology that is going to be applied in this study. We want to find the pmf (probability mass function) of $\widetilde{J}_n$, i.e., $f\{\widetilde{J}_n\}$, from the queue length distribution $f\{Q_n\}$. A discrete-time approach is used in the delay analysis, where time is slotted. It is assumed that the packets arrive and depart at the beginning of a slot, which leads to a discrete time queueing process.

Next, the various traffic modeling techniques are discussed. First, let's denote the heterogeneous network of steady state queue length distribution as the probability of queue size larger than *x*, i.e., $P(Q > x)$. Here, the input traffic at a high-speed multiplexer is characterized by a Gaussian process. Hence, the central limit theorem is applied to the traffic process. In high-speed networks, homogenous (i.i.d) and heterogeneous sources (independent but not identically distributed) are multiplexed at the multiplexer. Here, we are concerned about the delay jitter characteristic in heterogeneous traffic since real network traffic consists of multimedia traffic, which is a mixture of CBR voice, CBR video, real time VBR video, real time VBR voice, time critical data and non-time critical data. In networks, traffic management controls are on the level of traffic aggregates, and not for individual links. Therefore, even though each link is not normally distributed, the aggregated traffic distribution shows Gaussian characteristics. This can be verified by applying the Central Limit Theorem (CLT) that states that summing a large numbers of independent variables with finite variance will converge or weakly converge to a Gaussian random density [22].

The tail of the queue length distribution ($P(Q > x)$) at high-speed multiplexer can be approximated under a variety of Cramer type assumptions (exponentially bounded marginal and autocorrelations of the arrival process). Thus, $P(Q > x)$ of an infinite buffer is asymptotically exponential [26],

$$P(Q > x) \sim Ce^{-\eta x}, \tag{1.6}$$

where the positive constant η is called the asymptotic decay rate, and the positive constant C is called the asymptotic constant, in which (1.6) is suggested for large values of x [25-30].

In section 1.4, the discussion of some popular approaches to find the queue length distribution by applying Gaussian/Normal process is provided.

1.3 Queueing and Teletraffic Theory Applied in High-Speed Wide Area Networks (WANs) with Heterogeneous Type Traffic

Due to the fast growth of telecommunication technology over the last past ten years (based on demands for better mobile telephony and Internet services) there have been many studies on traffic behavior and network architechture. Hence, traffic modeling is one of the key issues in traffic engineering for efficient designing and controlling of networks, where a mathematical model analysis of a complex system yields useful information. Here, Queueing theory and Teletraffic theory are applied as a tool for traffic modeling.

Queueing theory concerns the usage of a network resource for which the demands are random. When the resource is not available to the users, the requested service demand is queued or declined. While Teletraffic theory is a branch of engineering mathematics that relies heavily (but not solely on) queueing theory, it can be viewed commonly as queueing theory applied for telecommunication systems. It is applied for evaluating network designs, performance, stability and routing protocols, and optimizing network resources through mathematical analysis and simulation.

Hence, the following topics are essential in traffic modeling for networks:

- Characterization of Internet traffic
- Long-range dependent traffic models (LRD)
- Heavy-tailed distribution traffic models
- Gaussian traffic models
- Effective bandwidth model
- Traffic models for mobile networks
- Traffic measurements
- Traffic estimation and prediction

Poisson modeling fails to properly characterize multiplexed network traffic, due to the fact that Internet and Wide Area Network (WAN) traffic is self-similar. The Poisson traffic model severely underestimates burstiness especially when traffic flows multiplex/aggregate. The self-similar characteristic can be analyzed using statistical analytic tools, and the most popular methods are R/S plot, Variance-Time plot, Wavelet method, Periodogram method, and Whittle estimator. The several traffic models that are being proposed are Pareto On/Off model, M/G/∞ input model (infinite source Poisson model with heavy-tailed service/processing time), Fractional Brownian Motion (fBm), Fractional Auto-Regressive Integrated Moving Average (ARIMA) process, which capture the heavy-tailed characteristic (burstiness) of self-similar traffic. Here, the Hurst parameter (*H*) is the measure of burstiness.

Discrete-time definition of self-similarity is a stationary times processes *X*, and we define the *m*-aggregated time series $X^{(m)} = \left\{ X^{(m)}(k), k = 0,1,2,\cdots \right\}$ and $X^{(m)}(k)$ by averaging the original times series *X* over a non-overlapping block of size *m*, which is given in (1). Thus, a process *X* is said to be self-similar with parameter β (0 <β<1) if for all *m* = 1, 2, … we have the following definitions:

Exactly second-order self-similar: Variance: $\text{var}(X^{(m)}) = \dfrac{\text{var}(X)}{m^{\beta}}$ (1.7)

$$\text{Autocorrelation:}\, R_{X^{(m)}}(k) = R_X(k)\,.$$

Asymptotically second-order self-similar: Variance: $\text{var}(X^{(m)}) = \dfrac{\text{var}(X)}{m^{\beta}}$ (1.8)

$$\text{Autocorrelation:}\, R_{X^{(m)}}(k) \to R_X(k) \;\; \text{as}\, m \to \infty\,.$$

Eq. (1.7) and (1.8) define that self-similarity characteristic exists as the autocorrelation of the aggregated process has the same form as the original process, where the degree of variability or burstiness at different time scales will be the same. Fig. 1.1 and 1.2 show a graphical representation of self-similarity, where we can see that the networks traffic is very bursty and self-similar across all time scales. That is Fig. 1.1 and Fig. 1.2 look similar to one another in a distribution sense, all of the plots involve a fair amount of burtiness. The traffic pattern in Fig. 1.2, which sampled at large scale (the data are aggregated over 10 folds time scale, i.e., from 1ms to 0.01s) does not smooth out.

Thus, the other models such as Markov Modulated Poisson Process (MMPP) and M/M/1/B that can be applied in wireless network traffic modeling may not be a good fit. In wired or wireless networks, the Poisson arrival may not be an appropriate model since the networks traffic is very bursty in nature due to its long-range dependency (LRD) and self-similarity characteristics.

Numerous studies have found that aggregated traffic at a node (router, switch for wired networks case) or base station (for wireless networks case) exhibit LRD, and such traffic are very bursty. Thus, many broadband traffic studies have evolved around Gaussian, and they show that Gaussian models indeed provide a good approximation to network traffic if the aggregated traffic is sufficiently large [13][14]. Hence, heterogeneous wireless network traffic processes can be modeled as a Gaussian process, where some of these models will be discussed in section 1.4. In extension of the MVA approach, we can derive the MVA Lower Bound and Upper Bound to provide a boundary condition on the traffic. Another popular approach in traffic modeling is by using Fractional Gaussian Noise (fGN), which is a stationary Gaussian process with LRD. Recent research [15-17] models the heterogeneous networks traffic in the wavelet domain, where self-similar traffic exhibits long-range-dependent (LRD) correlation and a non-Gaussian marginal distribution [15].

In recent years, the emerging Gaussian model is believed to be the possible model for high-speed networks, where traffic flows in the networks are highly multiplexed. The Gaussian model can be LRD by including H in the modeling process. The main reasons to consider Gaussian traffic modeling are Central Limit Theorem (CLT) and the Functional Central Limit Theorem (FCLT) [18]. As the network grows, the number of multiplexed/aggregated traffic flows at a node (e.g., router, switch) increase, and the shape of traffic distribution will become closer and closer to Gaussian [19][20]. The aggregated input traffic queueing processes will weakly converge to a Gaussian process [19][20], which is shown in Fig. 1.3 and 1.4. These two plots display a normal probability plot of the data, where if the data does come from a normal distribution, the plot will appear linear. As the traffic flows aggregate, the traffic becomes closer and closer to a Gaussian process, which illustrated in Fig. 1.4.

The aggregated traffic is just as bursty as before, where it has the exactly same Hurst parameter *H*, but it is smoother [21]. Let a traffic process denotes as *X*, the aggregate of L independent copies of X is $X^{\oplus L}$ [21]. To demonstrate this, Fig. 1.4 shows the trace of aggregated process (byte.txt, a data set provided at Murad S. Taqqu's website is used; the data is splitting up into non-overlapping blocks, and then looking at the aggregation of two or more of those blocks).

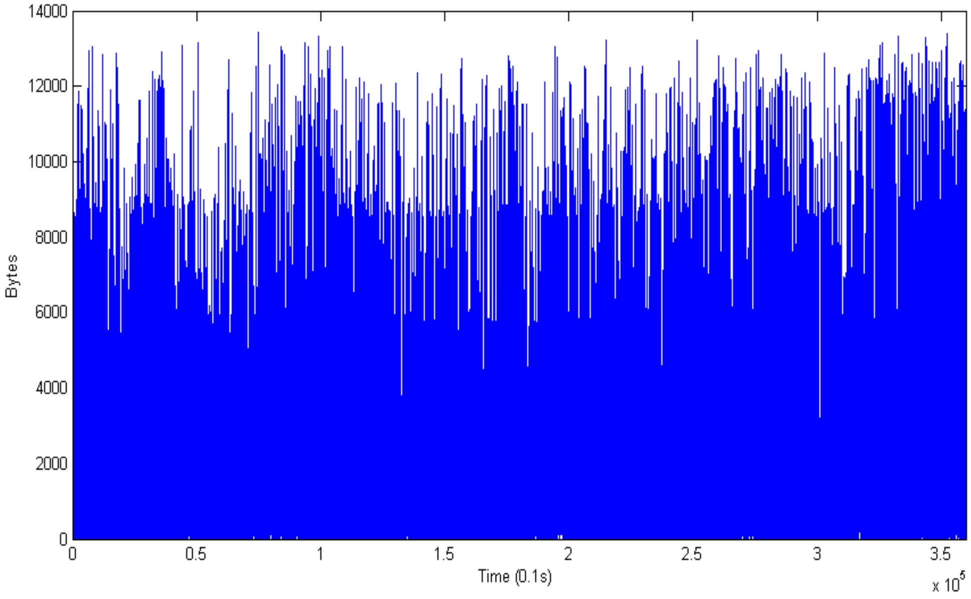


Fig. 1.1 The actual Ethernet traffic trace in original time scale (time unit = 1ms).

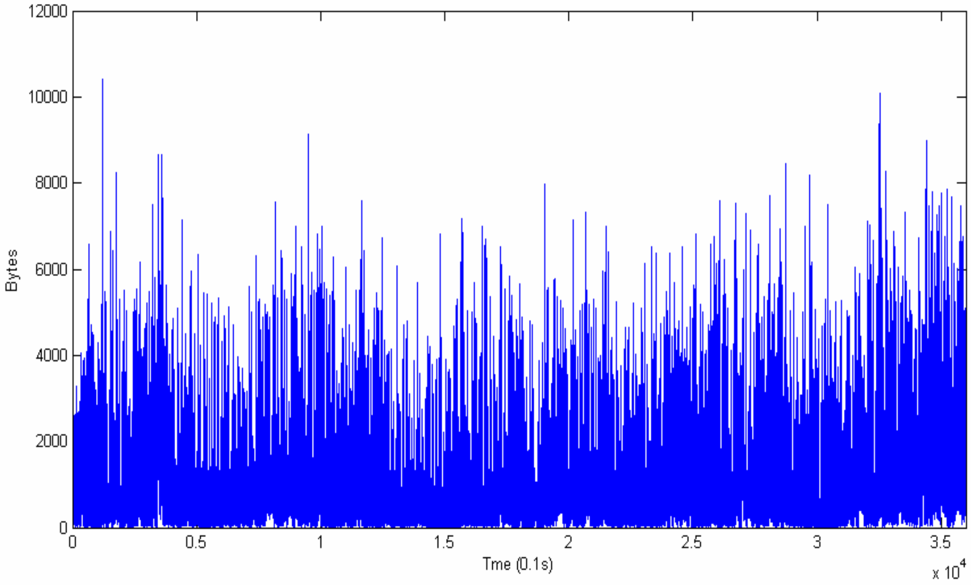


Fig. 1.2 The Ethernet traffic trace is aggregated 10 folds (time unit = 0.01s).

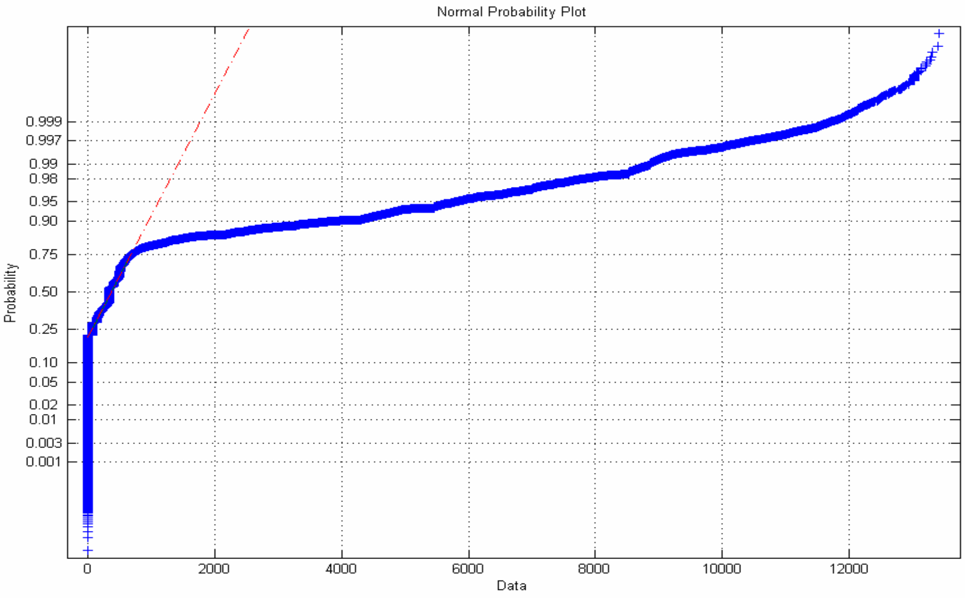


Fig. 1.3 Original Ethernet traffic.

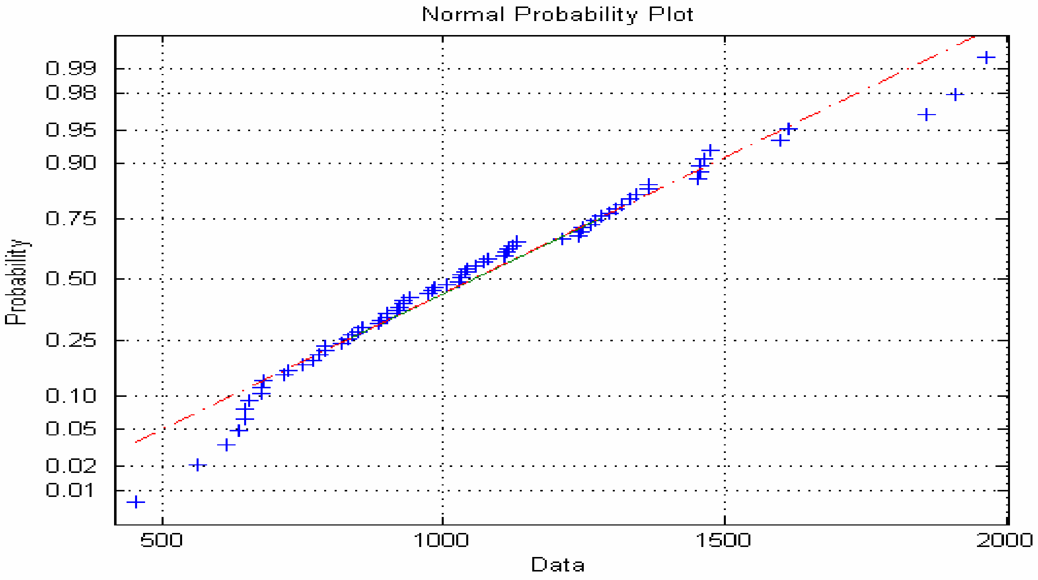


Fig. 1.4 Aggregated 100 Ethernet traffic flows.

1.4 Approximations for the Queue Length Distribution Gaussian Model for Broadband Traffic

In high-speed networks, the aggregated traffic becomes close to Gaussian characteristic by applying Central Limit Theorem (CLT) that stated that summing a large number of independent variables with finite variance is going to converge or weakly converge to a Gaussian random variable [22].

The Gaussian model is useful for two main reasons. First, any stationary Gaussian process can be completely characterized by its mean and auto-covariance. Second, the high-speed networks of today are highly complex and traffic is usually the superposition of some thousand network applications. By applying CLT, the aggregated traffic can be modeled as a Gaussian process; even though each single independent data source does not follow a Gaussian distribution. The only defect in this model is that there is a positive probability of a negative quantity of arriving traffic, which is impossible in real traffic. This significant weakness is counterbalanced by the fact that the CLT appeals as more and more traffic streams are aggregated to share a link, traffic becomes more Gaussian, and the case that the amount of negative traffic reduces as traffic is aggregated [24].

Many broadband traffic studies have evolved around Gaussian, and they show that Gaussian models indeed provide a good approximation to network traffic if the aggregated traffic is sufficiently large. In [22], [23], and [24], results show that the Gaussian traffic models could be the precise tool for analyzing the high-speed networks. Moreover, the Gaussian model is also a good fit for high-speed networks with differentiated services (DiffServ). In differentiated services networks, traffic management controls are on the level of traffic aggregates, and not for individual links [23].

The input traffic at a high-speed multiplexer can be characterized by Gaussian processes. Thus, the tail of the queue length distribution ($P(Q > x)$) at high-speed multiplexer can be approximated under a variety of Cramer type assumptions (exponentially bounded marginal and autocorrelations of the arrival process). Thus, $P(Q > x)$ of an infinite buffer is asymptotically exponential [26],

$$P(Q > x) \sim Ce^{-\eta x}, \tag{1.9}$$

where the positive constant η is called the asymptotic decay rate, and the positive constant C is called the asymptotic constant. Equation (1.9) is suggested for large values of x [25][27][28][29][30].

1.4.1 Single-Asymptotic Upper Bound [25]

The single-asymptotic upper bound can be represented as

$$P(Q > x) \leq \exp\left[-\left(\frac{2k}{S}\right)\left(\frac{x + kD}{S}\right)\right] \tag{1.10}$$

where k, S, and D are defined by the first two moments of a single Gaussian input process and the service rate μ per input, $D := 2\sum_{l=1}^{\infty} lC_{\gamma}(l)$ [25]. The $C_{\gamma}(l)$ denotes the auto-covariance function of a stationary Gaussian net input process $\gamma_n = \lambda_n - \mu$, where $\gamma_n = \lambda_n - \mu$ is the net amount of fluid input at time n and $(x)^{+} := \max\{0, x\}$ [25]. Since the mean and auto-covariance function of X_n can be computed in term of

$k := -E\{\gamma_0\}$ and $\mathrm{C}_\gamma(l)$ as $E\{X_n) = -kn,$ and $C_X(n_1, n_2) = \sum_{l_1=1}^{n_1} \sum_{l_2=1}^{n_2} C_\gamma(l_2 - l_1)$ by a change of variable $l = l_2 - l_1$, the variance can be expressed as a weighted sum of $\mathrm{C}_\gamma(l)$, i.e., $\mathrm{Var}\{X_n\} = nC_\gamma(0) + 2\sum_{l=1}^{n\text{-}1}(n\text{-}l)C_\gamma(l)$ [25].

1.4.2 Maximum Variance Asymptotic (MVA)[25][26]

For a discrete-time fluid queue, the amount of fluid in the buffer is denoted by Q_n, which can be defined using Lindley's equation [25]:

$$Q_n = (Q_{n-1} + \gamma_n)^+ \tag{1.11}$$

where $\gamma_n = \lambda_n - \mu$ is the net amount of fluid input at time n less the capacity μ, and $(x)^+ := \max\{0, x\}$ [25]. Let X_n represents a stochastic process $\{X_n : n = 0,1,2,\cdots\}$ defined by $X_n := \sum_{m=1}^{n} \gamma_{-m}$, the backward accumulation process (Note that the first two moment of X_n is defined as above in Single-Asymptotic Upper Bound section). Assume stationary and ergodicity of γ_n, and the stability condition, i.e., $E\{\gamma_n\} < 0$, the distribution of Q_n converges to a steady-state queue distribution. Hence, the aggregate arrival process, λ_n, can be characterized by a stationary Gaussian process. Thus, the steady-state queue length distribution can be represented as

$$P(Q > x) = \sup_{n \geq 0} P(X_n > x). \tag{1.12}$$

The function $P(X_n > x)$ achieves its maximum at a finite value of $n = \widehat{n}_x$:

$$P(Q > x) = P(X_{\widehat{n}_x} > x) \tag{1.13}$$

where in a Gaussian process, the time scale $\hat{n}_x$ is also the time at which $\sigma_{x,n}^2$ achieves its maximum value $\langle \sigma_x^2 \rangle$.

1.4.3 MVA Lower Bound [25]

By studying the asymptotic behavior of $P(Q > x)$, and with $\alpha = 1$ corresponding to the lower bound, (for a Gaussian process) the lower bound of $P(Q > x)$ can be written in terms of $\psi(z) := 1 \big/ \sqrt{2\pi} \int_z^\infty \exp[-(y^2/2)]dy$ (the tail function of the standard Gaussian distribution) as

$$P(Q > x) \geq \psi\left(\sqrt{\frac{x}{\langle \sigma_x^2 \rangle}}\right)$$

$$\sim \sqrt{\frac{\langle \sigma_x^2 \rangle}{2\pi x}} \exp\left(-\frac{x}{2\langle \sigma_x^2 \rangle}\right). \tag{1.14}$$

1.4.4 MVA Upper Bound [25]

The approach is to find a function of $z = \sqrt{x \big/ \langle \sigma_x^2 \rangle}$, which resembles $\psi(z)$ such that is similar to the asymptotic upper bound (1.11),

$$\psi(z) = \exp[-(z^2/2)]. \tag{1.15}$$

1.4.5 MVA general approach (not bound) [26]

In discrete time queue, the extreme value distribution of X_n is determined by determining the extreme value distribution of a new stochastic process Z_n, $\{Z_n : n = 0,1,2,\cdots\}$, which is defined as

$$Z_n = \frac{X_n - \mu(1-\rho)n}{\mu((1-\rho)n+k)} \tag{1.16}$$

where Z_n is a normal process with

$E\{Z_n\} = 0$, and

$$\mathrm{Var}\{Z_n\} = \frac{\mathrm{Var}\{X_n\}}{\mu^2((1-\rho)n+k)^2} . \tag{1.17}$$

Thus, the steady-state queue distribution can be rewritten as [26]

$$\begin{aligned} P(Q > x) &= \sup_{n\geq 0} P\left(X_n > x\right) \\ &= \sup_{n\geq 0} P\left(Z_n > 1\right). \end{aligned} \tag{1.18}$$

The $\mathrm{Var}\{Z_n\}$ achieves its maximum value for $n = \hat{n}_x$ [26]

$$\sup_{n\geq 0} P(Z_n > u) \approx P(Z_{\hat{n}} > u) \tag{1.19}$$

for sufficiently large *u*. If $\mathrm{Var}\{Z_n\} << 1$, then by using the approximation in (1.19) to obtain [26]

$$P(Q > x) \approx P\left(Z_{\hat{n}} > 1\right). \tag{1.20}$$

Thus, the tail probability $P(Q > x)$ can be defined as

$$P(Q > x) \approx P\left(Z_{\hat{n}} > 1\right) = \int_{\frac{1}{\sigma_{\max}}}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) dx \tag{1.21}$$

where $\sigma_{\max}$ is the maximum value of $\mathrm{Var}\{Z_n\}$, which can be computed by determining the maximum value of $\mathrm{Var}\{Z_n\}$,

$$\mathrm{Var}\{Z_n\} = \frac{nC_\gamma(0) + 2\sum_{l=1}^{n\text{-}1}(n\text{-}l)C_\gamma(l)}{\mu^2((1-\rho)n+k)^2}. \tag{1.22}$$

1.4.6 Long Range Dependent Gaussian model [27] [32]: The Discrete Gaussian Model

Assume a FIFO single server queue, and the time is divided into fixed-length sampling intervals (discrete time). The arrival process is considered to be stationary and ergodic, and Gaussian distributed. A discrete time queueing system has k statistically independent, but identical, traffic streams aggregated at a multiplexer. Let time be divided into fixed-length sampling intervals [28]. The amount of queued traffic in a buffer at time n is defined as V_n, with standard deviation σ, and net input rate m. H is the Hurst parameter of the input traffic [27]. Under the long-range dependent (LRD) case, the overflow probability (unfinished work distribution) for a Gaussian fractal queue can be approximated by [13][19][22][27][32],

$$P\{V_\infty > t\} \approx \frac{\sqrt{2\pi}}{\sigma}\psi(-m,\sigma)\times\exp\left(-\frac{2}{\sigma^2}\times|1-H|^{-2}\left(\frac{H}{|(1-H)m|}\right)^{-2H} t^{2-2H}\right) \tag{1.23}$$

where H is the Hurst parameter of the traffic, V_n denotes the amount of queued traffic in a buffer at time n, σ is the standard deviation of the arriving traffic, m is the net input rate to the buffer (which has to be negative for stability) [13][19][22][27][32]. The function $\psi(-m,\sigma)$ represents the mean of the random variable that is normally distributed with mean $-m$ and standard deviation σ [13][19][22][27][32],

$$\psi(m,\sigma) = \frac{\sigma}{\sqrt{2\pi}} e^{-\frac{m^2}{2\sigma^2}} - \frac{m}{2} erfc\left(\frac{m}{\sigma\sqrt{2}}\right). \tag{1.24}$$

1.4.7 The Continuous Gaussian Model

A superposition of independent Gaussian processes is still a Gaussian processes. Assume FIFO with single server queue.

Let us define the cumulative arrival processes of class *i*, which are independent continuous Gaussian processes with stationary increments, as $A_t^{\{i\}} = m_i t + Z_t^{\{i\}}$ [14][27]. The mean input rate is m_i and $Z^{\{i\}}$ is a centered continuous Gaussian process with variance $v_i(t) = \mathrm{Var} Z_t^{\{i\}}$ [14][27]. The server capacity is denoted as *c*, and the queue with input $A^{\{i\}}$ is denoted as $Q^{\{i\}}$ [14][27].

Applying the path space of the Gaussian processes, reproducing kernel Hilbert space, and based on the estimate of the buffer emptiness probability, the lower bound for the aggregated queue distribution can be derived using the 1-dimensional normal distribution as [27]

$$P\left(Q_0^{\{1,\ldots,k\}} \geq x\right) \geq P\left(A_{-t_x}^{\{1,\ldots,k\}} \geq x - (c - \sum_{i=1}^{k} m_i) t_x \right)$$

$$= \phi\left(\frac{(c - \sum_{i=1}^{k} m_i) t_x}{\sqrt{\sum_{i=1}^{k} v_i(t_x)}} \right) \tag{1.25}$$

where $t_x < 0$ is the value of *t* which minimize the expression [14][27]

$$\frac{(x+(c-\sum_{i=1}^{k} m_i)(-t))^2}{\sum_{i=1}^{k} v_i(t)}. \tag{1.26}$$

1.4.8 Large two-stage models: Heavy Traffic

Let the arrival rate be $\lambda > 0$, service rate $\mu > 0$, and offered load $m = \lambda / \mu$. Assume the arrival process is asymptotically normal. Let N_ℓ be a stationary number of customers in the system that has ℓ channels. Then *N* is approximately normal with mean *m* and variance *mz* when *m* is large. The distribution of N_ℓ can be obtained as [chapter 7 of 33]

$$P(N_\ell(i)) \approx (1/\sqrt{mz})\varphi\left(\frac{i-m}{\sqrt{mz}}\right) \bigg/ \phi\left(\frac{i-m}{\sqrt{mz}}\right), \tag{1.27}$$

where φ is the standard normal density function, and ϕ is the standard normal distribution [chapter 1 of 33]

$$\varphi(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-(x-\mu)^2 / 2\sigma^2\right). \tag{1.28}$$

In this chapter, the motivations of studying the jitter process and its concept have been discussed. The self-similarity characteristic of broadband networks traffic have also been studied. Then, several popular traffic modeling techniques for modeling the networks with heterogeneous type traffic have been investigated. In the next chapter, the methods used for determining the self-similarity of a network are being examined.

Chapter 2

Wide Area Networks and Internet Traffic Analysis

2.1 Introduction: Analyzing traffic traces

The statistical analysis is applied to a set of collected data, and then a statistical inference of the data set is made. Here, we are interested in traffic self-similarity, which is measured by the Hurst parameter. The methods used to estimate the traffic self-similarity by computing the Hurst parameter (*H*) are:

1. Aggregated Variance
2. R/S Analysis
3. Wavelet Estimate

2.2 Aggregated Variance Method

2.2.1 Description and Methodology

Aggregated variance method is also referred to as variance-time analysis.

- For each *m* = 1, 2, 3, …, data is divided into non-overlapping blocks of size *m* to obtain the aggregated process $X^{(m)}$. For instance, if the given input data set has length *N*:

$$X^{(m)}(k) = \frac{1}{m} \sum_{i=(k-1)m+1}^{km} X(i), \quad k = 1,2,\ldots,N/m\,. \tag{2.1}$$

- For the variance vs. time log-log plot, the variance is first normalized to the corresponding sample variance (Thus, the estimate $\hat{\beta}$ of the asymptotic slope will fall between –1 and 0, which suggests self-similarity). Then, we need to plot $\log_{10}(\text{var}(X^{(m)}))$ versus $\log_{10}(m)$. The variance is computed as

$$\widehat{\text{var}}(X^{(m)}) = \frac{1}{N/m}\sum_{k=1}^{N/m}\left(X^{(m)}(k) - \overline{X}\right)^2. \tag{2.2}$$

<u>*Recall*</u>:

The discrete-time definition of self-similarity is a stationary times *X*, and we define the *m*-aggregated time series $X^{(m)} = \{X^{(m)}(k), k = 0,1,2,\cdots\}$ and $X^{(m)}(k)$ by averaging the original times series *X* over non-overlapping block of size *m*, which is given in (2.1). Thus, a process *X* is said to be self-similar with parameter β $(0 < \beta < 1)$ if for all *m* = 1, 2, … we have the following:

Exactly second-order self-similar: Variance: $\text{var}(X^{(m)}) = \dfrac{\text{var}(X)}{m^{\beta}}$ (2.3.a)

Autocorrelation: $R_{X^{(m)}}(k) = R_X(k)$.

Asymptotically second-order self-similar: Variance: $\text{var}(X^{(m)}) = \dfrac{\text{var}(X)}{m^{\beta}}$ (2.3.b)

Autocorrelation: $R_{X^{(m)}}(k) \to R_X(k)$ as $m \to \infty$.

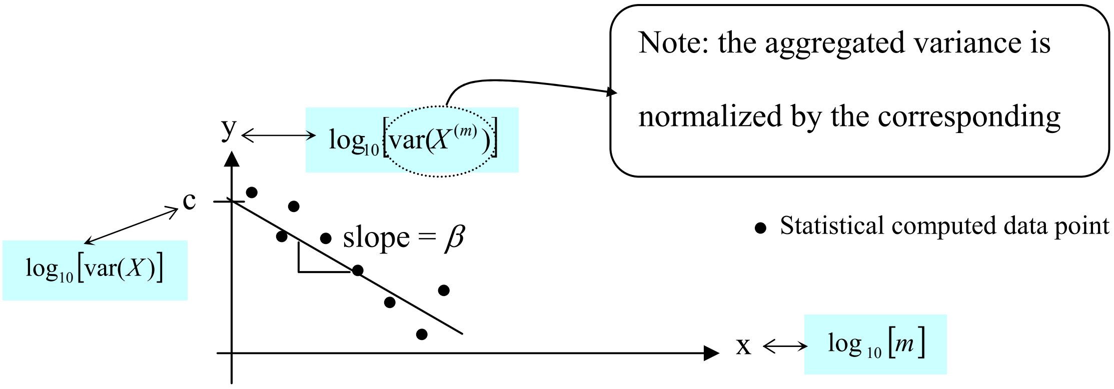
(2.3.a) and (2.3.b) define that self-similarity characteristic exists as the autocorrelation of the aggregated process has the same form as the original process, where the degree of variability or burstiness at different time scales will be the same.

Simple Least Squares Line Fitting:

Now, by taking $\log_{10}$ on both sides of the variance in (2.3): $\log_{10}[\mathrm{var}(X^{(m)})] = \log_{10}[\mathrm{var}(X)] - \beta \log_{10}[m]$, and it can be rewritten as

$$\log_{10}[\mathrm{var}(X^{(m)})] = -\beta \log_{10}[m] + \log_{10}[\mathrm{var}(X)] \qquad (2.4)$$

and $\mathrm{var}(X) = \sigma^2$ is a positive finite constant. Equation (2.4) can be fitted through a simple least squares line equation as $y = mx + c$. So, we try to fit a line through the points on the log-log plane. The graphical representation is given in Fig 1.



Note: the aggregated variance is normalized by the corresponding

y
$\log_{10}[\mathrm{var}(X^{(m)})]$
c
$\log_{10}[\mathrm{var}(X)]$
slope = β
● Statistical computed data point
x
$\log_{10}[m]$

Fig. 2.1 The relationship between aggregated variance equation and the least squares line.

From the log-log plot, we estimate $\hat{\beta}$ by computing the slope of the least line. Then, we can compute H using (6). The slope is formulated:

$$\hat{\beta} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - \left(\sum x_i\right)^2}, \qquad \text{where } n \text{ is the number of samples.} \quad (2.5)$$

(Note: To avoid computational round off errors (2.5) is used instead of a simple form of (2.4). This has been verified by comparing the results from both equations, with the later underestimating the slope value)

- Hurst parameter is estimated by fitting a simple least squares line through the resulting points in the plane. By ignoring the very low-end and very high-end *m* value, compute the slope ($\hat{\beta}$), and then the Hurst parameter (*H*) is computed as:

$$\hat{H} = (\hat{\beta} + 2)/2 = 1 - \hat{\beta}/2 . \tag{2.6}$$

(Note: capped symbol means an estimate of the true value.)

Simulation Results

Ethernet data set: byte.txt (Note: the byte.txt data set (1999) is more recently collected compared to BC-pAug89, but the this data set is smaller compared to BC-pAug89) is used to plot Fig. 2. According to [34], Ethernet data in 1989 (AUG89 trace) has *H* ~0.8, and the traffic is more self-similar from 1990 onwards with H~0.9 (FEB92 trace during high internet traffic hour of Feb. 1992 with *H*~0.96 [34]). Fig. 2 shows the same degree of self-similarity in the Ethernet traffic of year 1999, which once again confirms that *H*~(0.90,0.96) for Ethernet traffic in 90's. This may be due to the popularity of Internet usage, and more applications/services (e.g. digital video streaming, multimedia service) being able to be provided through Internet.

The aggregated variance plot of Ethernet traffic is showed in Fig. 2.2. It has the H value 0.9279, which indicates that the traffic do have the self-similarity characteristic.

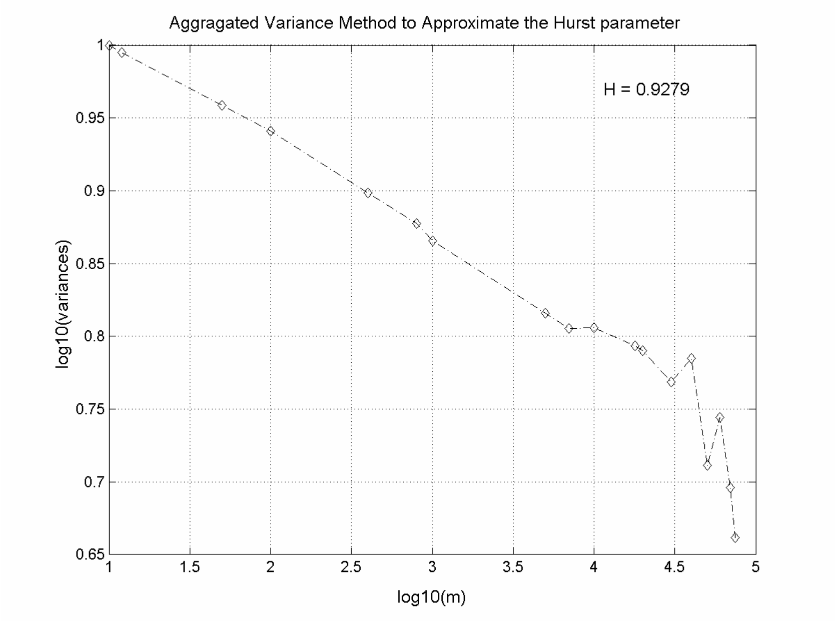


Fig 2.2 This figure shows that Ethernet traffic is very self-similar, which is indicated by a fairly large *H* value.

Star Wars IV, high quality: Terse_StarWarsIV.dat data set is used to plot Fig. 2.3.

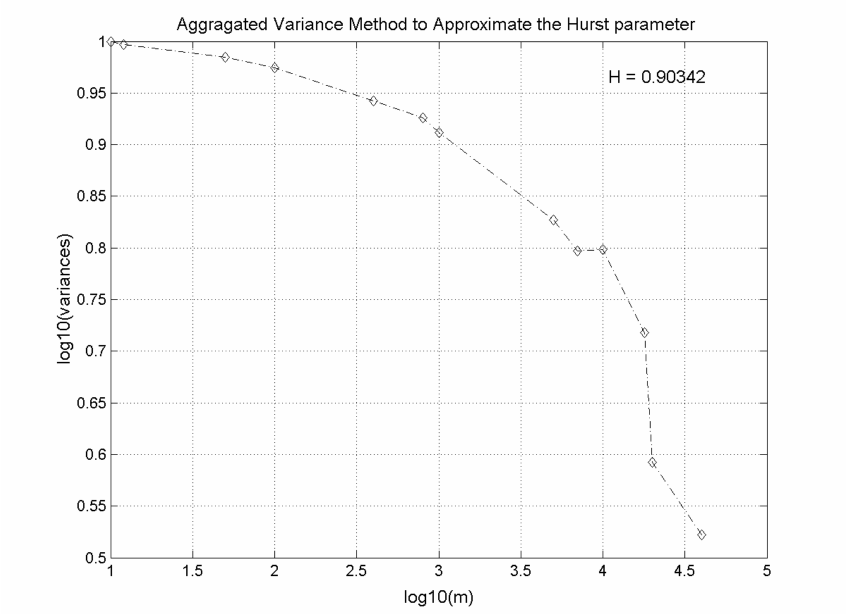
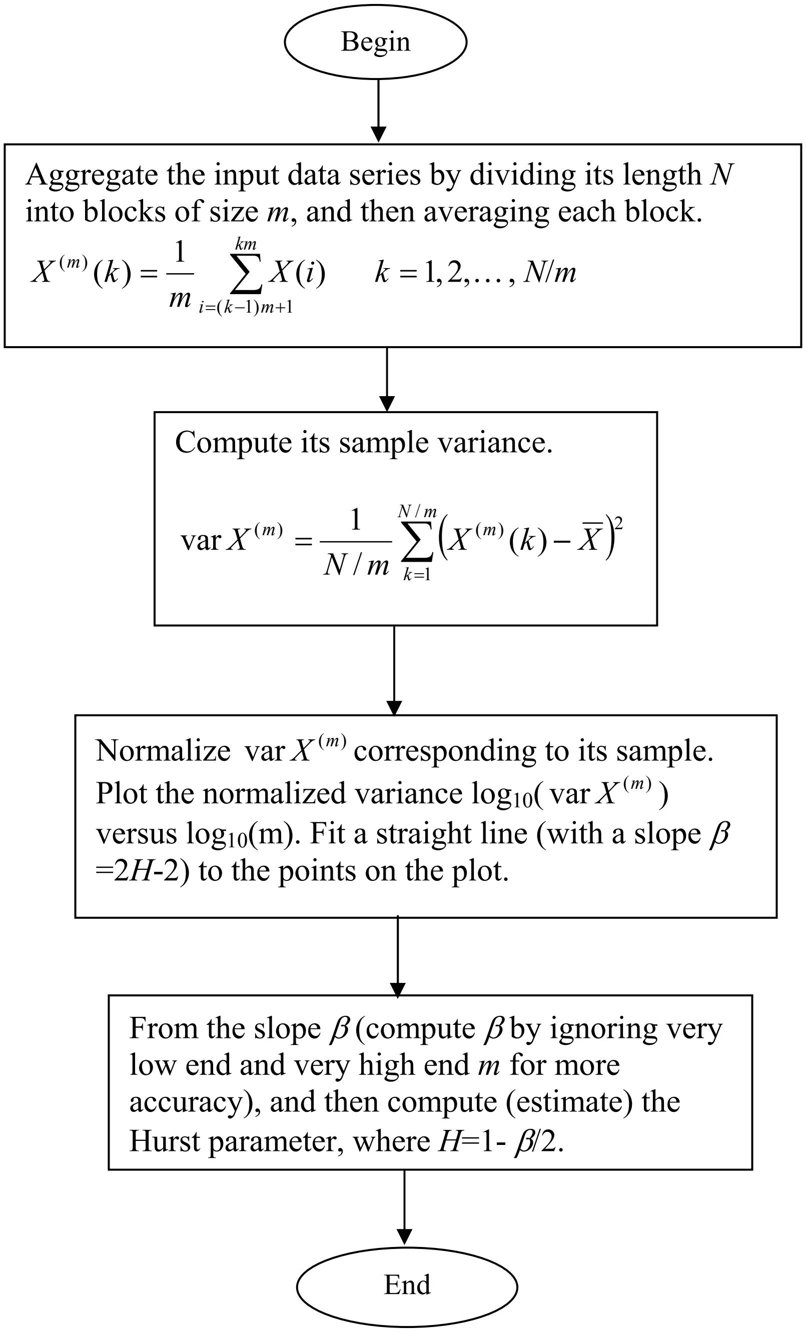


Fig 2.3 This figure shows that Star Wars IV data stream is self-similar, and the level of self-similarity is indicated by the H value.

2.2.3 Simulation Flow Chart



Begin

Aggregate the input data series by dividing its length *N* into blocks of size *m*, and then averaging each block.

$$X^{(m)}(k) = \frac{1}{m} \sum_{i=(k-1)m+1}^{km} X(i) \qquad k = 1, 2, \dots, N/m$$

Compute its sample variance.

$$\operatorname{var} X^{(m)} = \frac{1}{N/m} \sum_{k=1}^{N/m} \left(X^{(m)}(k) - \overline{X} \right)^2$$

Normalize $\operatorname{var} X^{(m)}$ corresponding to its sample. Plot the normalized variance $\log_{10}(\operatorname{var} X^{(m)})$ versus $\log_{10}$(m). Fit a straight line (with a slope β =2*H*-2) to the points on the plot.

From the slope β (compute β by ignoring very low end and very high end *m* for more accuracy), and then compute (estimate) the Hurst parameter, where *H*=1- β/2.

End

2.2 R/S Method

2.3.1 Description and Methodology

R/S statistic is also referred to the rescaled adjusted range, and is formulated as

$$\frac{R}{S}(d):=\frac{1}{S(d)}\left[\max_{0\le t\le d}\left(Y(t)-\frac{t}{d}Y(d)\right)-\min_{0\le t\le d}\left(Y(t)-\frac{t}{d}Y(d)\right)\right]. \qquad (2.7)$$

For a stationary time series, $X=\{X_i\}$, where $i\geq 1$, with sample sum $Y(d)=\sum_{i=1}^{d}X_i^2$,

and sample variance

$$S^2(d)=\frac{\sum_{i=1}^{d}X_i^2-Y^2(d)/d}{d}. \qquad (2.8)$$

For fractional Gaussian noise or fractional ARIMA:

$E[R/S(d)]\sim C_H d^H$ as $d\rightarrow\infty$ and C_H is a finite positive constant that is not dependent on *d*. Thus, we have

$$\log_{10}[E[R/S(d)]]=H\log_{10}[d\ ]+\log_{10}[C_H], \qquad (2.9)$$

From (2.9), *H* is directly proportional to the slope. Refer to the least squares line concept in section 2.2 for the same argument.

The algorithm: (An algorithm of the code also given in [35])

- For a data length *N*, divide it into series of *K* blocks.
- Then, for each lags *d* compute $R(t_i,d)/S(t_i,d)$, for different starting point t_i; the starting points must be $(t_i-1)+d\leq N$.
- Plot $\log_{10}[R(t_i,d)/S(t_i,d)]$ versus $\log_{10}(d)$.
- Fit a least squares line through the points on the log-log plane, the slope of the line is the estimator for Hurst parameter, *H* = slope of the line.

(Note: the slope is calculated by using the same equation as (2.5).)

The same Ethernet data set and the video file of Star Wars IV (high quality) are used to show that the multimedia source and Internet traffic do exist self-similarity, which illustrates in Fig. 2.4 and Fig. 2.5 using the R/S method. Since both figures have *H* value larger than 0.5, and closed to 1.

2.3.2 Simulation Results

The same Ethernet data set and the video file of Star Wars IV (high quality) are used to show that the multimedia source and Internet traffic do exist self-similarity, which illustrates in Fig. 2.4 and Fig. 2.5 using the R/S method. Since both figures have *H* value larger than 0.5, and closed to 1.

Compare to [35] $H = 0.903$ for K=10 and up to $\log_{10}(d) = 4.6$, the obtained results are closed to [35] but not exactly the same, this may due to only up to $\log_{10}(d) = 3.4$ estimate points are used.

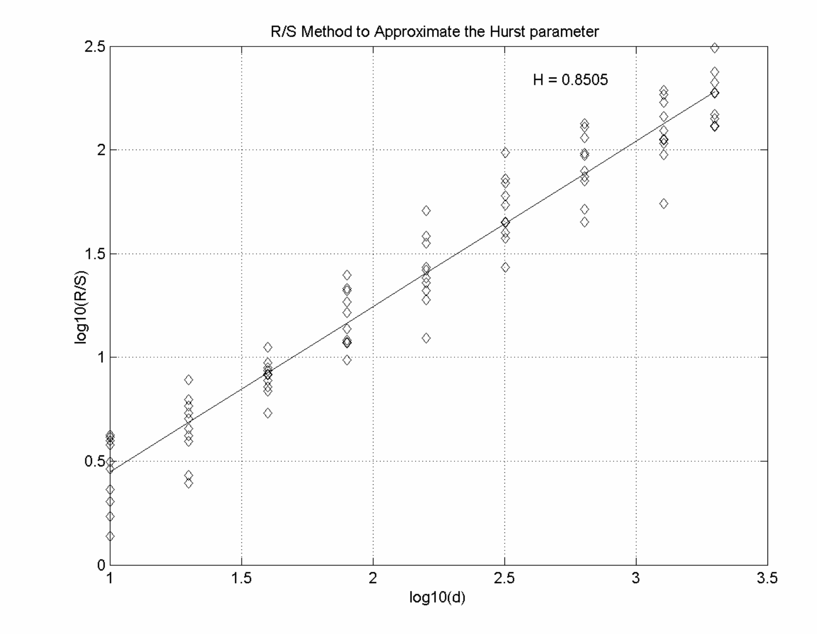


Fig 2.4 This figure shows that the data set is self-similar, and the level of self-similarity is indicated by the H value. Here, K = 10 and upto $\log_{10}$(d)=3.4 is used.

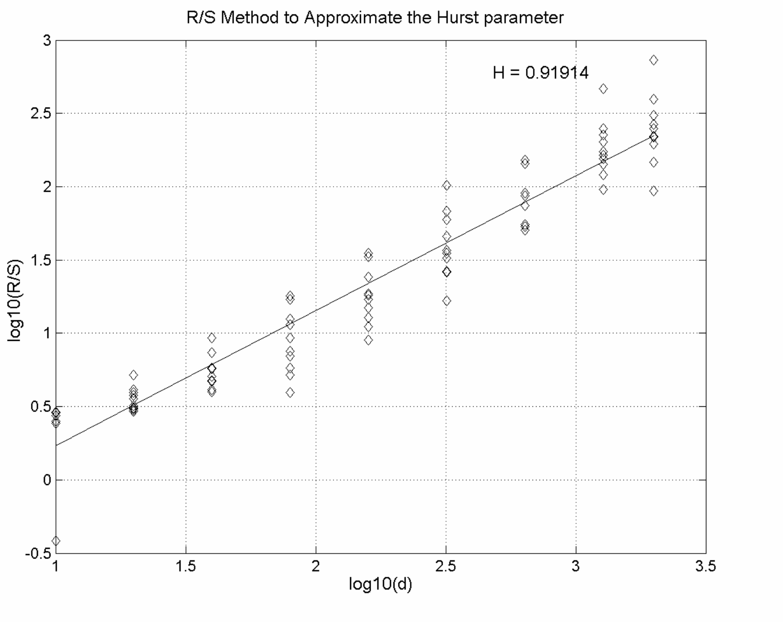
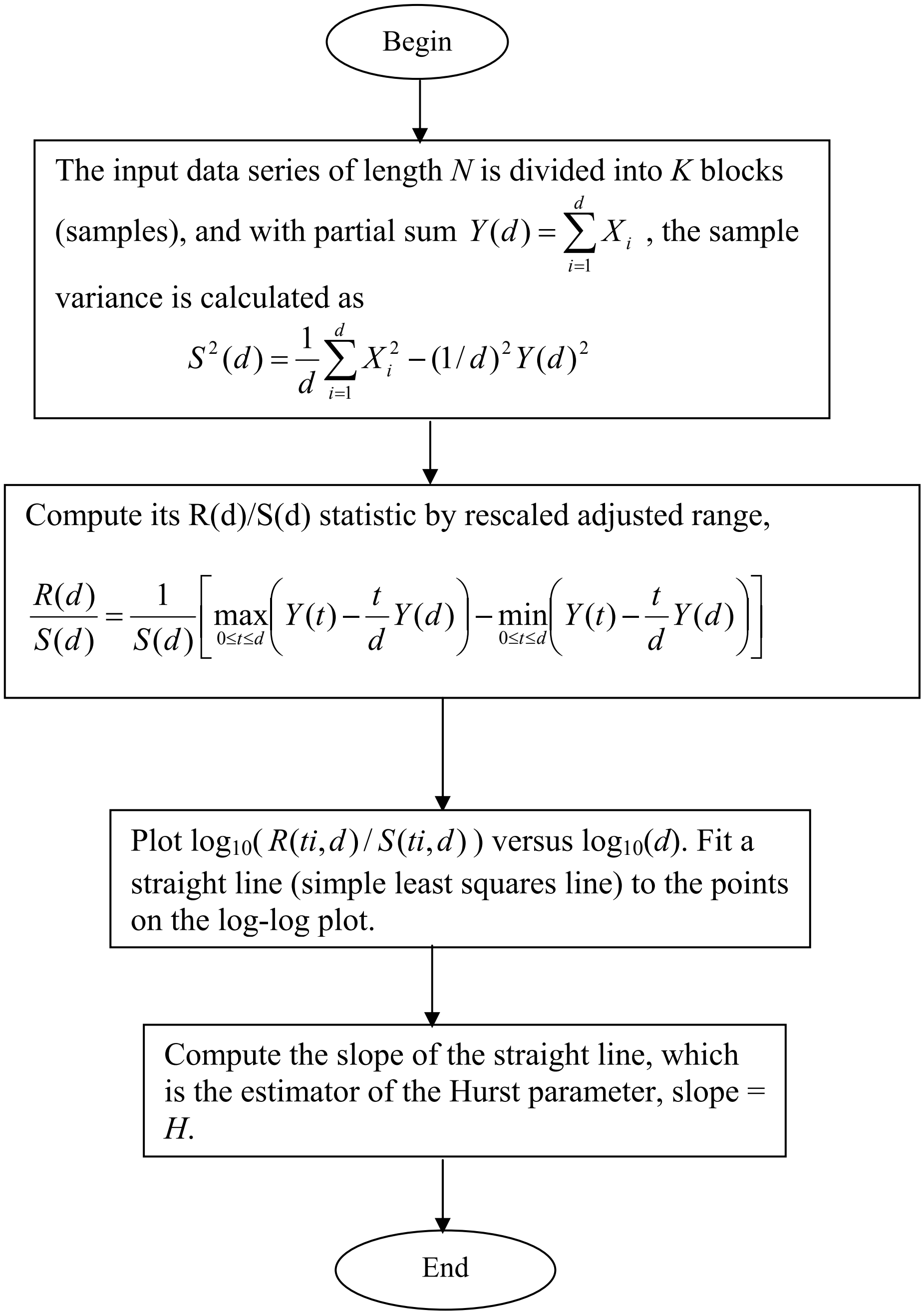


Fig 2.5 This figure shows that Star Wars IV data stream is self-similar, and the level of self-similarity is indicated by the H value. K=10 and up to $\log_{10}$(d)=3.4 is used.

2.3.3 Simulation Flow Chart



The input data series of length *N* is divided into *K* blocks (samples), and with partial sum $Y(d)=\sum_{i=1}^{d} X_i$, the sample variance is calculated as

$$S^2(d)=\frac{1}{d}\sum_{i=1}^{d} X_i^2-(1/d)^2 Y(d)^2$$

Compute its R(d)/S(d) statistic by rescaled adjusted range,

$$\frac{R(d)}{S(d)}=\frac{1}{S(d)}\left[\max_{0\le t\le d}\left(Y(t)-\frac{t}{d}Y(d)\right)-\min_{0\le t\le d}\left(Y(t)-\frac{t}{d}Y(d)\right)\right]$$

Plot $\log_{10}$($R(ti,d)/S(ti,d)$) versus $\log_{10}(d)$. Fit a straight line (simple least squares line) to the points on the log-log plot.

Compute the slope of the straight line, which is the estimator of the Hurst parameter, slope = *H*.

2.4 Estimating the Hurst Exponent using Wavelet Spectral Density

2.4.1 Description and Methodology

This method transforms the time series into discrete wavelets. In Matlab, Discrete Wavelet Transform can be use to decompose the input data to scaling and wavelet functions.

1. The Hurst exponent for a set of data is calculated from the wavelet spectral density,

$$P_j = \frac{1}{2^j}\sum_{i=1}^{2^j} c_i^2 \qquad \text{where } c_i \text{ is the wavelet coefficients.} \tag{2.10}$$

2. Regression line through the normalized spectral density :

$$H = \text{abs}((\text{slope} - 1\ )/2). \tag{2.11}$$

Recall:

The spectrum of a long-range dependent process $X(t)$ with the discrete wavelet transform coefficients c_X show the following behavior $E[c_X^2(j,l)] = 2^{j(1-2H)}C$ [43]. By taking log2 on both sides,

$$\log_2[E[c_X^2(j,l)]] = (1-2H)\log_2[2^j] + \log_{10}[C\ ]. \tag{2.12}$$

From (2.12):

$$slope = (1-2H) \rightarrow H = \frac{slope-1}{2} \rightarrow H = \left|\frac{slope-1}{2}\right| \tag{2.13}$$

which follows the same least squares line argument. Refer to the least square line concept in section 2.2.

Algorithm:

- Compute the power spectrum by averaging the squares of the coefficients of the transform $P(j)$.
- Plot $\log_2(P)$ versus $\log_2(2^j)$.
- Fit a simple least squares line through the points on the log-log plane.
- The Hurst parameter can be computed as $H = \left|\frac{slope - 1}{2}\right|$

.(Note: the slope is calculated by using the same equation as in (2.5).)

The same Ethernet data set and the video file of Star Wars IV (high quality) are used to show that the multimedia source and Internet traffic do exist self-similarity, which illustrates in Fig. 2.6 and Fig. 2.7 using the wavelet method. Using this method, both of the data sets have calculated H value larger than 0.5, and closed to 1.

2.4.2 Simulation Results

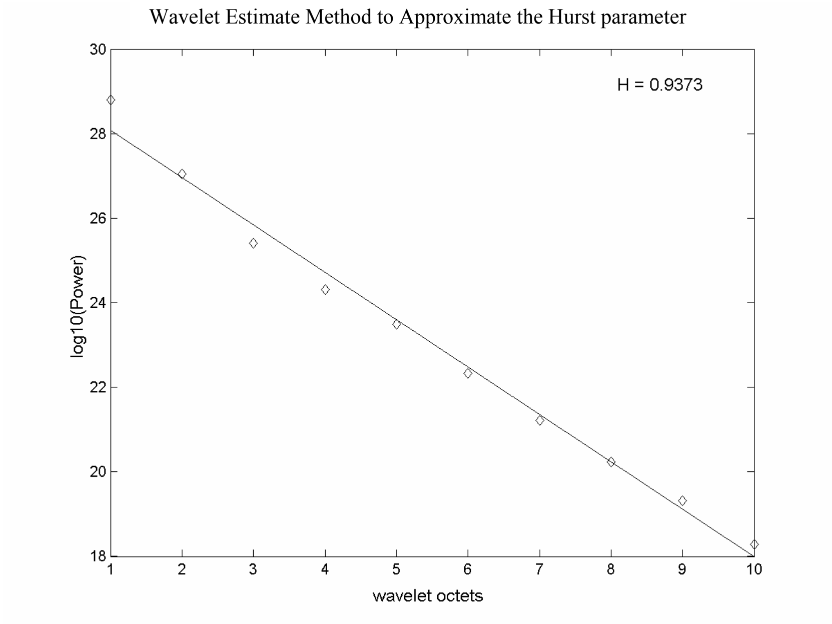


Fig 2.6 The figure shows that Ethernet traffic is self-similar, and the level of self-similarity is indicated by the H value.

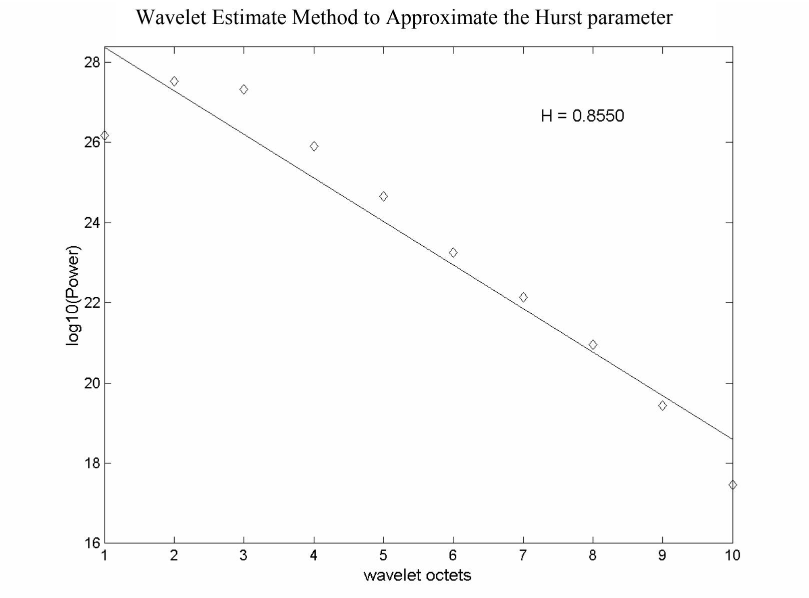
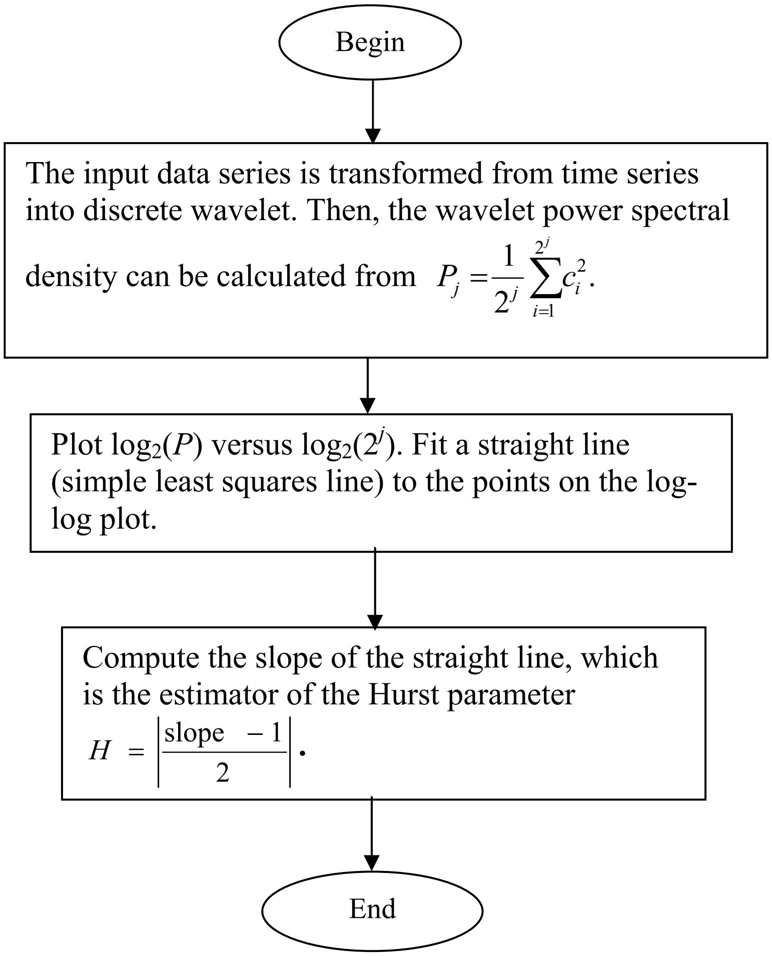


Fig 2.7 The figure shows that Star Wars IV data stream is self-similar, and the level of self-similarity is indicated by the H value.

2.4.3 Simulation Flow Chart



The input data series is transformed from time series into discrete wavelet. Then, the wavelet power spectral density can be calculated from $P_j = \frac{1}{2^j}\sum_{i=1}^{2^j} c_i^2$.

Plot $\log_2(P)$ versus $\log_2(2^j)$. Fit a straight line (simple least squares line) to the points on the log-log plot.

Compute the slope of the straight line, which is the estimator of the Hurst parameter $H = \left|\frac{\text{slope} - 1}{2}\right|$.

2.5 Observation and Discussion

All three graphical estimation methods provide a good estimate of the Hurst parameter (Fig. 2.8), *H*, and allow us to detect self-similarity in an empirical data set. These three methods provide a closed estimation of *H*, which is between (0.85, 0.93) for Ethernet traffic, and a value between (0.85, 0.91) for Star Wars data set. However, these three methods provide point estimate to *H*, and we may be interested in a more refined data analysis, e.g., confidence intervals for *H*. Therefore, an interval estimation can be done with maximum likelihood-type estimates (MLE) and related methods based on periodogram [34]. The slope of the graph is the graphical tool to compute the Hurst parameter, where the slope of the straight line will provide the information about the Hurst parameter. The critical step in this aspect is choosing the statistical computed points on the plane that should be used to compute the slope. Usually, the very low end and very high end points are not used [36].

Fundamentally, the goal of traffic modeling in this dissertation is:

- To find a general/universal traffic model that is suitable for broadband networks.
- The model could be appropriate to real traffic and to all type of traffic.
- Performance of real networks transporting the traffic of the model could be estimated with adequate accuracy.

By understanding the characteristics of Internet and multi-media traffic, we would be able to estimate the jitter process of the broadband networks with adequate accuracy, where the networks with heterogeneous type traffic are greatly different than the networks with homogeneous type traffic. The following two chapters have derived the mathematical equations for evaluating the jitter process in those networks.

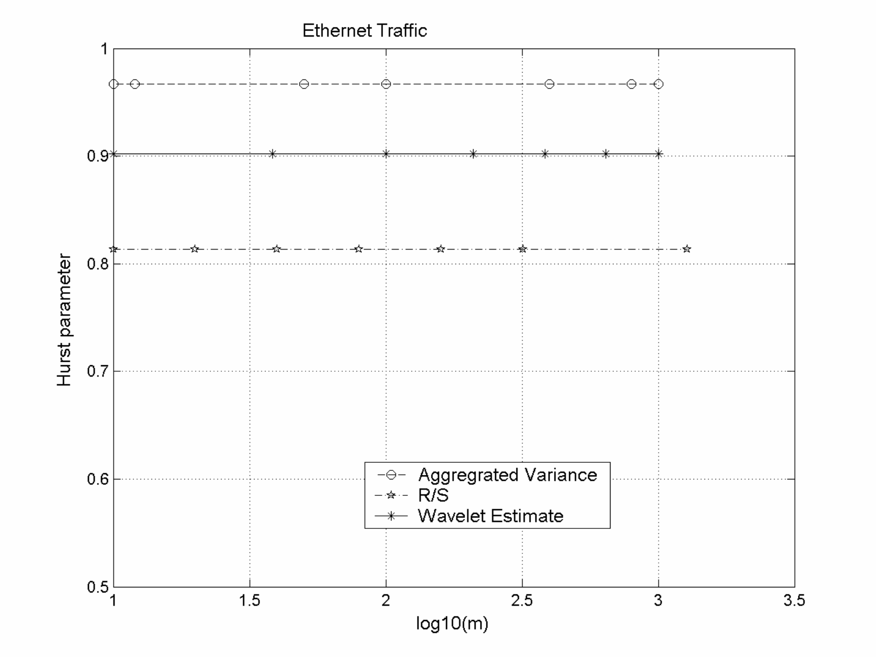


Fig 2.8 The comparison between the three methods is shown, where the level of self-similarity is indicated by the H value. The estimation range is between (0.82, 0.93) for Ethernet traffic for low m aggregated level, where all the three methods have proved that the Ethernet traffic is very self-similar.

Chapter 3

Theoretical Analysis of Delay Jitter of Homogeneous Traffic in ARQ Wireless Networks

3.1 Abstract

This chapter provides a theoretical analysis of delay jitter for homogeneous time constrained traffic and proposes a call admission control (CAC) scheme for wireless differentiated services (DiffServ) networks that apply Selective-Reject (SREJ) automatic retransmission request (ARQ). The CAC scheme regulates the lower class traffic to provide the required levels of delay jitter for the higher priority classes.

3.2 Introduction

DELAY JITTER [1] (which is equivalents to interarrival packet jitter) is one of the most important parameters in quality of service (QoS) support measurement for real-time data transfer. Due to this fact, there are several papers that have analyzed the performance of delay jitter [6][7][61]. In [6], the time division multiple access (TDMA)/time division duplex (TDD) scheduling scheme is used. In [7], the jitter is analyzed for mobile ad hoc networks (MANET) based on the ad hoc on-demand distance vector (AODV) routing protocol, where the authors propose a novel handover mechanism. Both [6] and [7] provide results on jitter performance analysis based on computer simulation for the wireless network models. In [61], the authors provide a jitter analysis on wireless networks involving ARQ error recovery, where the delay jitter is calculated using the window-length generating function and the numerical results are verified through simulation. The delay jitter research of [9] extends the results of [1] and [10] for DiffServ networks, but all three papers do not consider the channel error probability or the ARQ error control scheme, which makes the results difficult to apply to wireless networking applications. Therefore, this paper provides a novel analysis of the per-class jitter performance of DiffServ networks based on wireless channels that experience packet errors, assuming a non-preemptive head-of-the-line (HOL) priority scheme. The derivations provide a direct method to analyze/evaluate the per-class jitter based on the DiffServ network, retransmission time constraints, and network packet error parameters [12]. In addition, this paper evaluates the effects on the delay jitter in reference to the priority control scheme of the ARQ traffic for two cases of: 1) the ARQ traffic has a priority over the original transmission traffic; and 2) the ARQ traffic has no priority over the original transmission traffic.

3.3 Analytical Model

3.3.1 Network and System Model Assumptions

In this dissertation's chapter, we investigate a wireless communication link that is framed for a fixed packet size or is time slot based (e.g., time-division multiple access (TDMA)) transmission. The examples of mobile cellular communications based on time-coordinated slotted data transmissions include Global Systems for Mobile Communications (GSM), IS-54 and IS- 136 of the North American Digital Cellular

(NADC), and Integrated Digital Enhanced Network (iDEN), which are all TDMA based protocols, and the hybrid code-division multiple access (CDMA) based TDMA systems. Based on these models, in this paper, the traffic delay jitter characteristics are based on a discrete time (slotted) queueing structure. The network traffic model is assumed to be slot-based homogenous and statistically time division multiplexed. The ARQ model applied in this paper is a selective-reject/repeat topology, where only the packets that are erroneous are retransmitted.

Additionally, assume class 1 has priority over class 2, class 2 has priority over class 3, and so on. Among the same priority the first in first out (FIFO) transmission ordering applies to the first transmission packets. In this paper, two kinds of ARQ traffic priority control schemes are investigated: 1) when the ARQ traffic has a priority over the original transmission traffic; and 2) when the ARQ traffic has no priority over the original transmission traffic. The network controller model that gives a non-preemptive priority to the ARQ traffic is based on existing network schedulers that process admission control in reference to time delay bounds, such as the earliest deadline first (EDF) scheduler, where a delivery time deadline is applied to the data packets, thereby limiting the number of possible retransmissions of an erroneous packet [61]. For these types of schedulers, the ARQ packets will commonly have less of a time bound remaining for delivery, since a part of their delay bound would have been used in its former transmission(s), which would result in having a higher priority over the original transmission traffic [12], [61]. Based on these assumptions, in the following, the mathematical foundation of the analysis is established.

3.3.2 Theoretical Formulations

Assume the packets arrive and depart at the beginning of a slot, which leads to a discrete time queueing process. A slot is defined as the time interval [t-1, t), where t is a non-negative integer (i.e., $t \geq 0$). Here we define the mth time slot interval (cycle) of a tagged stream's mth packet as [mT, (m+1)T), where m and T are non-negative integers.

Any individual periodic $T^{'}$ traffic stream of interest among the different class of service (CoS) could be the tagged stream. Thus, as we investigate the jitter of a specific class, namely the nth class traffic stream, and define the traffic as a mixture 2 of the specified periodic $T^{'}$ tagged stream and the combined background traffic of numerous $T^{'}$ period streams from different sources. The number of packets that arrive in each slot will not only depend on the data sources but also on the number of ARQ packets that need to be retransmitted. Hence, the period of each renewal cycle is $T = T^{'} + T^{'}P_e$, where the retransmitted packets will also consume the channel capacity. The term P_e represents the packet retransmission probability, which takes into account that an ARQ retransmission packet can also be erroneous again. Assume the packet errors are independent from packet to packet. Given the packet error probability p_{error}, the probability that a packet transmission is successfully received is $(1 - p_{\text{error}})$. If up to r retransmissions are limited within the time bound, each of the j consecutive packet transmissions fail, and a successful transmission then follows, where j = 1, 2, · · ·, r, and r = 1, 2, · · · , T . Therefore, P_e becomes $P_e = \sum_{i=1}^{T+1} \frac{(i-1)}{i} p_{error}^{i-1}(1 - p_{error})$, where T is the transmission window size (processing rate) in packets (slots). The following notations are applied throughout this paper. The network's packet processing rate is μ packets/s and the average arrival rate is λ_{total}, and it is assumed that there are N classes of service. Hence, for a stable system the utilization is required to satisfy

$$\rho = \frac{\lambda_{\text{total}} + P_e \lambda_{\text{total}}}{\mu} = \frac{\left(\lambda_N + \lambda_{(N-1)} + \cdots + \lambda_1\right) + P_e \lambda_{\text{total}}}{T} \leq 1 . \tag{3.1}$$

The utilization of a specific class n (i.e., $n = 1, 2, \ldots, N$) is noted as ρ_n, and it can be obtained from $\rho_n = \left(\mu_n / \mu\right) = \left[\left(\lambda_n + P_e \lambda_n\right) / \mu\right]$, which considers the original transmission traffic and the retransmission traffic of the class n data packets.

The CAC will control the amount of new admissions given to the class n sources such that the aggregated arrival rate of class n traffic does not exceed $\lambda_{n,\max} = \left(\mu\rho_n / \left(1 + P_e\right)\right)$. When the amount of ARQ packets increase the CAC will regulate the traffic loads of the lower classes by giving out less admissions, to maintain the required QoS of the higher priority classes.

In order to investigate the effects of delay jitter in DiffServ networks, the derivation of the probability of jitter (i.e $P\{\widetilde{J} = j\}$) becomes an essential building block.

Let the random variable $\widetilde{J}_{m,n} = j$, where $n = 1, 2, \ldots, N$ class and cycle $m \geq 1$ [10], denotes the process of the normalized jitter or position difference between the mth and (m+1)th packet of the nth class tagged stream that originates from the same source. The term $A_{(m+1),n}$ indicates the total number of class n packets that arrive in the (m +1)th time slot group. The average number of higher priority packets and ARQ packets that arrive in the (m+1)th period is represented as $b_{(m+1),(\overline{n-1})}$, which can be obtained from $b_{(m+1),(\overline{n-1})} = b_{(m+1),(n-1)} + b_{(m+1),(n-2)} + \cdots + b_{(m+1),1}$. The term $b_{(m+1),(n-1)}$ denotes the number of all the equivalent ($n-1$)th priority packets that enter the buffer before the class ($n-1$)

tagged stream's (m+1)th packet. The probability that a jitter size of j will exist for the class n tagged stream is given by [9]

$$P\left\{\widetilde{J}_{m,n} = j\right\} = P\left\{\widetilde{J}_{m,n} = j \mid A_{(m+1),n} - 1 = k,\ b_{(m+1),\,\overline{(n-1)}} = a\right\}$$
$$\cdot P\left\{A_{(m+1),n} - 1 = k \mid b_{(m+1),\,\overline{(n-1)}} = a\right\} \cdot P\left\{b_{(m+1),\,\overline{(n-1)}} = a\right\}. \tag{3.2}$$

In (3.2), the term $P\left\{A_{(m+1),n} - 1 = k \mid b_{(m+1),\,\overline{(n-1)}} = a\right\}$ is based on the consideration that the assignment probability of the class n packets is dependent to the number of the higher priority packets (i.e., variable a) that arrive within the time slot group specified by $b_{(m+1),\,\overline{(n-1)}}$. We subtract 1 from the total number of class n packets that arrive in the (m+1)th time slot group (i.e., $A_{(m+1),n}$) due to the fact that the tagged frame is also among the overall traffic. The binomial distribution used in (3.2) is based on

$$P\left\{A_{(m+1),n} - 1 = k \mid b_{(m+1),\,\overline{(n-1)}} = a\right\} = B_k\left[p, (T-a-1)\right]$$
$$= \binom{T-a-1}{k} (p)^k (1-p)^{(T-a-1)-k} \tag{3.3}$$

where $k \in [1, (T-a-1)]$, $a \le |j|$, and each class n packets have a successful (errorless) arrival probability p. The $P\left\{b_{(m+1),\,\overline{(n-1)}} = a\right\}$ term represents the probability of the higher priority (class 1 to class (n-1)) packets occupying a number of slots among the first j slots of jitter offset before the class n tagged packet enters service, and is calculated by the probability mass function of a bilateral geometric function (BGF) [9]. The discrete triangle distribution is applied to the term $P\left\{\widetilde{J}_{m,n} = j \mid A_{(m+1),n} - 1 = k,\ b_{(m+1),\,\overline{(n-1)}} = a\right\}$, which is a symmetric distribution, which represents the probability that the time slot position of two sequentially arriving frames will be offset by a jitter amount of +j or −j slot(s) for the range [−k, k]

$$P\left\{\widetilde{J}_{m,n} = j \mid A_{(m+1),n} - 1 = k,\ b_{(m+1),\overline{(n-1)}} = a\right\} = \Delta_k(j-a)$$

$$= \begin{cases} \dfrac{1}{k+1} - \dfrac{|j-a|}{(k+1)^2}, & \text{for } |j| \le k \\ 0, & \text{otherwise} \end{cases}, \qquad (3.4)$$

among the overall x possible time slots, $0 \le x \le T$.

A special case of this is the zero jitter case (i.e., $j = 0$) [9], which is simply

$$P\left\{\widetilde{J}_{m,n} = 0\right\} = 1 - 2\sum_{j=1}^{(T-1)} P\left\{\widetilde{J}_{m,n} = j\right\}. \qquad (3.5)$$

3.3.3 Jitter Analysis with Non-Priority ARQ Traffic

In this section a comparison between the different packet error probabilities for wireless networks deploying a non-priority ARQ scheme with fixed (limited) channel capacity is presented. The non-priority ARQ scheme represents the case where the ARQ packets do not have a priority over the regular traffic. The probability of jitter for this case can be derived as,

$$P\left\{\widetilde{J}_{m,n} = j\right\} = \sum_{a=0}^{|j|} f_1(a) \sum_{k=(|j|-a)}^{T-a-1} B_k\left[\frac{\rho_n}{T' + T'P_e}, (T-a-1)\right] \cdot \Delta_k(j-a), \qquad (3.6)$$

where $1 \le |j| \le (T-1)$. The results are illustrated in Fig. 3.1. The $P\left\{b_{(m+1),\overline{(n-1)}} = a\right\} = f_1(a)$ term can be obtained from the probability mass function of a BGF [9],

$$f_1(a) = \frac{1}{2}(1 - p_{\overline{(n-1)}})(p_{\overline{(n-1)}})^a, \quad a \le | j |. \qquad (3.7)$$

The term $p_{\overline{(n-1)}}$ is the probability of arrival of the higher priority packets and their ARQ packets compared to class *n* in a frame of *T* slots (i.e., $p_{\overline{(n-1)}} = \sum_{i=1}^{n-1}\rho_i$). The jitter probability for class 1 packets can be easily obtained from (3.6)

$$P\{\tilde{J}_{m,1} = j\} = \sum_{k=|j|}^{T-1} B_k\left[\frac{\rho_1}{T}, (T-1)\right]\Delta_k(j),\ \ 0 \le |j| \le (T-1). \qquad (3.8)$$

3.3.4 Jitter Analysis with Priority ARQ Traffic

In this section a comparison between the different packet error probabilities for wireless networks deploying a priority ARQ scheme with fixed (limited) channel capacity is presented. The probability of jitter for this case can be derived as,

$$P\{\tilde{J}_{m,n} = j\} = \sum_{a=0}^{|j|} f_2(a) \sum_{k=(|j|-a)}^{T-a-1} B_k\left[\frac{(\lambda_n/\mu)}{T^{'} + T^{'}P_e}, (T-a-1)\right]\Delta_k(j-a) \qquad (3.9)$$

where $1 \le |j| \le (T-1)$. In each slot there is a probability P_{error} of receiving a negative acknowledgement of the message. If a negative acknowledgement is received, the transmitter immediately retransmits the packet that was received in error instead of a new packet. The $P\{b_{(m+1),\overline{(n-1)}} = a\} = f_2(a)$ term can be obtained from the probability mass function of a BGF [9],

$$f_2(a) = \frac{1}{2}(1 - p^{'}_{\overline{(n-1)}})(p^{'}_{\overline{(n-1)}})^a, \quad a \le |\, j\,|. \qquad (3.10)$$

The term $p^{'}_{\overline{(n-1)}}$ is the probability of arrival of the higher priority packets, which includes the retransmission packets of classes 1 through *n*, which can be obtained from $p^{'}_{\overline{(n-1)}} = \frac{P_e\lambda_n}{\mu} + \sum_{i=1}^{n-1}\rho_i$.

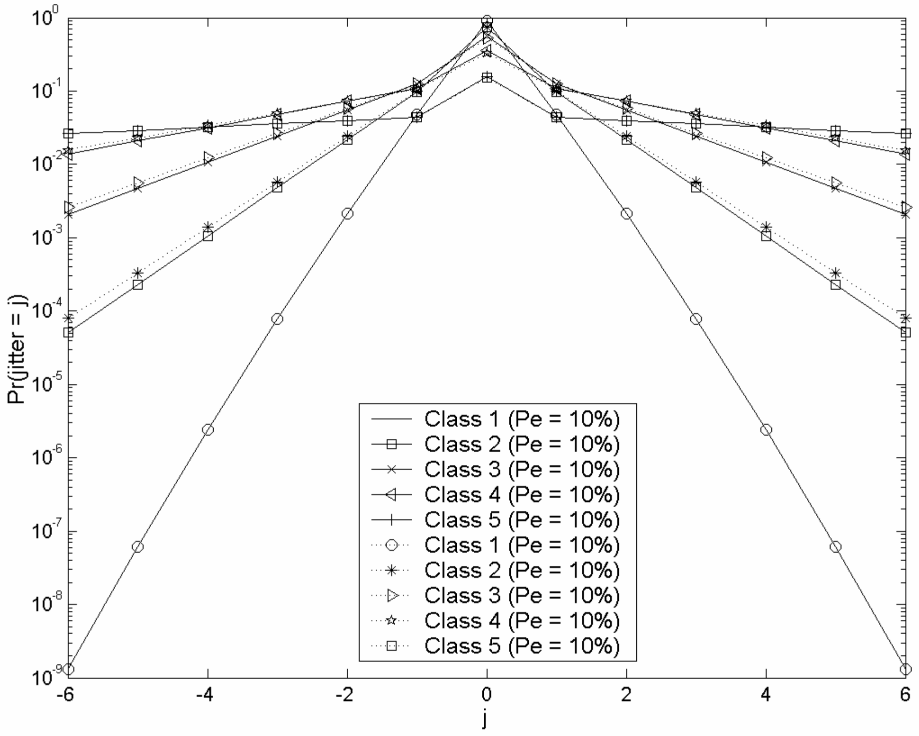


Fig. 3.1 Comparison between wireless networks deploying the priority and non-priority ARQ scheme with Pe = 10%. The solid line represents the non-priority ARQ scheme, and the dotted line represents the priority ARQ scheme.

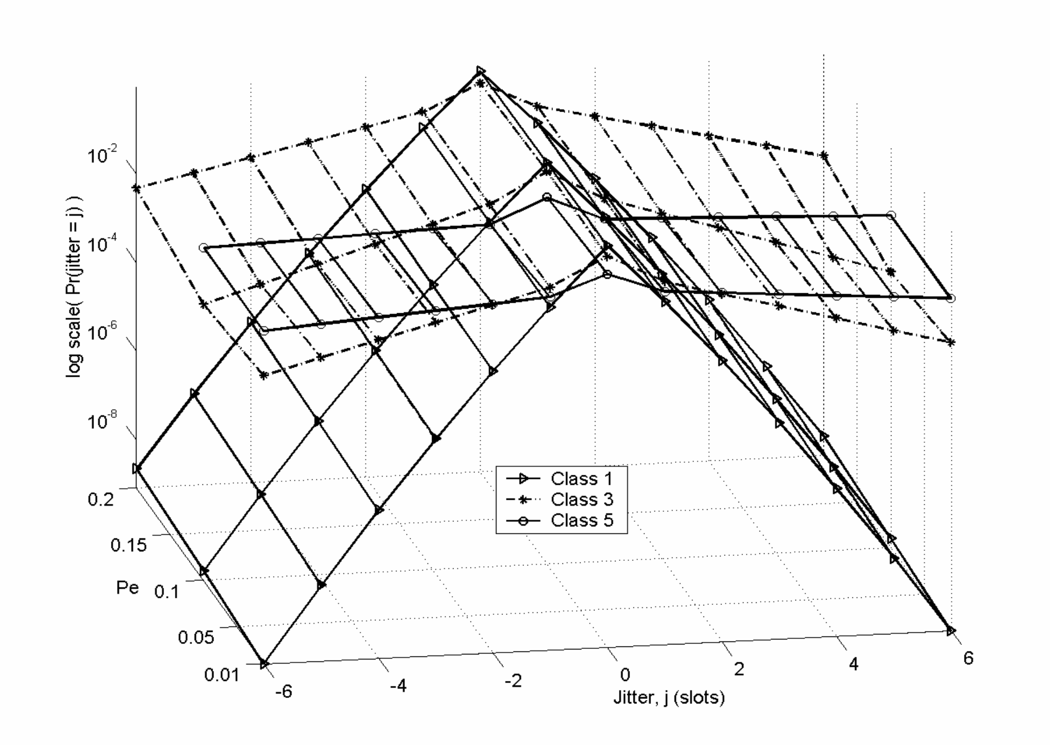


Fig. 3.2 Comparison between wireless networks non-priority ARQ scheme with Pe = 20%.

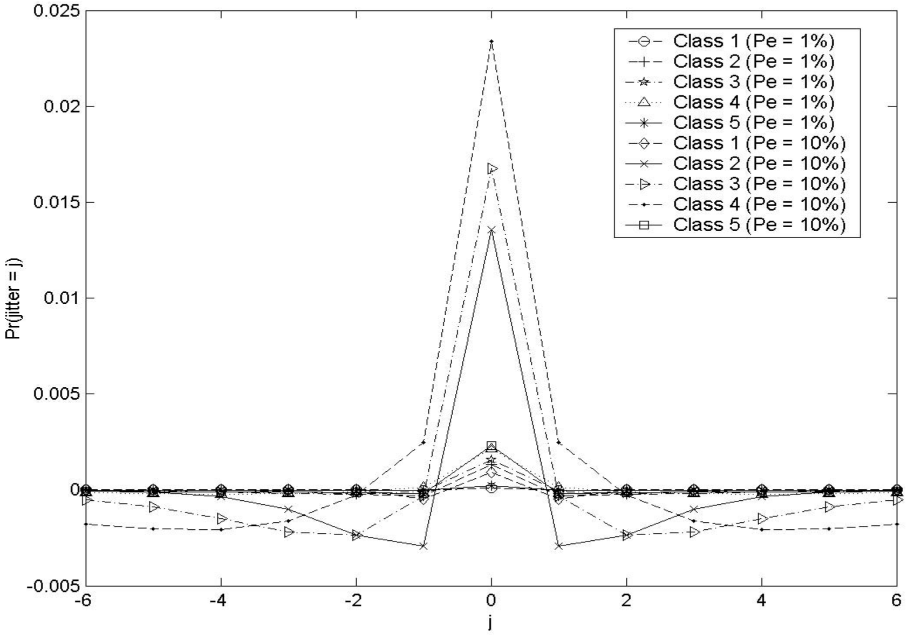


Fig. 3.3 Comparison between wireless networks deploying the priority and non-priority ARQ scheme with Pe = 1% and 10% for class 1 through class 5.

3.4 Discussions and Conclusions

This paper provides a novel analysis of the per-class jitter probability of homogeneous traffic streams performance of DiffServ networks based on wireless channels that experience packet errors, assuming a non-preemptive head-of-the-line (HOL) priority scheme. The investigation has been conducted for the case where the ARQ retransmission packets have priority over the first time transmission packets. The mathematical results were compared with simulation data, and the results show an accurate match, which verifies the derivations. The results of Fig. 3.1 show that for both the priority and non-priority ARQ cases, the probability of jitter for the same class can be controlled to be approximately the same disregarding the channel packet error rate for the traffic of the higher priority classes if the CAC of the wireless network can obtain accurate packet error rate information and respond effectively in regulating the admission access. The call admission controller will dynamically manage the network traffic loads by regulating admissions based on the amount of retransmission traffic that is anticipated. Thus, the traffic load can be regulated dynamically to maintain the required jitter performance.

Based on the observations of Fig. 3.1, Fig. 3.2, and Fig. 3.3, it can be observed that the normalized jitter probability of the priority ARQ scheme is larger than the non-priority ARQ scheme. The difference in performance becomes more significant as the probability of packet error increases (e.g., when P_e increases from 1% to 10%, and 20%). However, class 1 and class 2 still maintain a relatively low probability of jitter compared to the traffic of the lower classes. In the ARQ scheme with priority control, the ARQ packets will block the new packets from being transmitted during the retransmission process. Thus, the normalized jitter probability of the wireless network deploying ARQ with the priority scheme is worse compared to the ARQ with the non-priority scheme, especially for the low priority packets. Although the ARQ priority scheme sacrifices the delay jitter performance for the low priority classes, it may be needed especially for real-time service applications in which the packet might be useless unless a successful reception of the packet is obtained within its time bound [12], [61]. In Fig.3.2, as P_e is 20%, no left-over channel capacity is available for class 5 (the lowest priority class) since a significant part of the channel capacity is used for retransmission of the higher priority class packets. In Fig.3.2, the negative values represent the normalized jitter probability of the priority ARQ scheme is larger than the non-priority ARQ scheme, where probability of jitter for ARQ priority scheme minus non-priority ARQ scheme. These results demonstrate the influence and the effectiveness of DiffServ in limiting the delay jitter for the high priority users.

Chapter 4

Jitter Characteristics of Heterogeneous Wired Communication Networks, Gaussian Traffic Modeling and Queue length approximation using the MVA Approach

In real-time wired or wireless communications, each transmitted packet may experience different time delays during arrival at the destination. This variation in delay is often referred as delay jitter or simply jitter. Jitter is an unsurprising result of packet transferring in high-speed switched networks. The variable delay results from the processing and queueing at each node through the multi-hop network. This is because the packets transmitted between a given source and destination may vary in length, may take different routes, and may experience different delay at the switches. To compensate the packet delay jitter, the incoming packet are buffered at the destination, and then replayed at a constant rate based on the controller that regenerates the audio/real-time traffic (playback mechanism). The regenerated packets are therefore smoothed out.

However, the compensation provided by the delay buffer is limited, where any incoming real-time traffic packets that have been delayed more than the replay delay bound limit, then the packets will be discarded. Thus, by controlling the delay jitter within the network can further reduce the timing distortion of real-time applications. In order to examine the heterogeneous traffic network jitter behavior in a more general environment, we develop a generalized analytic approach to approximate the jitter probability density function (pdf) of a stationary tagged stream that is multiplexed on a high-speed IP network. The main objective of this chapter is to examine the relationship between the jitter and traffic characteristics, such as traffic load/utilization, and period *T* of the tagged stream. This insight may aid the design of the scheduling control of the buffer in the playback mechanism at each node. For a generalized approach to analyze a queueing process, the heterogeneous traffic arrivals at a node are modeled as a Gaussian process.

In high-speed networks, it has been shown that aggregated traffic has a direct relation to Gaussian characteristics through Central Limit Theorem (CLT), which states that summing a large number of independent variables with finite variance can converge or weakly converge to a Gaussian random variable [22]. The networks' self-similar traffic exhibits long-range-dependent (LRD) correlations, and it is very close to being Gaussian and strongly LRD (i.e., approximately Fractional Gaussian noise), which is caused by both small and large file transfers over limited bandwidth links. Some of the traffic may follow a non-Gaussian marginal distribution. However, by applying CLT, the aggregated traffic can be modeled as a Gaussian process; even though each single independent data source does not follow a Gaussian distribution.

The Gaussian model is useful for two main reasons. First, any stationary Gaussian process can be completely characterized by its mean and autocovariance. Second, as today high-speed networks are highly complex and traffic is usually heavily multiplexed of thousands of network applications. By applying CLT, the aggregated traffic can be modeled as Gaussian process; even each single independent data source does not follow a Gaussian distribution. The only defect in this model is that there is a positive probability of a negative quantity of arriving traffic, which is impossible to happen in real networks traffic. This significant weakness is counterbalanced by the fact that the CLT appeals as more and more traffic streams are aggregated to share a link; traffic becomes more Gaussian, and the case that the amount of negative traffic reduces as traffic is aggregated [24].

Many broadband traffic studies have evolved around the Gaussian model, and they show that Gaussian models indeed provide a good approximation to networks traffic if the aggregated traffic is sufficiently large. In [22], [23], and [24], the results shown that the Gaussian traffic model could be the precise tool for analyzing high-speed networks. Moreover, the Gaussian model is also a good fit for high-speed networks with differentiated services (DiffServ). In differentiated services networks, traffic management controls are on the level of traffic aggregates, not for individual links [23][54][55].

4.1 Introduction

In order to evaluate delay jitter, the queue length distribution of the output queue was first analyzed. We can relate the transmission delay and the queue length (number of packets) ahead of the tagged packet (i.e., Q_n represents the number of customers ahead of the *n*th tagged packet). Let any customer's service time be equal to one slot and follows FIFO order, Q_n denotes the number of customers, regardless of their class affiliation, that are waiting for service just ahead of the arrival of the *n*th packet of the *p*-

customer. Therefore, the normalized jitter relative to tagged stream period T is (see Chapter 1 for more detail discussion)

$$\tilde{J}_n = Q_{n+1} - Q_n, \quad n = 1, 2, \ldots \ . \tag{4.1}$$

$$J_n = T + (Q_{n+1} - Q_n) = T + \Delta Q(n) \ . \tag{4.2}$$

The normalized jitter with respect to the original data stream period (inter-departure time between nth and (n+1)th packets transmitted from the source), T:

$$\tilde{J}_n = J_n - T = \Delta Q(n), \quad n = 1, 2, \ldots, \text{ where } \Delta Q(n) \overset{def}{=} Q_{n+1} - Q_n \ . \tag{4.3}$$

From (4.3), we can observe the study of jitter turns into an equivalent study of the queue size variation $\Delta Q(n)$ of a selected user's packet arrival, which is the methodology that is going to be applied in this study. We want to find the pdf of $\tilde{J}_n$, i.e., $\Pr\{\tilde{J}_n\}$, from the queue length distribution of the steady state probability of the buffer with queue length q ($P(Q>q)$), q = 1, 2, … . A discrete-time approach is used in the delay analysis, which a node/multiplexer is modeled as a discrete time fluid queue where the time is slotted. It is assumed that the packets arrive and depart at the beginning of a slot, which leads to a discrete time queueing process.

In summary, the approach here is

- First, apply the Gaussian traffic modeling using the MVA approach [25][25] to conduct the queue length analysis. Obtain the tail probability $P\{Q>q\}$ based on the MVA approach for the discrete time queueing model [25][26] to conduct the queue length analysis, which is used in heterogeneous jitter analysis. This is done in section 4.2. At a node in high-speed networks, the packet arrival that consist of heterogeneous type traffic includes real time and non-real time integrated traffic streams from different sources (e.g., voice, audio, video, image, plain text and data of other newly developed Internet applications). A single link will carry hundreds or even thousands of applications, which apparently lead to the application of Central Limit Theorem, that the network traffic can be modeled by a Gaussian stochastic process.

- Second, find the pdf of the jitter probability from the queue length distribution. This is done in section 4.3.

4.2 Discrete Time Fluid Queue Model and Extreme Value [26]

The multiplexer is modeled as a fluid queue serving a large number of independent sources.

In the discrete time model, the fluid is permitted to flow in and out of the buffer only at discrete time intervals denoted in *k*. Let *N* be the total number of sources served by the multiplexer, and $X_k^{(s)}$ be the amount of traffic that arrives from source *s* into the buffer at time *k*, *s* =1, 2, …, *N*. Further, let $X_k = \sum_{s=0}^{N} X_k^{(s)}$ and $X_k^{(0)} \overset{\Delta}{=} -\mu$, which means that X_k is the aggregated fluid arriving at time *k*, and less the system capacity μ. Let $Q_k \geq 0$ be the amount of fluid in the buffer at time *k*, less than the capacity μ, and the Q_k for the infinite buffer case can be represented using Lindley's equation [26]

$$Q_k = (Q_{k-1} + X_k)^+ = \max\{Q_{k-1} + X_k, 0\}. \tag{4.4}$$

Lets say that the fluid queue has always been operational, that is $k \in \mathbf{z}$ = {…, -1, 0, 1, …} (instead of time starting at *k* = 0). As long as X_k is a stationary and ergodic process, and with $\overline{X} = E\{X_k\} < 0$ (the stability condition for an infinite buffer queuing system, where the arrival rate must be less than the service rate), (4) can uniquely determine a stationary process ($\widetilde{Q}_k$) that satisfies the equation, where for all $k \in \mathbf{Z}$ = {…, -1, 0, 1, …} [26][45] ($k \in \mathbf{Z}$ from this point on). Due to X_k being an ergodic stationary process, the tail probability of buffer queue length q at time k ($P(Q_k > q)$) converges to a steady state tail probability of buffer queue length q ($P(Q \ \ > q)$) regardless of the initial condition *Q₀*. Thus, for all $k \in \mathbf{Z}$, $P(\widetilde{Q}_k > q)$ is equal to $P(Q \ \ > q)$, since the tail probability does not depend on the initial condition *Q₀*, and the tail probability of a stationary process is the tail probability itself, where $P(\widetilde{Q}_k > q) = P(Q \ \ > q)$ [26][45]. The relation between an ergodic stationary process X_k and the corresponding stochastic process $\widetilde{Q}_k$, where $k \in \mathbf{Z}$ [26]

$$\widetilde{Q}_k = \sup_{m \geq 0} \sum_{i \geq 1}^{m} X_{k-i} \tag{4.5}$$

where m = 0, 1, 2, … .Since, $P(\widetilde{Q}_k > q) = P(Q \ \ > q)$ for all k [26],

$$P(Q > q) = P(\widetilde{Q}_k > q)$$

$$= P\left\{ \sup_{m \geq 0} \sum_{i \geq 1}^{m} X_{-i} > q \right\}. \tag{4.6}$$

From (4.5) and (4.6), we can see that the steady state tail probability $P(Q > q)$ can be determined by summing all previous stationary aggregate fluid arrival process. In other words, the probability of the tail probability $P(Q > q)$ is the same as the tail probability of the supremum of a stochastic process denoted by $W_m = \sum_{i=0}^{m} X_{-i}$, where $P(Q \ > q) = P\left\{ \sup_{m \geq 0} W_m > q \right\}$, m = 0, 1, 2, … [26]. This stochastic process, W_m, is called the backward accumulation process, and it is not stationary. The first two moments of W_m are defined as below [26]

$$E\{W_m\} = m\overline{X} \tag{4.7.a}$$

$$Var\{W_m\} = lC_X(0) + 2\sum_{i=1}^{m-1}(m-i)C_X(i) \tag{4.7.b}$$

where $C_X(0) \overset{\Delta}{=} E\left\{(X_k - \overline{X})(X_{k+l} - \overline{X})\right\}$ is the auto-covariance function of X_k. Hence, $P(Q \ > q)$ can be calculated by determining the tail probability of the supremum of the backward accumulative process W_m, where the Extreme Value Theory is applied to this supremum of a stochastic process to obtain a simple approximation for $P(Q \ > q)$ [26], which is shown in the next subsection below.

Evaluating *P*(*Q*>*q*) at a multiplexer with Gaussian Arrival Process using the MVA approach

We can obtain a simple approximation for *P*(*Q*>*q*) by applying the results from the theory in section 4.2, where determining the tail probability *P*(*Q*>*q*) of a queueing system is mapped to that of determining the tail probability of the supremum of the backward accumulation process W_m (here only the discrete case is our concern) [26].

Let $X_k^{(0)} \overset{\Delta}{=} -\mu$ denote the link capacity of a multiplexer (e.g., an ATM multiplexer). Assume that $X_k^{(s)}$, the fluid arrival process of source *s* at time *k*, *s* = 1, …, *N*, are independent mean ergodic stationary arrival processes with finite mean $\overline{X}_s$, and auto-covariance $C_s(l)$. $A_k \overset{\Delta}{=} \sum_{s=1}^{N} X_k^{(s)}$ is defined to be the *aggregate arrival process*. Then, A_k is also a mean ergodic stationary arrival process with mean $\overline{A} = \sum_{s=1}^{N} \overline{X}_s$ and auto-covariance $C_A(l) = \sum_{s=1}^{N} C_s(l)$. So, $X_k = \sum_{s=0}^{N} X_k^{(s)} = A_k - \mu$ is a mean ergodic stationary process with mean $\overline{X} = \overline{A} - \mu$ (where $\overline{X} = E\{X_k\} = \sum_{s=0}^{N} X_k^{(s)} = \sum_{s=1}^{N} X_k^{(s)} + (-\mu)$, since $X_k^{(0)} \overset{\Delta}{=} -\mu$) and auto-covariance $C_X(l) = C_A(l)$. In other words, X_k is the aggregate fluid arriving at time *k*, less the capacity μ. As long as $\overline{X} < 0$, which means that the arrival rate must be less than the service rate, the infinite buffer queueing system is stable.

Now, model the arrival process A_k as a stationary *normal process*. The main reason to consider Gaussian traffic modeling is the CLT [25][26]. As the network grows, the number of independent traffic sources aggregating at a node (e.g., router, switch, multiplexer) increase, and the shape of the traffic distribution will become closer and closer to Gaussian. A_k has both negative and positive components. The negative component in [26] imply that the buffer is empty due to the arrival process, which is not possible in reality. However, as the mean $\bar{A}$ of the aggregate process is typically several times larger than its standard deviation $\sqrt{C_A(0)}$, the probability of this negative flow is negligible.

Since A_k is modeled as a normal process, both X_k and W_m are normal processes. Now, rewrite the arrival process, where determining the extreme value distribution of W_m

is equivalent to determining the extreme value distribution of a new stochastic process $\{Z_m : m = 0,1,\ldots\}$, which is defined as [26]

$$Z_m \overset{\Delta}{=} \frac{W_m + \mu(1-\rho)m}{\mu((1-\rho)m+k)} \tag{4.8}$$

where $\rho \overset{\Delta}{=} \frac{\bar{A}}{\mu}$ is the utilization of the ATM multiplexer, $k \overset{\Delta}{=} \left\lceil \frac{q}{\mu} \right\rceil$ is the time it takes for the fluid in the buffer at level q to empty, and μ is the service rate. Z_m is the standardized maximum value, where $\mu((1-\rho)m+k)$ and $\mu(1-\rho)m$ are the scale and location parameters, respectively. Also, Z_m is a normal process with [26]

$$E\{Z_m\} = 0 \tag{4.9.a}$$

$$Var\{Z_m\} = \frac{Var\{W_m\}}{\mu^2((1-\rho)m+k)^2}\,. \tag{4.9.b}$$

The extreme value distribution is used to quantify the probabilistic behavior of unusual or rare events, and provide a better fit to the data set.

Justification of the new stochastic process $\boldsymbol{Z_m}$: Applying Central Limit Theorem, assume that $\{B_t : t \geq 0\}$ is a Gaussian process with stationary increments such that $B_0 = 0$, and define $\varsigma := -E(B_t) = -E\left(\sum_{n=1}^{N} B_t^{(n)}\right)$, and $v(t) := \mathrm{var}(B_t)$, where each $B_t^{(n)}$ is an individual stream. Expressing $P(B_t > q)$ in terms of ς, $v(t)$, and the standard Gaussian tail function $\psi(\alpha) := \int e^{-z^2/2} dz \Big/ \sqrt{2\pi}$, we obtain a multivariate version of the CLT of the normalized process [58]

$$P(B_t > q) = \psi\left(\frac{q + \varsigma t}{\sqrt{v(t)}}\right). \tag{4.10}$$

Now, apply the Dominant Time-Scale t_q and the Maximum Variance Asymptotic approach. The term $\frac{v(t)}{(q+\varsigma t)^2}$ should attain its maximum value at some finite point $t = t_q$, at which $P(B_t > q)$ attains it maximum. For Gaussian processes, t_q, the dominant time scale is also the time at which it attains its maximum variance value $v(t)$. $P(Q > q) = P(B_t > q)$ is largely dominated by $P(B_{t_q} > q)$ [25][26][58].

The t_q should be a local maximum point of $\log[v(t)/(q+\varsigma t)^2]$ that lies in the open set $\{t : v(t) > 0\}$, t_q must satisfy [58][59]

$$0 = \left[\frac{d}{dt}\log\frac{v(t)}{(q+\varsigma t)^2}\right]_{t=t_q}. \tag{4.11}$$

Note: Eq. (4.10) and (4.11) will hold for discrete time processes.

For notation simplicity, under (4.10) and (4.11), the new stochastic process Z_m is defined as

$$Z_m = \frac{W_m + \chi m}{q + \chi m} \tag{4.12}$$

and define $k \overset{\Delta}{=} \left\lceil \frac{q}{\mu} \right\rceil$ and $\rho = \frac{\overline{A}}{\mu}$

where

$$\begin{aligned}\chi m &:= -E(W_m) = -mE(\overline{X})\\ &= -\left[E(\overline{A} - \mu)\right]m\\ &= -\left[E(\overline{A}) - \mu\right]m\\ &= -m\left[\rho\mu - \mu\right]\end{aligned}$$

$$= -\mu[\rho - 1]m$$

$$= \mu[1 - \rho]m \tag{4.13}$$

Substitute (4.13) into (4.12), we have

$$Z_m = \frac{W_m + \mu[1-\rho]m}{q + \mu[1-\rho]m}$$

$$= \frac{W_m + \mu[1-\rho]m}{k\mu + \mu[1-\rho]m}$$

$$= \frac{W_m + \mu[1-\rho]m}{\mu([1-\rho]m + k)}. \tag{4.14}$$

Proposition 1 [26] $W_{m>q}$ if and only if $Z_m > 1$

From proposition 1 and (4.6), the relationship between the steady state tail probability of the buffer queue length *q* and the extreme value distribution of Z_m and W_m

$$P(Q > q) = P\left(\sup_{m\geq 0} W_m > q\right)$$

$$= P\left(\sup_{m\geq 0} Z_m > 1\right). \tag{4.15}$$

From the Dominated Convergence Theorem [26][48], the equations (4.16) and (4.17) show that $Var\{Z_m\} \to 0$ as $m \to \infty$ if $\sum_{k=-\infty}^{\infty} |C_X(k)| < \infty$ (a sufficient condition for the ergodicity of a stationary process [49]),

$$\lim_{m\to\infty} \frac{1}{m} Var\{W_m\} = \sum_{k=-\infty}^{\infty} C_X(k) \tag{4.16}$$

and

$$\lim_{m\to\infty} mVar\{Z_m\} = \frac{\sum_{k=-\infty}^{\infty} C_X(k)}{\mu^2(1-\rho)^2}. \tag{4.17}$$

Fig. 4.1 illustrates a plot of the variance of Z_m versus m, it shows that $Var\{Z_m\}$ must reach its maximum value at some finite $m = m_{\max} \geq 0$. Next, the extreme/supremum value distribution will be applied to approximate the tail probability.

Maximum Variance Approximation**:** "For a zero mean normal process Z_m that is correlated in time and whose variance $Var\{Z_m\}$ achieves a maximum value for some finite $m_{\max}$ [26],"

$$P\left(\sup_{m\geq 0} Z_m > u\right) \approx P\left(Z_{m_{\max}} > u\right) \tag{4.18}$$

for *sufficiently large u* [26]. Fig. 4.1 illustrates this point by plotting $Var\{Z_m\}$ vs. m. For eq. (4.18), the right hand side approximation is a lower bound of the left hand side, that is $P\left(\sup_{m\geq 0} Z_m > u\right) \geq P\left(Z_{m_{\max}} > u\right)$ [26]. "Typically, $u > 1.5\sqrt{Var\{Z_{m_{\max}}\}}$ is sufficiently large for the approximation to work well [26]". Thus, if $Var\{Z_m\} << 1$, then the tail probability $P(Q>q)$ can be approximated as [26]

$$P(Q > q) \approx P\left(Z_{m_{\max}} > 1\right)$$

$$= \int_{\frac{1}{\sigma_{\max}}}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) dx. \tag{4.19}$$

This is because the probability that $P(Q>q)$ is much less that 1, the condition $Var\{Z_m\} << 1$ readily hold for this analysis [26].

Theorem A (Theorem A in [56], Theorem D.4 in [57]) "Let $\{\varsigma_t : t \in [L,U]\}$ be a zero mean (centered) Gaussian process, and suppose that there exist constants a and γ such that $E\{(\varsigma_s - \varsigma_t)^2\} \le a|t-s|^\gamma$ for all t, $s \in [L,U]$. Then, there exists a constant K determined only by a and γ, such that for any $A \subset [L,U]$ and y,

$$P(\{\langle\varsigma\rangle_A > y\}) \le K(U-L)y^{2/\gamma}\psi\left(\frac{y}{\langle\sigma\rangle_A}\right) \tag{A.1}$$

where $\langle\sigma\rangle_A = \sup_{t\in A}\sqrt{Var\{\varsigma_t\}}$".

Summary of the algorithm for obtaining the tail probability *P*(*Q*>*q*) based on the MVA approach for discrete time queueing model [26]:

1. Given individual arrival processes with mean $\overline{X}_s$ and auto-covariance $C_s(l)$, compute the aggregate mean $\overline{A} = \sum_{s=1}^{N}\overline{X}_s$, and auto-covariance $C_A(l) = \sum_{s=1}^{N}C_s(l)$ of the aggregate arrival process.
2. Compute $Var\{Z_m\}$ using Eq. (7.b) and (9.b) as

$$Var\{Z_m\} = \frac{mC_A(0) + 2\sum_{i=1}^{m-1}(m-i)C_A(i)}{\mu^2((1-\rho)m+k)^2}$$

   where $\mu > \overline{A}$ is the link rate, $\rho = \frac{\overline{A}}{\mu}$, and $k \overset{\Delta}{=} \left\lceil \frac{q}{\mu} \right\rceil$.
3. Determine $\sigma_{\max}^2$, the maximum value of $Var\{Z_m\}$.
4. Finally, compute the tail probability $P(Q>q) \approx P(Z_{m_{\max}} > 1) = \int_{\frac{1}{\sigma_{\max}}}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}dx$,

   where $Z_{m_{\max}}$ is the normal random variable.

The typical plot of $Var\{Z_m\}$ versus m is shown in Fig. 4.1. This analysis is done at an ATM multiplexer with the following assumptions: q=1000, μ=0.61, N=100, where individual arrival process has a mean $\overline{X}_s$= 10 cells/sec, and auto-covariance of Gaussian process $C_s(l) = 100e^{-0.04l}$.

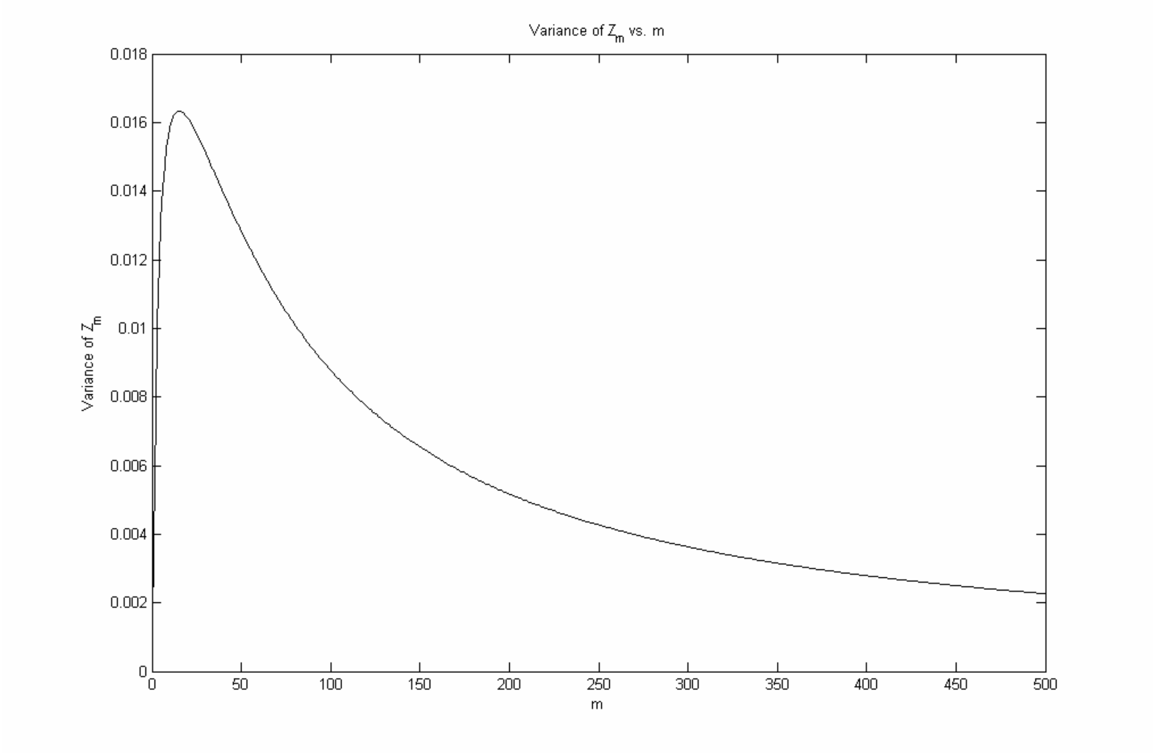


Fig. 4.1[1] Plot of variance of Z_m versus m. The variance reaches a maximum value and then begins to converge to 0. The parameters used are q=1000, μ=0.61, N=100, each individual arrival process with mean $\overline{X}_s$= 10 cells/sec, and auto-covariance $C_s(l) = 100e^{-0.04l}$.

[1] This figure is attempted to reproduce the result of Figure 3 in [26]. Both graphs have the same distribution shape but values may not be exactly the same, because the parameters used in [26] to plot Figure 3 are not provided by reference [26]. So, we cannot regenerate the plot with the same set of parameters used in [26] (which is unknown to us). The attempt here is to use the MVA theory and plot the Fig. 4.1 using a set of predicted parameters to regenerate the plot that is similar to Figure 3 in [26].

4.3 Jitter Analysis of Heterogeneous Traffic Networks

Assume the packets arrive and depart at the beginning of a slot, which leads to a discrete time queueing process. The time is slotted with the size being equal to the transmission of a packet. An arrival time slot is defined as the time interval [t-1, t), where t is a non-negative integer (i.e., $t \geq 0$). Here we define the nth packet arrival time slot interval as [nT, (n+1)T) for the packet from the tagged stream (stream of interest), where n and T are non-negative integers. The background traffic is the superposition of all the other traffic competing for the resources with the tagged stream at a node. In heterogeneous high-speed network, jitter experienced by real-time traffic can be worsening if traffic is not being regulated or some type of congestion control mechanism is not applied. Thus, a service discipline similar to the distortion-reducing peak output-rate enforcing (PORE) [50] is used to prevent the delay jitter increasing without bound. Let's refer to this as a tagged stream adaptive PORE (APORE) service discipline. The APORE service strategy guarantees the packets belong to a tagged stream are transmitted at a minimum spacing of $A_{\min}$ slots, where $A_{\min} = T$ -1. Whenever the queue level $q < A_{\min}$, the server delays providing service to the tagged stream, instead provides first-come-first-serve service to the background traffic. Thereby, it ensures the tagged stream experiences minimum jitter (see Fig. 4.4).

Proposition: The probability that a jitter size of j will exist for the nth packet of the tagged stream is given by:

$$P\{\tilde{J}_n = j\} = P\{\tilde{J}_n = j \mid A_{n+1} - 1 = q\} \cdot P\{A_{n+1} \ - 1 = q\}$$

$$= \frac{1}{2} P(Q > q) f_q(j), \qquad 0 < q \leq A_{\min} \tag{4.20}$$

where $1 \le |j| \le (T-1)$ and A_{n+1} indicates the total number of packets that arrive in the (n+1)th time slot group. In (4.20), the probability of q number of data packets that arrive among the available (T-1) slots will be served.

A special case of (4.20) is the zero jitter case (i.e., j = 0), which is simply

$$\mathrm{P}\left\{\widetilde{J}_n = 0\right\} = 1 - 2\sum_{j=1}^{(T-1)} P\left\{\widetilde{J}_n = j\right\}. \tag{4.21}$$

***Derivations & Explanations**:* In (4.20), the term $P\left\{A_{n+1} \; -1 = q\right\}$ is the aggregated traffic arrival distribution that approximates the number of arrivals in the arrival slot of a tagged stream packet. It is based on the consideration of the probability the packets that arrive within the time slot group, which will be served ahead of the tagged packet. Subtract 1 from the total number of aggregated packets that arrive in the (n+1)th time slot group (i.e., $A_{(n+1)}$) due to the fact that the tagged frame is also among the overall traffic. The traffic distribution in eq. (4.20) is based on

$$P\left\{A_{n+1} \; -1 = q\right\} = \frac{1}{2} P(Q > q)$$

$$\approx \frac{1}{2} P\left(\left\{Z_{m_{\max}} \; > 1\right\}\right) = \frac{1}{2} \int_{\frac{1}{\sigma_{\max}}}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx \tag{4.22}$$

where $q \in [1, (T-1)]$, calculated using the MVA approach discussed in section 4.2. The discrete triangle distribution applied is a symmetric distribution, which represents the probability that the time slot position of two sequentially arriving frames will be offset by a jitter amount of +j or –j. Thus, the triangle distribution for the range [-q, q] is applied as

$$P\left\{\widetilde{J}_n = j \mid A_{(n+1)} - 1 = q,\right\} = f_q(j)$$

$$= \begin{cases} \dfrac{1}{q+1} - \dfrac{j}{(A_{\min}+1)^2}, & \mathit{for}\ |j| \le q \\ 0, & \mathit{otherwise} \end{cases} \tag{4.23}$$

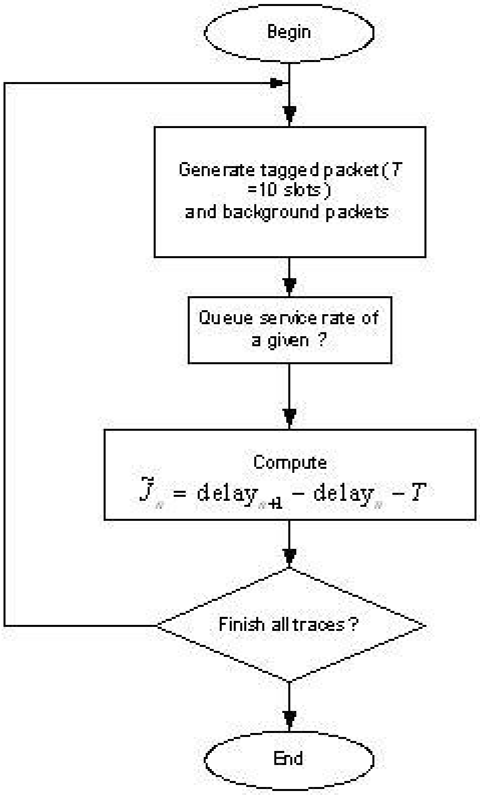
which provides the normalized jitter probability of j given that q packet arrivals enter the server before the (n+1)th tagged packet of the tagged stream, among the overall x possible time slots, $0 \le x \le T$.

4.3.1 Simulation and Numerical Study

The following simulation and numerical results focus on the delay jitter in terms of pdf on the performance of a tagged stream under different traffic conditions. For the illustrative purpose, the bursty Internet traffic trace is used in this evaluation. The simulation environment has been set up in such a way that the information bit stream (e.g., video, multi media, digitized voice, etc) is packetized into fixed-length ATM cells (packets). The cells are transmitted using fixed inter-arrival time but the bit length is varied from cell-to-cell (packet-to-packet). Accordingly, the delay jitter is defined as the delay jitter experienced between two successive packets (e.g., the nth and n+1th packets from the tagged stream), where the delay jitter is calculated based on (4.3) (see Flowchart 1). For the experiment, a real Internet traffic trace is used, **BC-pAug89** packet traces of LAN and WAN traffic seen on an Ethernet. We assume that Ethernet traffic sources are multiplexed on to an ATM network, where traffic is segmented into small and fixed sized cell/packet in order to achieve small delay and delay jitter [51][52]. The reason we base our analysis on ATM networks is because still ATM remains as the most widely used Internet backbone protocol at this time. The traces are in byte units, which is converted into ATM cell 53 bytes. The variable size Ethernet packets are segmented in to constant size of 53 bytes packets (cells). Thus, the traffic transmitting in the network is fixed size packets and one packet is transmitted in each time slot.

Flowchart 1 is the algorithm of delay jitter calculation at a multiplexer/node. The queue consists of a tagged stream packet and background stream packets. The tagged stream and background stream are segmented into ATM 53 bytes packets. There are a random number of background stream packets that enter the queue before the tagged stream packet. Whenever the queue level $q < A_{\min}$, the server delays providing service to the tagged stream, instead provides first-come-first-serve service to the background traffic. The traffic trace of **BC-pAug89** is re-sampled at 10 ms. Assume the average processing rate and overhead delay is 1ms per cell, the inter-arrival time between two consecutive cells is 10 ms, which will be equivalent to 10 time slots (for generate $T = 10$ tagged stream). Through the numerical example, we demonstrate the accuracy of the closed form delay jitter approximation of eq. (4.20) shown in Fig. 4.2 of a 95% confidence interval (CI), for self-similar traffic of heterogeneous networks. Here, we calculate the 95% CI for the delay jitter proportion between the real data set and the numerical analysis (eq. (4.20)). The width of the 95% CI reveals the degree of uncertainty in the estimate of the treatment effect; in another words, this is the interval which includes the true value with 95% certainty. The real data set is used to obtain the empirical pdf function for the delay jitter. Then, eq. (4.20) is used to approximate the delay jitter pdf function of the data set. Subsequently, confidence intervals for the estimations are computed. The confidence interval bounds on the standard deviation between the empirical and numerical values.

In Fig 10, we compare the approximation eq. (4.20) and the simulation results of the jitter distribution at a multiplexer serving various Internet applications. Fig. 4.2 shows that the delay jitter increases when the traffic utilization increases from 50% to 90%, for T equals to 10. In Fig. 4.3, we can observe that the tagged stream with a large period will tend to experience large delay variation as the utilization increases from 50% to 70%, where the T increases from 10 from 20. Fig. 4.4 illustrates the effectiveness of the control algorithm, the APORE service strategy that guarantees the packets belong to a tagged stream are transmitted at a minimum spacing of $A_{\min}$ slots, where $A_{\min} = T - 1$. If we choose the $A_{\min}$ to be larger than T -1, then we will introduce more delay variation to the tagged stream, as shown in Fig. 4.4 the probability of jitter increases as $A_{\min} > T-1$. We will not consider $A_{\min} < T-1$, because this will increase the possibility that a packet from the tagged stream may not be arrived to a node even it is its turn for entering service.


Begin
Generate tagged packet (T =10 slots)
and background packets
Queue service rate of a given ?
Compute
$J_n = \text{delay}_{n+1} - \text{delay}_n - T$
Finish all traces ?
End

Flowchart 1 The simulation algorithm used to calculate delay jitter at a multiplexer.

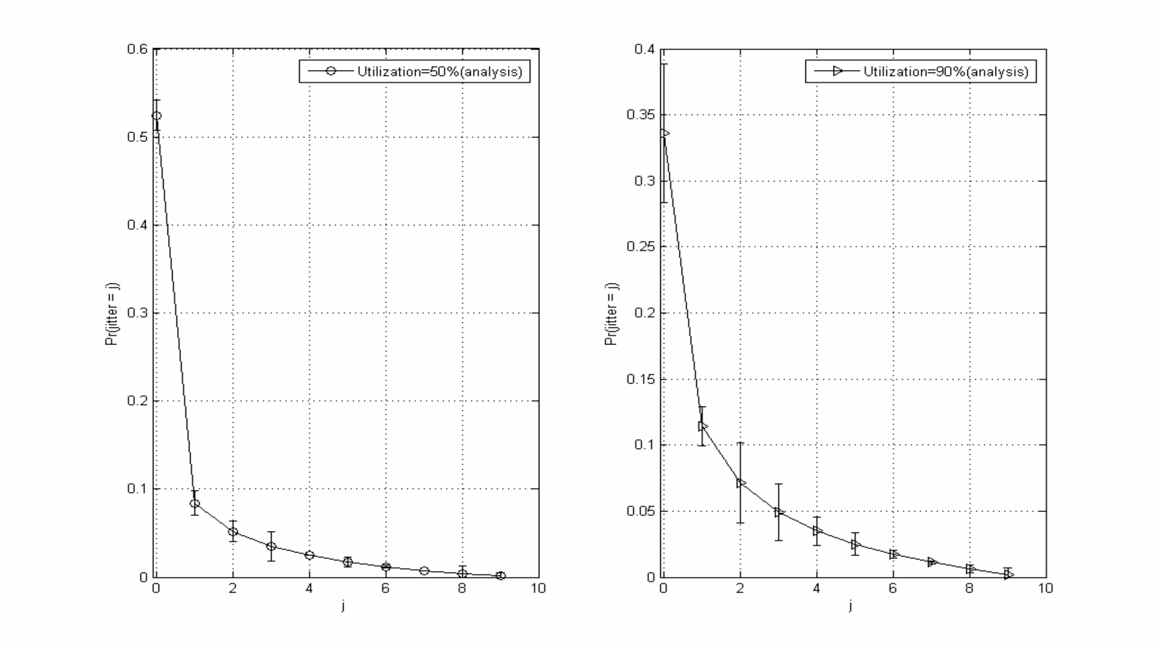


Fig. 4.2 Comparison between the approximation equation and the simulation results of the jitter distribution for utilization of 50% and 90% with 95% CI, for the case of $T = 10$.

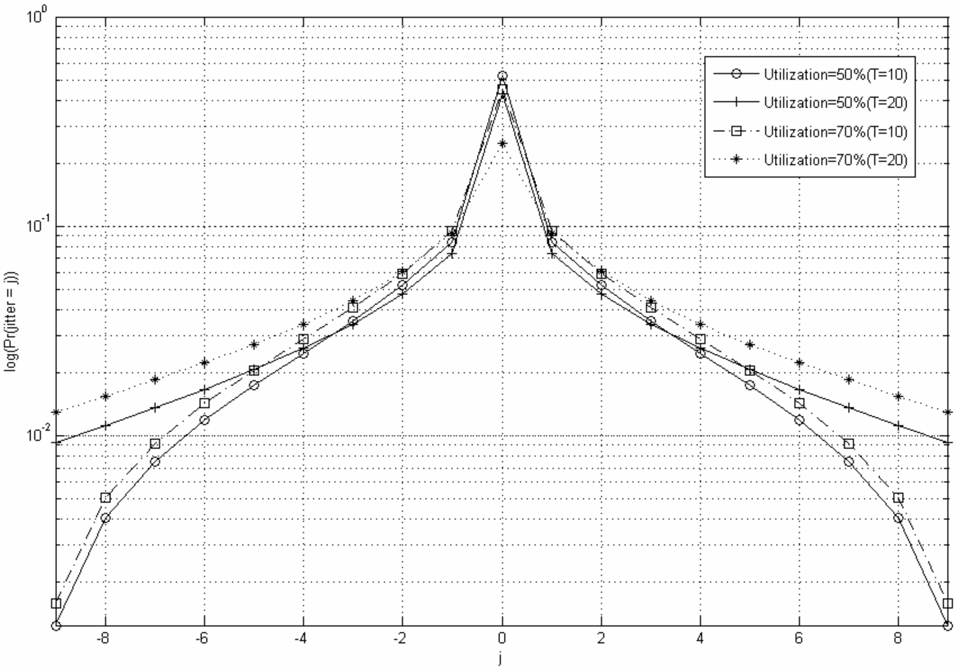


Fig. 4.3 Distribution of jitter for the comparison between utilization of 50% and 70%, and for the comparison between $T = 10$ and 20.

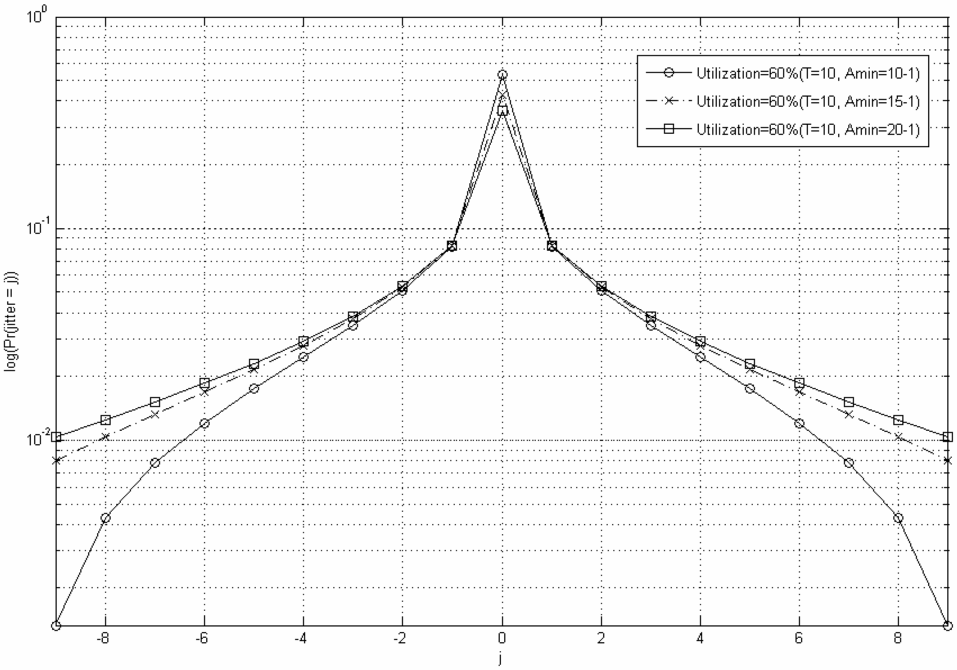


Fig. 4.4 Distribution of jitter for utilization of 60%, T=10, and various $A_{\min}$ values.

4.4 Jitter Analysis of Heterogeneous Traffic Networks with Differentiated Services

In this section, the jitter characteristic of the priority class in the DiffServ network is analyzed.

In order to provide high quality service, end-to-end quality of service (QoS) support is required. Several network models and mechanisms have been proposed by the Internet Engineering Task Force (IETF) to improve the QoS of integrated service networks by deploying differentiated service (DiffServ), traffic engineering (TE), and constraint-based routing (CR) [53]. The realization of DiffServ in networks is one of the essential focuses of TE technologies under development. Recently, in wide area networks (WANs), protocols like multiprotocol label switching (MPLS) and generalized MPLS (GMPLS) are being developed with this focus of enabling effective TE with DiffServ capabilities.

A basic differentiated service scheme can be provided by a set of priority scheduling algorithms. The traffic flows are aggregated according to their belonging to a certain class type called the per-hop behavior (PHB) [54]. The traffic classification is done in qualitative terms, which are based on vendor concerns [54][55]. For example, the network administrator may configure the service rate of 50%, 35%, and 15% for classes 1, 2, and 3 respectively. Class 3 may correspond to delay jitter insensitive traffic, where the least bandwidth is being enforced for this class (e.g., best effort Internet traffic).

Therefore, the research presented in this dissertation proposes a methodology to directly compute the QoS performance quantity (e.g., probability distribution of inter-arrival jitter) of network deploying differentiated services. The analytic models are applied in analyzing the contribution of multiple-class of services in Internet networks with respect to inter-arrival packet jitter. In addition, the relationship between inter-arrival packet jitter and the system characteristics, such as the server utilization and priority, is also investigated. This tool will help to provide an analysis of network routers and network streams for various system server topologies in order to assist network system designers in the hardware/firmware development and estimate the quality of service of DiffServ networks with heterogeneous type traffic, which includes real time and non-real time integrated traffic streams from different sources (e.g., voice, audio, video, image, plain text and data of other newly developed Internet applications).

DiffServ capable networks provide end-to-end QoS control by classifying and assigning different classes of services onto the incoming packets. Hence, at the base station's router/switch, the packets are buffered before being multiplexed onto the high-speed link following a non-preemptive HOL priority scheme, where the packets that have been postponed in services will wait at the head of the line of its equivalent class until all higher priority packets have been cleared out of the server. The APORE service control algorithm strategy that guarantees the packets belong to a tagged stream are transmitted at a minimum spacing of $A_{\min}$ slots, where $A_{\min} = T$ -1. Assume class 1 has priority over class 2, class 2 has priority over class 3, and so on. Among the same priority the first in first out (FIFO) transmission ordering applies.

In the discrete time model, the fluid is permitted to flow in and out of the buffer only at discrete time interval denoted by *k*. Let N be the total number of sources served by the multiplexer. Define $X_{k,d}$ to be the aggregated fluid of class *d* arriving at time *k*, and less the system capacity μ_d of class *d*, where $X_{k,d} = \sum_{s=0}^{N} X_{k,d}^{(s)}$ and $X_{k,d}^{(0)} \overset{\Delta}{=} -\mu_d$. Here, we assume that each fluid arrival process of class *d* corresponding to source s, $X_{k,d}^{(s)}$, is an independent mean ergodic stationary process with mean $\overline{X}_{s,d}$ and auto-covariance $C_s(l)$. Let's denote $A_{k,d} \overset{\Delta}{=} \sum_{s=1}^{N} X_{k,d}^{(s)}$ to be the *aggregate arrival process*. Then, A_k is also a mean ergodic stationary arrival process with mean $\overline{A}_d = \sum_{s=1}^{N} \overline{X}_{s,d}$ and auto-covariance $C_{A_d}(l) = \sum_{s=1}^{N} C_{s,d}(l)$. Thus, $X_{k,d} = A_{k,d} - \mu_d$ is a mean ergodic stationary process with mean $\overline{X}_d = \overline{A}_d - \mu_d$ (where $\overline{X}_d = E\{X_{k,d}\} = \sum_{s=0}^{N} X_{k,d}^{(s)}$) and auto-covariance $C_{X_d}(l) = C_{A_d}(l)$. Let $Q_{k,d} \geq 0$ be the amount of fluid corresponding to class d in the buffer at time k, less than the capacity μ_d, and $Q_{k,d}$ for the infinite buffer case can be represented using Lindley's equation

$$Q_{k,d} = (Q_{k-1,d} + X_{k,d})^+ = \max\{Q_{k-1,d} + X_{k,d}, 0\}. \tag{4.24}$$

Let say that the fluid queue has always been operational, that is $k \in \mathbf{Z} = \{\ldots, -1, 0, 1, \ldots\}$ (instead of time starting at $k = 0$). Due to $X_{k,d}$ being an ergodic stationary process, the tail probability of the buffer queue length q at time k ($P(Q_{k,d} > q)$) converges to a steady state tail probability of buffer queue length q ($P(Q_d > q)$) regardless of the initial condition. The steady state queue length distribution corresponds to the supremum distribution of an ergodic stationary process $X_{k,d}$ and the corresponding stochastic process $\widetilde{Q}_{k,d}$. The maximum amount of fluid in the system at time k can be expressed as

$$\widetilde{Q}_{k,d} = \sup_{m \geq 0} \sum_{i=0}^{m-1} X_{k-i,d} \tag{4.25}$$

where m = 0, 1, 2, … . On the supremum distribution of Integrated Stationary Gaussian process with linear drift, the limiting (steady state) queue length distribution corresponds to the supremum distribution,

$$P(Q_d > q) := \lim_{k\to\infty} P(\widetilde{Q}_{k,d} > q)$$

$$= P\left(\left\{\sup_{m\geq 0}\sum_{i=0}^{m-1} X_{-i,d} > q\right\}\right). \qquad (4.26)$$

Let the supremum of a stochastic process for class d defined by $W_{m,d} \overset{\Delta}{=} \sum_{i=0}^{m-1} X_{-i,d}$. Based on the Extreme Value Theory, we can study the supremum distribution of Q_d through the supremum stochastic process $W_{m,d}$. Note that this stochastic process $W_{m,d}$ is not stationary. Now, $A_{k,d}$ is modeled as a stationary normal process, then both X_k and W_m are normal processes. This is justified by applying the Central limit Theorem, which mentioned in the introduction that the multiplexer will serve a large number of independent sources. Thus, we study the supremum distribution of $W_{m,d}$ through the new supremum stochastic process $Z_{m,d}$, where $Z_{m,d}$ is a centered (zero mean) Gaussian process. So, determining the extreme distribution of $W_{m,d}$ is equivalent to determining the extreme distribution of $Z_{m,d}$. Assume there is a total of D class of services, the extreme value distribution of a new stochastic process of class $d\{Z_{m,d} : m = 0,1,\ldots;\ d = 1,2,\ldots,D\}$, which is defined as

$$Z_{m,d} \overset{\Delta}{=} \frac{W_{m,d} - \mu_d(1-\rho_d)m}{\mu_d((1-\rho_d)m + k)} \qquad (4.27)$$

where $\rho_d \overset{\Delta}{=} \frac{\bar{A}_d}{\mu_d} = \sum_{c=1}^{d} \rho_c = \frac{\sum_{c=1}^{d} \sum_{s=1}^{N} \overline{X}_{s,c}}{\mu_d}$ is the load of class *d* at the multiplexer, $k \overset{\Delta}{=} \left\lceil \frac{q}{\mu_d} \right\rceil$ is the time it takes for the fluid in the buffer at level *q* to empty, and μ_d is the service rate of class *d*. Based on this we can define $Z_{m,d}$ as a normal process with

$$E\{Z_{m,d}\} = 0 \tag{4.28.a}$$

$$Var\{Z_{m,d}\} = \frac{Var\{W_{m,d}\}}{\mu_d^{\ 2}((1-\rho_d)m+k)^2}. \tag{4.28.b}$$

Based on the Maximum Variance Approximation, the tail probability of class *d*, $P(Q_d > q)$, can be approximated as

$$P(Q_d > q) \approx P(Z_{m_{\max},d} > 1)$$

$$= \int_{\frac{1}{\sigma_{m_{\max},d}}}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) dx. \tag{4.29}$$

This is based on the same argument proved in [26], which is explained in section 4.2 above. Due to the probability that $P(Q_d > q)$ is much less that 1, the condition $Var\{Z_{m,d}\} << 1$ readily hold for this study.

Proposition: The probability that a jitter size of *j* will exist for the *n*th packet of the tagged stream for a class *d* traffic source is given by

$$P\{\tilde{J}_{n,d} = j\} = P\{\tilde{J}_{n,d} = j \mid A_{n+1,d} - 1 = q\} \cdot P\{A_{n+1,d} - 1 = q\}$$

$$= \frac{1}{2} P(Q_d > q) f_{q,d}(j), \qquad 0 < q \le A_{d_{\min}} \tag{4.30}$$

where $1 \le |j| \le (T-1)$ and $A_{n+1,d}$ indicates the total number of packets that arrive in the (*n*+1)th time slot group of class *d* and higher priority classes compare to class *d*. In eq.

(4.30), the probability of *q* number of data packets that arrive among the available (*T*-1) slots will be served.

A special case of (4.30) is the zero jitter case (i.e., *j* = 0), which is simply

$$\mathrm{P}\left\{\widetilde{J}_{n,d}=0\right\}=1-2\sum_{j=1}^{(T-1)}\sum_{c=1}^{d}P\left\{\widetilde{J}_{n,c}=j\right\}. \tag{4.31}$$

***Derivations & Explanations**:* In (4.30), the term $P\{A_{n+1,d}-1=q\}$ is the aggregated traffic arrival distribution of class 1 up to class *d*, that approximates the number of arrivals in the arrival slot of a class *d* tagged stream packet. It is based on the consideration that the assignment probability of the packets that arrive within the time slot group, which will be served ahead of the tagged packet. Subtract 1 from the total number of aggregated packets that arrive in the (*n*+1)th time slot group (i.e., $\boldsymbol{A}_{(\boldsymbol{n}+1),\boldsymbol{d}}$) due to the fact that the tagged frame is also among the overall traffic. The traffic distribution in (4.30) is based on

$$P\{A_{n+1,d}-1=q\}=\frac{1}{2}P(Q_d>q)$$

$$\approx\frac{1}{2}P\left(\left\{Z_{m_{\max},d}>1\right\}\right)=\frac{1}{2}\int_{\frac{1}{\sigma_{\max,d}}}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}dx \tag{4.32}$$

where $q\in[1,(T-1)]$, calculated using the MVA approach discussed in section 4.2. The discrete triangle distribution applied is a symmetric distribution, which represents the probability that the time slot position of two sequentially arriving frames will be offset by a jitter amount of +*j* or –*j*. Thus, the triangle distribution for the range [-*q*, *q*] is applied as

$$P\left\{\widetilde{J}_n=j\mid A_{(n+1),d}-1=q,\right\}=f_{q,d}(j)$$

$$= \begin{cases} \dfrac{1}{q+1} - \dfrac{j}{(A_{\min}+1)^2}, & \mathit{for}\ |j| \le q \\ 0, & \mathit{otherwise}, \end{cases} \qquad (4.33)$$

which provides the normalized jitter probability of j given that q packet arrivals enter the server before the (n+1)th tagged packet of tagged stream, among the overall $A_{\min}$ possible time slots.

4.4.1 Simulation and Numerical Study

The jitter probability of heterogeneous traffic streams with HOL priority control in DiffServ networks has been investigated. The derivations of this section provide a direct method to analyze the per-class jitter based on the parameters ρ and T. Fig. 4.5 and 14 show the jitter probability distribution of three different classes for the case of $T = 10$ slots and ρ equal to 50%, and also assuming that the probability of traffic arrival in the T slot observation window for class 1, class 2, and class 3 are the same. Fig. 4.5 has shown the 95% CI for the delay jitter proportion between the real data set and the numerical analysis. The width of the 95% CI reveals the degree of uncertainty in the estimate of the treatment effect; in another words, this is the interval which includes the true value with 95% certainty. The wider the width the more uncertainty it is, and more data set that collected from the same population is needed to increase the confidence level. In Fig. 4.6, the approximation equation results are comparing to the simulation results of the jitter distribution. In Fig. 4.5 and 4.6, it can be observed that the probability distribution function of the interarrival packet jitter widens as the class priority is descending, thus resulting in a higher probability of packet jitter for low priority classes. This is an expected result, where the newly derived equations provide a method to directly compute this physical probabilistic quantity. Also, for class 1 as $\rho << 1$ the probability of jitter at zero (no jitter) increases while the probability of jitter for the non-zero values is reduced.

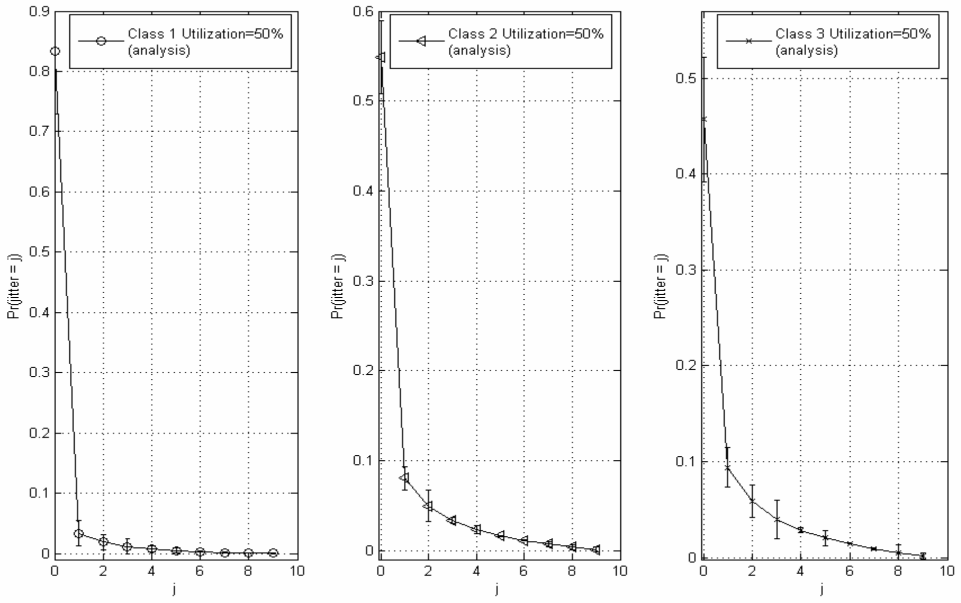


Fig. 4.5 Comparison among class 1, class 2 and class 3 with respect to the approximation equation and the simulation results of the jitter distribution for utilization of 50% with 95% CI, and $T = 10$.

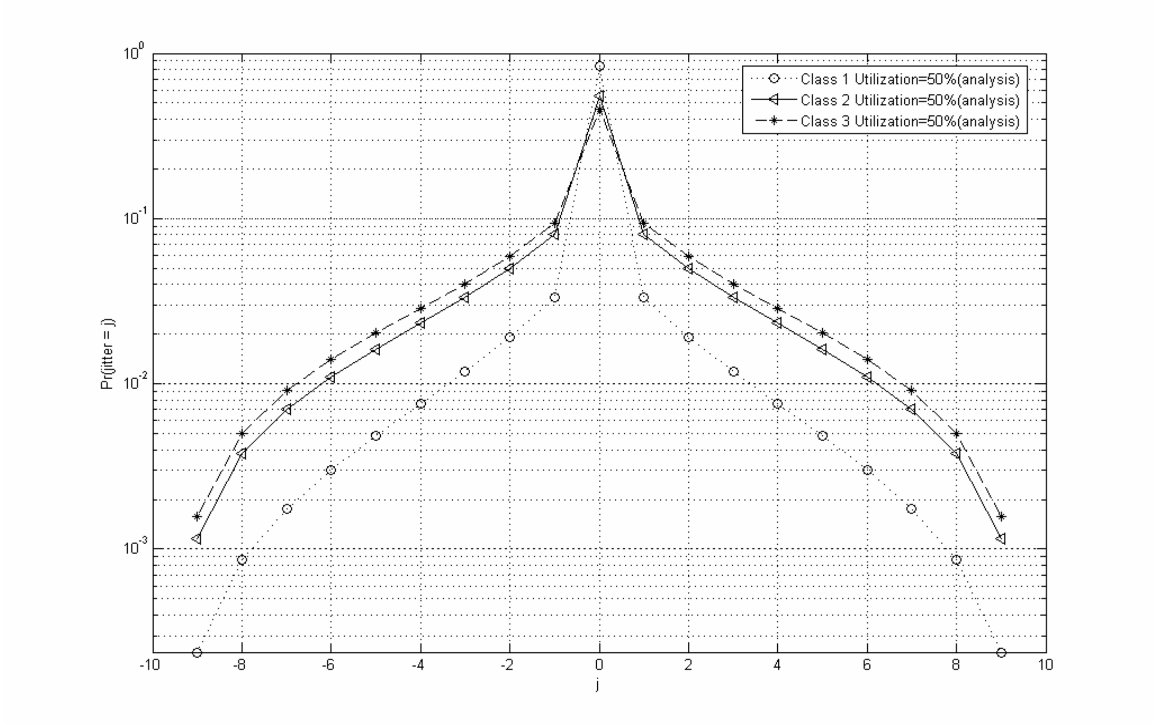


Fig. 4.6 The comparison of probability jitter distribution among class 1, class 2, and class3, where utilization is 50% and *T* is 10.

Chapter 5

Conclusions

In this dissertation, the following research has been conducted

1. Investigate the delay jitter performance of homogenous wireless networks that apply ARQ error recovery with time constraints have been developed. The effects on the delay jitter in reference to the priority control scheme of the ARQ traffic for the two cases are evaluated: i) the ARQ traffic has a priority over the original transmission traffic; and ii) the ARQ traffic has no priority over the original transmission traffic.
2. Investigate the issues of traffic jitter characteristics in heterogeneous wired communication networks deploying different scheduling algorithms:
   - Obtain the tail probability $P\{Q>q\}$ based on the MVA approach for discrete time queueing model [25][26] to conduct the queue length analysis, which is used in heterogeneous jitter analysis.
   - To find the pmf (probability mass function) of inter-arrival packet jitter from the queue length distribution.
3. The objective of this dissertation is to investigate and analyze the various possible traffic modeling techniques and evaluate the challenges in characterizing the diverse statistical properties of heterogeneous wireless networks [13-43][56-60]:

- Study the characteristics of the traffic: self-similarity, heavy-tailed distribution, Gaussian traffic distribution.
- Comparisons among the popular traffic models.

4. Apply the Gaussian traffic modeling using the MVA approach [25][26] to conduct the queue length analysis, which will be further used in heterogeneous jitter analysis [1-12].
5. Analyze the difference between jitter probability of multiple priority queues and switches [1-12]:
   - The head-of-line (HOL) priority queueing mechanism is applied at the queueing and scheduling control.
6. Develop a service discipline called the tagged stream adaptive distortion-reducing peak output-rate enforcing to control and avoid the delay jitter increases without bound.

In conclusion, using the Gaussian traffic modeling technique combined with the MVA approach for self-similar network traffic, the delay jitter was shown with the 95% CI for the delay jitter proportion between the real data set and the numerical analysis.

For future research, this analysis will be extended to multiple priority queueing case. The multi-hop jitter of wireless and wired network analysis can be conducted, where end-to-end congestion control is needed at a router for fair bandwidth allocation per-flow. The Core-Stateless Fair Queueing (CSFQ) [47] can be applied to allocate fair bandwidth per-flow, and performance is evaluated (the number of congested links).

References

1. C. C. Bisdikian, W. Matragi, and K. Sohraby, “A Study of the Jitter in ATM Multiplexers,” *High Speed Networks and their performance: IFIP Trans. C, Commun. Systems, C-21*, pp. 219-235, New York, 1994.
2. T. C. Wong, J. W. Mark, K. C. Chua, and B. Kannan, “Delay jitter performance of voice traffic in a cellular wireless ATM network,” *Proc. IEEE 55th Veh. Technol. Conf.-Spring 2002*, vol. 1, pp. 90-94, May 2002.
3. D. Zhao, X. Shen, and J. W. Mark, “QoS performance bounds and efficient connection admission control for heterogeneous services in wireless cellular networks,” *Wireless Networks*, vol. 8, pp.85-95, 2002.
4. C.-S. Chang, K.-C. Chen, M.-Y. You, and J.-F. Chang, “Guaranteed quality-of-service wireless access to ATM networks,” *IEEE J. Selected Areas in Commun.*, vol. 15, no. 1, Jan. 1997.
5. D. Zhao, X. Shen, and J. W. Mark, “Efficient call admission control for heterogenous services in wireless mobile ATM networks,” *IEEE Commun. Mag.*, vol. 38, pp. 72-78, Oct. 2000.
6. W. K. Wong and V. C. M. Leung, “Scheduling for integrated services in next generation packet broadcast networks,” *Proc. IEEE Conf. Wireless Commun. and Netw.*, vol. 3, pp. 1278-1282, Sept. 1999.

7. S. Marwaha, C. K. Tham, and D. Srinivasan, “A novel handover mechanism for wireless mobile ad hoc networks,” *Proc. ICCS* 2002, vol. 2, pp. 1063-1067, Nov. 2002.

8. U. Lambrette, L. Bruhl, and H. Meyr, “ARQ protocol performance for a wireless high data rate link,” *Proc. IEEE 47th Veh. Technol. Conf.-Spring 1997*, vol. 3, pp. 1538-1542, May 1997.

9. J.-M. Chung and H. M. Soo, “Jitter Analysis of Homogeneous Traffic in Differentiated Services Networks,” *IEEE Commun. Lett.*, vol. 7, no. 3, pp. 130-132, Mar. 2003.

10. A. Privalov, and K. Sohraby, “Per-Stream Jitter Analysis in CBR ATM Multiplexors,” *IEEE/ACM Trans. on Netw.*, vol. 6, no. 2, Apr. 1998.

11. J.-M. Chung, H. M. Soo and W.-C. Jeong, “Jitter Analysis of Homogeneous Traffic in Wireless Differentiated Services Networks,” *Proc. IEEE LANMAN 200*4, CA, USA, Apr. 25-28, 2004.

12. J. J. Metzner and J.-M. Chung, “Efficient Energy Utilization with Time Constraint in Mobile Time Varying Channels,” *IEEE Trans. Veh. Techno*l, vol. 49, no. 4, pp. 1169-1177, July 2000.

13. R. G. Addie, “On the applicability and utility of Gaussian models for broadband traffic,” *Proc. IEEE Intl. Conf. on ATM*, 1998.

14. P. Mannersalo and I. Norros, “GPS schedulers and Gaussian traffic,” *Proc. IEEE INFOCOM*, 2002.

15. S. Sarvotham. R. Riedi, and R. Baraniuk, “Connection-level Analysis and Modeling of Network Traffic,” *IMW’01*, CA, Nov. 2001, pp. 99-103.

16. S. Ma and C. Ji, "Modeling Heterogeneous Netwrok Traffic in Wavelet Domain," *IEEE/ACM Trans. Networking*, vol. 9, no. 5, Oct. 2001, pp. 634-649.

17. H. Fei and W. Zhimei, "A Novel Traffic Based on Wavelet analysis," *Proc. ICCT 2003*, vol. 2, April 2003, pp. 1392-1395.

18. Petteri Mannersalo, Gaussain and multifractal process in teletraffic theory, 2003.

19. R. G. Addie, "Traffic will be more Gaussian in Future," Australian Telecom. Networks and Applications Conference. December, Melbourne, 1996.

20. R. G. Addie, On weak convergence of long-range dependent traffic processes, Technical Report SC-MC-9816, University of Soutern Queensland, 1998.

21. Damon Wischik, "*Implications of long-range dependence*," notes on the implications of long-range dependence for optical networks, Feb. 2001.

22. R. G. Addie, "On the weak convergence of long range dependent traffic process," *Proc. of the International Workshop on Long Range Dependent*, Jan. 1997.

23. P. Mannersalo and I. Norros, "GPS schedulers and Gaussian traffic," in *Proc. IEEE INFOCOM*, 2002.

24. R. G. Addie, "On the applicability and utility of Gaussian models for broadband traffic," *Proc. IEEE Intl. Conf. on ATM*, 1998.

25. J. Choe and N. B. Shroff, "A central-limit-theorem-based approach for analyzing queue behavior in high-speed networks," *IEEE/ACM Trans. On Networking*, *vol. 6*, *no. 5*, Oct. 1998, pp 659-671.

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

27. R. G. Addie, M. Zukerman, and T. D. Neame. “Broadband traffic modeling: simple solutions to hard problems,” *IEEE Communications Magazine*, August. 1998, pp 88-95.

28. J. Abate, G. L. Choudhury, and W. Whitt, “Exponential approximations for tail probabilities in queues-I: Waiting times,” *Oper. Res.*, vol. 43, no. 5, pp. 885-901, 1995.

29. R. G. Addie and M. Zukerman, “An approximation for performance evaluation of stationary single server queues,” *IEEE Trans. Commun.*, vol. 42, pp. 3150-3160.

30. R. Guerin, H. Ahmad, and M. Naghshineh, “Equivalent capacity and its application to bandwidth allocation in high-speed networks,” *IEEE J. Select. Areas Commun.*, vol. 9, pp. 968-981, Sept. 1991.

31. A. Simonian, “Stationary analysis of a fluid queue with input rate varying as an Ornstein-Uhlenbeck process,” *SIAM J. Appl. Math.*, vol. 51, pp. 828-842. 1991.

32. R. G. Addie, M. Zukerman, and T. Neame, “*Performance of a single server queue with self similar input*,” *IEEE* 1995.

33. R. W. Wolff, *Stochastic Modeling and the Theory of Queues*, Prentice Hall, New Jersey, 1989.

34. Will E. Leland, Murad S. Taqqu, Walter Willinger, and Daniel V. Wilson, “On the self-similar nature of Ethernet traffic (extended version),” *IEEE/ACM Trans. Net.*, vol. 2, no.1, Feb. 1994, pp. 1-15.

35. F. H. P. Fitzek and M. Reisslein, “MPEG-4 and H.263 video traces for networks performance evaluation,” *IEEE Net.*, Nov. 2001, pp. 40-54.

36. Murad S. Taqqu, http://math.bu.edu/people/murad/methods/index.html

37. Arifler and B. L. Evans, “Modeling the self-similar behavior of packetized MPEG-4 video using wavelet-based methods,” *Proc. Image Processing 2002*, 2002, vol. 1, pp. 848-851.

38. Thomas Karagiannis and Michalis Faloutsos, “SELFIS: A Tool For Self-Similarity and Long-Range Dependence Analysis,” 1*st Workshop on Fractals and Self-Similarity in Data Mining*: *Issues and Approaches* (*in KDD*) *Edmonton*, Canada, July 23, 2002.

39. Thomas Karagiannis , http://www.cs.ucr.edu/~tkarag/Selfis/Selfis.html

40. D.R.Avresky, V. Shurbanov, R. Horst and P. Mehra, “Performance Evaluation of ServerNet R SAN under Self-Similar Traffic,” 13*th International Parallel Processing Symposium and 10th Symposium on Parallel and Distributed Proc.*, Apr. 12 - 16, 1999, San Juan, Puerto Rico, pp. 143-147.

41. The trace/data set, Star Wars IVat,
http://www-tkn.ee.tu-berlin.de/research/trace/trace.html

42. The trace/data set, Ethernet traffic,
http://math.bu.edu/people/murad/methods/index.html

43. R. C. Garcia and J.-M. Chung, “Network Anomaly Detection Using A Self-Similar Traffic Model Report”, NSF DMI-0339471, Jan. 31, 2004.

44. W. Stallings, *High-Speed Networks TCP/IP and ATM Design Principles*, Prentice Hall, New Jersey, 1998.

45. R. M. Loynes, “The Stability of a Queue with Nonindependent Inter-arrival and Service Times,” Proc. Cambridge Philos Soc., vol. 58, pp. 497-520, 1962.

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

47. I. Stoica, S. Shenker, and H. Zhang, "Core-Stateless Fair Queueing: A Scalable Architecture to Approximate Fair Bandwidth Allocations in High-Speed Networks," *IEEE/ACM Trans. Net.*, vol. 11, no.1, Feb. 2003, pp. 33-46.

48. P. Billingsley, *Probability and Measure*, New York: John Wiley & Sons, 1979.

49. E. Wong and B. Jajek, Stochastic *Processes in Engineering Systems*, New York: Springer-Verlag, 1985.

50. R. Landry and Ionnis Stavrakakis, "Study of Delay Jitter With and Without Peak Rate Enforcement," *IEEE/ACM Trans. On Commun.*, vol. 5, no. 4, Aug. 1997, pp. 543-553.

51. C.-N. Chuah, R. H. Katz, Network provisioning and resource management for IP telephony, University of California, Berkeley, Report no. UCB/CSD-99-1061, Sept. 1, 1999.

52. L. Zhang, L. Zheng, and K. S. Nge, "Effect of delay and delay jitter on voice/video over IP," *Elsevier Computer Commun.*, vol. 25, 2002, pp. 863-873.

53. X. Xiao and L. M. Ni, "Internet QoS: A Big Picture," *IEEE Network*, March/April 1999, pp. 8-18.

54. S. Blake, D. Black, M. Carlson, E. Davies Z. Wang, and W. Weiss, "An Architecture for Differentiated Services," RFC 2475, Dec. 1998.

55. T. Li, and Y. Rekhter, "A Provider Architecture for Differentiated Services and Traffic Engineering (PASTE)", RFC 2430, Oct. 1998.

56. Jinwoo Choe, Ness B. Shroff, "Use of the Supremum Distribution of Gaussian Processes in Queueing Analysis with Long-Range Dependence and Self-Similarity," *Communications in Statistics - Stochastic Models*, vol. 16, no. 2, pp. 209~231, 2000.

57. Piterbarg, V. I., Asymptotic Methods in the Theory of Gaussian Processes and Fields, American Mathematical Society, Providence, RI.

58. J. Choe and N. B. Shroff, "Queueing analysis of high-speed multiplexers including long-range dependent arrival processes," *Proc. INFOCOM* '99, *vol. 2*, *Mar*. 1999, pp 617-624.

59. J. Choe and N. B. Shroff, "Use of supremum distribution of Gaussian processes in queueing analysis with long-range dependence and self-similarity," *Stochast. Models*, *vol. 16*, *no. 2*, Feb. 2000.

60. J. Choe and N. B. Shroff, "On the supremum distribution. of integrated stationary Gaussian processes with negative linear drift," *Advances in Applied Probability*, *vol. 31*, *no. 4*, 1999,pp. 135–157.

61. U. Lambrette, L. Bruhl, and H. Meyr, "ARQ protocol performance for a wireless high data rate link," *Proc. IEEE 47th Veh. Technol. Conf.-Spring 1997*, vol. 3, pp. 1538-1542, May 1997.

VITA

Hooi Miin Soo

Candidate for the Degree of

Doctor of Philosophy

Thesis: INVESTIGATION OF DELAY JITTER OF HETEROGENEOUS TRAFFIC IN WIRED NETWORKS

Major Field: Electrical Engineering

Biographical:

Education: Received Bachelor of Science degree in Electrical Engineering from Oklahoma State University, in May 2000, and the Master of Science degree with a major in Electrical Engineering at Oklahoma State University in May, 2003.

Experience: Research Assistant at Oklahoma State University, Stillwater, OK, from Aug. 2000 to present.
Teaching Assistant at Oklahoma State University, Stillwater, OK, ECEN 3913 Solid State Electronics Device, from Jan. 2001 to May 2001.
Teaching Assistant at Oklahoma State University, Stillwater, OK, ECEN 4503 Random Signal & Noise, from Jun. 2001 to Aug. 2001.

Professional Memberships: Student Member of IEEE, Eta Kappa Nu, Golden Key National Honor Society, and Tau Beta Pi National Honor Society.

Name: Hooi Miin Soo Date of Degree: Dec, 2006

Institution: Oklahoma State University Location: Stillwater, Oklahoma

Title of Study: INVESTIGATION OF DELAY JITTER OF HETEROGENEOUS TRAFFIC IN BROADBAND NETWORKS

Pages in Study: 89 Candidate for the Degree of Doctor of Philosophy

Major Field: Electrical Engineering

Scope and Methodology of Study: A critical challenge for both wired and wireless networking vendors and carrier companies is to be able to accurately estimate the quality of service (QoS) that will be provided based on the network architecture, router/switch topology, and protocol applied.. As a result, this thesis focuses on the theoretical analysis of QoS parameters in term of inter-arrival jitter in differentiated services networks by deploying analytic/mathematical modeling technique and queueing theory, where the analytic model is expressed in terms of a set of equations that can be solved to yield the desired delay jitter parameter. In wireless networks with homogeneous traffic, the effects on the delay jitter in reference to the priority control scheme of the ARQ traffic for the two cases of: 1) the ARQ traffic has a priority over the original transmission traffic; and 2) the ARQ traffic has no priority over the original transmission traffic are evaluated. In wired broadband networks with heterogeneous traffic, the jitter analysis is conducted and the algorithm to control its effect is also developed.

Findings and Conclusions: First, the results show that high priority packets always maintain the minimum inter-arrival jitter, which will not be affected even in heavy load situation. Second, the Gaussian traffic modeling is applied using the MVA approach to conduct the queue length analysis, and then the jitter analysis in heterogeneous broadband networks is investigated. While for wireless networks with homogeneous traffic, binomial distribution is used to conduct the queue length analysis, which is sufficient and relatively easy compared to heterogeneous traffic. Third, develop a service discipline called the tagged stream adaptive distortion-reducing peak output-rate enforcing to control and avoid the delay jitter increases without bound in heterogeneous broadband networks. Finally, through the analysis provided, the differential services, was proved not only viable, but also effective to control delay jitter. The analytic models that serve as guidelines to assist network system designers in controlling the QoS requested by customer in term of delay jitter.

ADVISOR'S APPROVAL: Dr. Jong-Moon Chung