

INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each original is also photographed in one exposure and is included in reduced form at the back of the book.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

UMI

A Bell & Howell Information Company
300 North Zeeb Road, Ann Arbor MI 48106-1346 USA
313/761-4700 800/521-0600

RICE UNIVERSITY

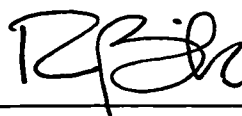
**Wavelet-Based Queuing Analysis of Gaussian and
NonGaussian Long-Range-Dependent Network
Traffic**

by

Vinay Ribeiro

A THESIS SUBMITTED
IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE
Master of Science

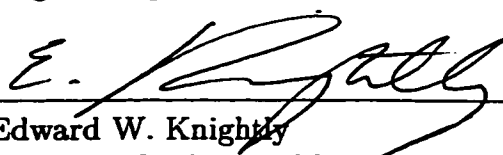
APPROVED, THESIS COMMITTEE:



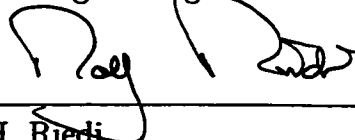
Richard G. Baraniuk, Chair
Associate Professor of Electrical and
Computer Engineering



Don H. Johnson
Professor of Electrical and Computer
Engineering



Edward W. Knightly
Assistant Professor of Electrical and
Computer Engineering



Rudolf H. Riedi
Research Associate, Department of
Electrical and Computer Engineering

Houston, Texas

May, 1999

UMI Number: 1394259

UMI Microform 1394259
Copyright 1999, by UMI Company. All rights reserved.

**This microform edition is protected against unauthorized
copying under Title 17, United States Code.**

UMI
300 North Zeeb Road
Ann Arbor, MI 48103

Wavelet-Based Queuing Analysis of Gaussian and NonGaussian Long-Range-Dependent Network Traffic

Vinay Ribeiro

Abstract

In this thesis, we develop a simple and powerful multiscale model for the synthesis of nonGaussian, long-range-dependent (LRD) network traffic. The wavelet transform effectively decorrelates LRD signals and hence is well-suited to model such data. However, wavelet-based models have generally been used for modeling Gaussian data which can be unrealistic for traffic. Using a multiplicative superstructure atop the Haar wavelet transform, we exploit the decorrelating properties of wavelets while simultaneously capturing the positivity and “spikiness” of nonGaussian traffic. We develop a queuing analysis for our model by exploiting its multiscale construction scheme. We elucidate our model’s ability to capture the covariance structure of real data and then fit it to real traffic traces. Queuing experiments demonstrate the accuracy of the model for matching real data and the precision of our theoretical queuing result. Our results indicate that a Gaussian assumption can lead to over-optimistic predictions of tail queue probability.

Acknowledgments

I have many people to thank for making this thesis possible. I am greatly indebted to Matthew Crouse whose work this thesis is based on and Rolf Riedi for both his advice and the long hours he spent discussing various ideas with me. Many thanks to my dynamic advisor, Dr. Richard Baraniuk, for his encouragement, patience and lively character that make doing research with him a pleasure. I must also thank Dr. Edward Knightly, Prof. Don Johnson and Dr. Dennis Cox for the enlightening discussions I had with them.

My list would be incomplete if I left out all those who have made it possible for me to study at a prestigious research institution like Rice University. I thank God for all His blessings, my parents for their love and support and my teachers for the education they gave me.

I also must thank my cooking group members Dins, Amit, Rahul, Pots, Neelsh, Chis and Felix for their company and of course excellent food. I also thank Dinesh and Ravi for being excellent room-mates and Justin and Tao for being office-mates par excellence. My bible study companions were a source of strength and guidance and I especially thank C. J. Fretheim, Nathan Tallent, Justin Lokey and Damian Dobric.

Last but not least, I must thank Texas Instruments, the National Science Foundation and DARPA for their financial support that made this thesis possible.

Contents

Abstract	ii
Acknowledgments	iii
List of Illustrations	vi
List of Tables	vii
1 Introduction	1
2 Wavelet Models for LRD Processes	6
2.1 Long-range dependence	6
2.2 Wavelet transform	7
2.3 Wavelet-domain Independent Gaussian (WIG) model	10
3 Multifractal Wavelet Model	12
4 Queuing Analysis of Wavelet-Based Models	16
4.1 Queue size and multiple time scales	17
4.2 Queuing analysis	19
4.2.1 Queuing formula for tree-based models	19
4.2.2 Queuing analysis of the WIG	21
4.2.3 Queuing analysis of the MWM	21
5 Experimental Results	23

5.1	Real data	23
5.2	Matching of Real Data	24
5.3	Queuing experiments	25
6	The MWM is a Cascade	30
6.1	Cascades	30
6.2	Multifractal analysis	31
6.3	Multifractal spectrum and higher-order moments	31
7	Conclusions	34
A	Proof of Lemma	35
	Bibliography	37

Illustrations

1.1	Real WAN traffic and synthetic MWM and WIG data at different aggregation levels	5
2.1	The Haar wavelet system and the Wavelet-domain Independent Gaussian (WIG) model construction	9
3.1	Multifractal Wavelet Model (MWM) construction	14
5.1	Histograms of real WAN data and synthetic MWM and WIG data at different aggregation levels	27
5.2	Comparison of variance-time plots of real data and synthetic WIG and MWM data	28
5.3	Comparison of queuing performance of real traces and synthetic WIG and MWM traces	28
5.4	Validation of theoretical formula for tail queue probability of the MWM and WIG models	29
6.1	Cascades and the Multifractal Spectrum	33

Tables

3.1	Comparison of the tree-based WIG and MWM models	15
-----	---	----

Chapter 1

Introduction

Traffic models play a significant rôle in the analysis and characterization of network traffic and network performance. Accurate models capture important characteristics of traffic and enhance our understanding of these complicated signals and systems by allowing us to study the effect of various model parameters on network performance through both analysis and simulation.

One key property of modern network traffic is the presence of *long-range dependence* (LRD) which was demonstrated convincingly in the landmark paper of Leland et. al. [24]. There, measurements of traffic load on an Ethernet were attributed to *fractal* behavior or *self-similarity*, i.e., to the fact that the data “looked statistically similar” (“bursty”) on all time-scales. These features are inadequately described by classical traffic models such as Markov or Poisson models. In particular, the LRD of data traffic can lead to higher packet losses than that predicted by classical queuing analysis [24, 12].

These findings were immediately followed by the development of new fractal traffic models [39, 25, 4]. *Fractional Brownian motion* (fBm), the most broadly applied fractal model, is the unique Gaussian process with stationary increments and the following scaling property: for all $a > 0$

$$B(at) \stackrel{fd}{=} a^H B(t), \quad (1.1)$$

with equality in the sense of finite-dimensional distributions. The discrete increment process $G(k) := B((k+1)\Delta) - B(k\Delta)$, called *fractional Gaussian noise* (fGn), has

an autocorrelation of the form

$$r_G[k] = \frac{\sigma^2}{2} |\Delta|^{2H} (|k+1|^{2H} - 2|k|^{2H} + |k-1|^{2H}), \quad (1.2)$$

where Δ is a constant and the parameter H , $0 < H < 1$, is known as the *Hurst parameter*. Gaussianity and the strong scaling (1.1) make rigorous analytical studies of queueing behavior possible [6, 29, 11, 30, 15], thus increasing the popularity of the fBm/fGn models.

Processes approximating fBm/fGn can be synthesized almost effortlessly using the amazing decorrelating capability of the *wavelet transform* [14, 44, 41, 22]. We simply generate independent Gaussian wavelet coefficients with variances decaying with scale as a power law and then invert the wavelet transform. Generalizations of fBm/fGn with a more flexible scaling relation than (1.1) are easily generated as well. Such models can approximate both the long and short-term correlations of a target data set and have been used by a number of authors [27, 23]. We will term all such models *wavelet-domain independent Gaussian* (WIG) models. The WIG synthesizes N -point data sets in a fast $O(N)$ algorithm.

Though traffic models based on fBm/fGn are appropriate in some cases [8, 43], they have severe limitations. First, real-world traffic traces do not exhibit the strict self-similarity of (1.1) or (1.2) and are at best merely asymptotically self-similar. In other words, the single parameter H does not sufficiently capture the complicated correlation structure of real network processes. Indeed, convincing evidence has been produced establishing the importance of short-term correlations for buffering [10, 38, 17] and so-called relevant time scales have been discovered [38, 28, 16]. This shortcoming is surmounted by more versatile models such as the WIG [27] and fractional ARIMA [40].

Second, the Gaussianity of fBm/fGn/WIG models can be unrealistic for certain

types of traffic, for instance when the standard deviation of the data approaches or exceeds the mean. In this case, considerable parts of the fBm/fGn/WIG output are negative (see Figure 1(a) and (b)).

Third, in many networking applications, we are nowhere near the Gaussian limit, in particular on small time scales. Indeed, various authors have observed heavy-tailed marginals in traffic [37, p. 364],[3].

In this thesis, we propose a new model for network traffic, the *multifractal wavelet model* (MWM), based on a multiplicative cascade in the wavelet domain that by design guarantees a positive output. Since each sample of the MWM process is obtained as a product of several positive independent random variables, the MWM's marginal density is approximately lognormal, a distribution with heavier tails than the Gaussian. The MWM is thus a more natural fit for positive arrival processes. This is especially true when the standard deviation is much larger than the mean (as observed in the traces we have studied). Fitting the MWM to real traffic traces results in an excellent match, far better than the WIG model, visually (see Figure 1(c)) and, as we will see, in the burstiness as measured by the multifractal spectrum, in the marginals and the queueing behavior. It thus appears that the multiplicative MWM approach is more appropriate than an additive Gaussian one.

In its simplest form, the MWM is closely related to the wavelet-based construction of fBm/fGn, having as few parameters (mean, variance, H). However, the MWM framework boasts the flexibility, if desired, to additionally match the short-term correlations like the WIG model.

Apart from matching crucial properties of real network traffic, a good traffic model must possess an accurate and simple queueing analysis. By restricting our analysis to data at time scales of powers of two, we exploit the inherent binary tree structure of the MWM in deriving an easy-to-use — and as numerous experiments verify —

a close approximation to the tail queue probability. As a consequence, the MWM becomes viable for applications like call admission control. Our analysis is applicable to tree-based models in general including the WIG.

After introducing wavelets and explaining the WIG model in Chapter 2, we describe the MWM in Chapter 3. We then perform a queuing analysis of the WIG and MWM in Chapter 4. Chapter 5 provides empirical evidence for the accuracy of the MWM in modeling real data as well as for the accuracy of our queuing analysis. We give an intuitive introduction to multifractal cascades in Chapter 6 and close with conclusions in Chapter 7. The proof of a lemma instrumental in our analysis appears in the Appendix.

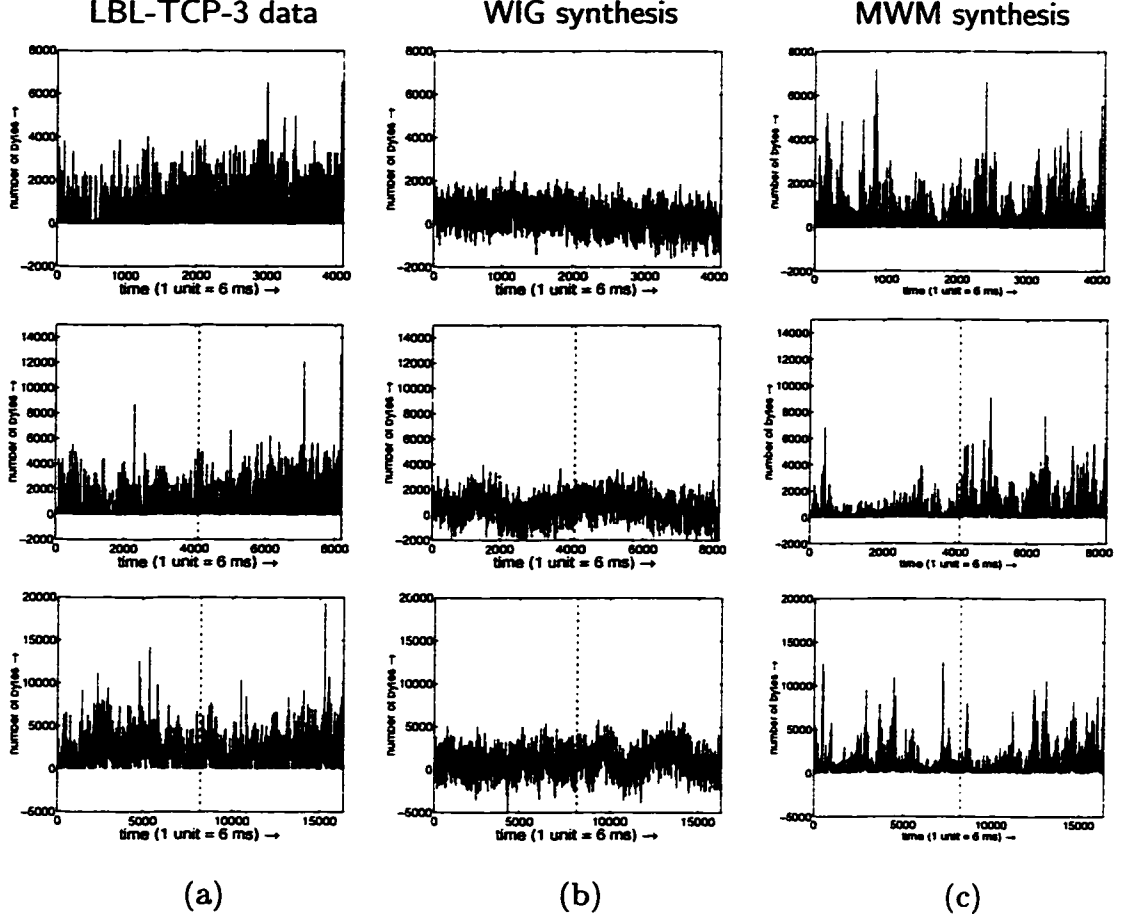


Figure 1.1 : Bytes-per-time arrival process at different aggregation levels for (a) wide-area TCP traffic at the Lawrence Berkeley Laboratory (trace LBL-TCP-3) [33], (b) one realization of the state-of-the-art wavelet-domain independent Gaussian (WIG) model [27], and (c) one realization of the multifractal wavelet model (MWM) synthesis. The top, middle and bottom plots correspond to bytes arriving in intervals of 6ms, 12ms and 24ms respectively. The top and middle plots correspond to the second half of the middle and bottom plots, respectively, as indicated by the vertical dotted lines. The MWM traces closely resemble the real data closely, while the WIG traces (with their large number of negative values) do not.

Chapter 2

Wavelet Models for LRD Processes

2.1 Long-range dependence

The discovery of LRD in data traffic [24, 33] has incited a revolution in network design, control and modeling. Intuitively, the strong correlations present in a LRD process are responsible for its “bursty” nature that causes excessive buffer overflows not predicted by traditional non-LRD traffic models such as Poisson and Markov models [12].

Consider a discrete-time, wide-sense stationary random process $\{X_t, t \in \mathbb{Z}\}$ with auto-covariance function $r_X[k] = \text{cov}(X_t, X_{t+k})$. A change in time scale can be represented by forming the aggregate process $X_t^{(m)}$, which is obtained by averaging X_t over non-overlapping blocks of length m and replacing each block by its mean

$$X_t^{(m)} = \frac{X_{tm-m+1} + \cdots + X_{tm}}{m}. \quad (2.1)$$

Denote the auto-covariance of $X_t^{(m)}$ by $r_X^{(m)}[k]$. The process X is said to exhibit LRD if its auto-covariance decays slowly enough to render $\sum_{k=-\infty}^{\infty} r_X[k]$ infinite [7]. Equivalently, $m r_X^{(m)}[0] \rightarrow \infty$ as $m \rightarrow \infty$ and the power spectrum $S_X(f)$ is singular near $f = 0$.

An important class of LRD processes are the *asymptotically second-order self-similar processes*, which are defined by the property $r_X[k] \sim k^{2H-2}$ for some $H \in (1/2, 1)$, or equivalently (see [7])

$$r_X^{(m)}[k] \rightarrow (1/2)r_X^{(m)}[0](|k+1|^{2H} - 2|k|^{2H} + |k-1|^{2H}) \quad (2.2)$$

as $m \rightarrow \infty$. In words, these processes “look similar” on all scales, at least from the point-of-view of second-order statistics. An example of such a process is the fGn where the *Hurst parameter* H in (1.1) is exactly the scaling parameter (2.2).

To estimate H by the *variance-time plot* method, we fit a straight line through the plot of an estimate of $\log \text{var}(X^{(m)})$ against $\log(m)$. More reliable estimators of H have been devised [40], in particular an unbiased one based on wavelets [1].

2.2 Wavelet transform

The discrete wavelet transform provides a multiscale signal representation of a one-dimensional signal $c(t)$ in terms of shifted and dilated versions of a prototype bandpass wavelet function $\psi(t)$ and shifted versions of a lowpass scaling function $\phi(t)$ [5, 9]. For special choices of the wavelet and scaling functions, the atoms

$$\psi_{j,k}(t) := 2^{j/2} \psi(2^j t - k), \quad \phi_{j,k}(t) := 2^{j/2} \phi(2^j t - k), \quad j, k \in \mathbb{Z} \quad (2.3)$$

form an orthonormal basis, and we have the signal representation [9]

$$c(t) = \sum_k u_{J_0,k} \phi_{J_0,k}(t) + \sum_{j=J_0}^{\infty} \sum_k w_{j,k} \psi_{j,k}(t). \quad (2.4)$$

Here the wavelet coefficients $w_{j,k}$ and the scaling coefficient $u_{J_0,k}$ are given by

$$w_{j,k} := \int c(t) \psi_{j,k}(t) dt, \text{ and } u_{J_0,k} := \int c(t) \phi_{J_0,k}(t) dt. \quad (2.5)$$

Without loss of generality, we will assume $J_0 = 0$.

In this representation, k indexes the spatial location of analysis and j indexes the *scale* or resolution of the wavelet analysis — larger j corresponds to higher resolution and $j = 0$ indicates the coarsest scale or lowest resolution of analysis. In practice, we work with a sampled or finite-resolution representation of $c(t)$, replacing the semi-infinite sum in (2.4) with a sum over a finite number of scales $0 \leq j \leq n-1$, $n \in \mathbb{Z}_+$.

Using filter-bank or pyramid algorithm techniques, the forward and inverse wavelet transforms of an N -point signal can be computed in $O(N)$ operations.

In this paper, we restrict our attention to the simplest wavelet system, the *Haar*. The Haar scaling and wavelet functions are given by (see Figure 2.1(a) for $\phi_{j,k}(t)$ and $\psi_{j,k}(t)$)

$$\phi(t) = \begin{cases} 1, & 0 \leq t < 1 \\ 0, & \text{else} \end{cases} \quad \text{and} \quad \psi(t) = \begin{cases} 1, & 0 \leq t < 1/2 \\ -1, & 1/2 \leq t < 1 \\ 0, & \text{else.} \end{cases} \quad (2.6)$$

The Haar scaling coefficients $u_{j,k}$ are obtained by integrating the product of the signal $c(t)$ and the rectangular scaling function $\phi_{j,k}(t)$ as in (2.5). They represent the *local mean values* of the signal in the time intervals $[k2^{-j}, (k+1)2^{-j}]$. The continuous signal $c(t)$ can thus be approximated at resolution j by the discrete signal $u_{j,k}$, $k \in \mathbb{Z}$. The support of $\phi_{j,k}(t)$ for different values of j and k are the intervals $[k2^{-j}, (k+1)2^{-j}]$. By design they are nested within each other and the relationship between coefficients $u_{j,k}$ is well-captured by a binary tree (Figure 2.1(b)). Nodes at horizontal levels in the tree correspond to representations of the signal at different approximations with finer resolutions situated lower in tree.

The wavelet coefficients $w_{j,k}$ represent the *difference* between local means in the time intervals $[k2^{-j}, (k+1/2)2^{-j}]$ and $[(k+1/2)2^{-j}, (k+1)2^{-j}]$. They provide the detail information required to move from level j on the scaling coefficient tree to the next finer level $j+1$. Thus, the Haar wavelet transform of a signal can be computed recursively starting from its finest-scale scaling coefficients via [9]

$$u_{j-1,k} = 2^{-1/2}(u_{j,2k} + u_{j,2k+1}), \quad w_{j-1,k} = 2^{-1/2}(u_{j,2k} - u_{j,2k+1}). \quad (2.7)$$

This corresponds to moving up the binary tree and storing the detail information $w_{j,k}$ lost while going from a finer to coarser resolution of the data (Figure 2.1(b)).

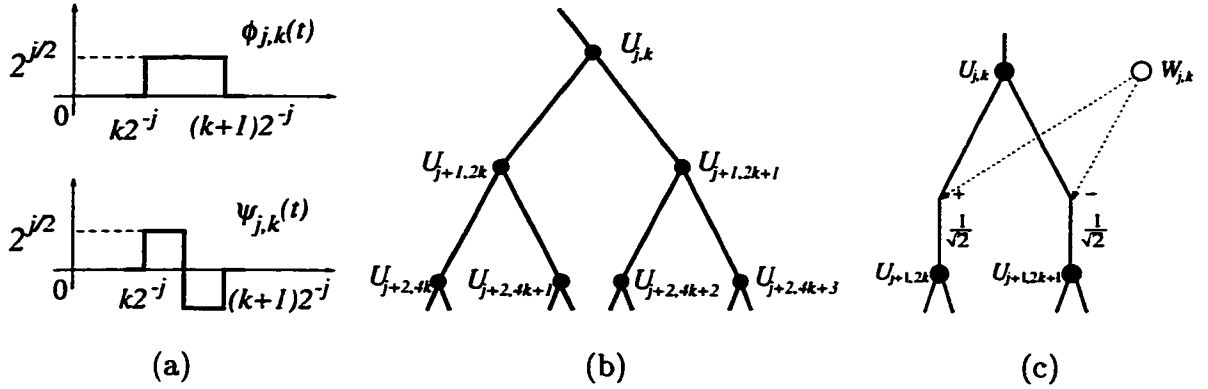


Figure 2.1 : (a) The Haar scaling and wavelet functions $\phi_{j,k}(t)$ and $\psi_{j,k}(t)$. (b) Binary tree of scaling coefficients from coarse to fine scales. (c) Recursive scheme for calculating the Haar scaling coefficients $U_{j+1,2k}$ and $U_{j+1,2k+1}$ at scale $j+1$ as sums and differences (normalized by $1/\sqrt{2}$) of the scaling and wavelet coefficients $U_{j,k}$ and $W_{j,k}$ at scale j . For the WIG model, the $W_{j,k}$'s are mutually independent and identically distributed within scale according to $W_{j,k} \sim N(0, \sigma_j^2)$.

The inverse Haar wavelet transform, computes finer scale scaling coefficients from coarser scale scaling and wavelet coefficients via

$$u_{j,2k} = 2^{-1/2}(u_{j-1,k} + w_{j-1,k}) \quad \text{and} \quad u_{j,2k+1} = 2^{-1/2}(u_{j-1,k} - w_{j-1,k}) \quad (2.8)$$

and is equivalent to moving down the scaling coefficient tree to get finer representations of the signal by adding in the wavelet coefficient terms.

We introduce three different processes: the continuous-time signal $c(t)$, its integral $D(t)$, and a discrete-time approximation $C[k]$ to $c(t)$. These three signals are related by

$$C[k] := \int_{k2^{-n}}^{(k+1)2^{-n}} c(t) dt = D((k+1)2^{-n}) - D(k2^{-n}). \quad (2.9)$$

In this paper, $C[k]$ and $D(t)$ will play rôles analogous to fGn and fBm, respectively.

For notational simplicity, we will assume that both $c(t)$ and $D(t)$ live on $[0, 1]$ and that $C[k]$ is a length- 2^n discrete-time signal. Thus, there is only one scaling coefficient $U_{0,0}$ in (2.4). Note that we use capital letters when the underlying variables

are random. (A more general case with multiple scaling coefficients at the coarsest scale is treated in [36].) $C[k]$ relates directly to the finest-scale scaling coefficients:

$$C[k] = 2^{-n/2} U_{n,k}, \quad k = 0, 1, \dots, 2^n - 1. \quad (2.10)$$

We will focus on modeling $C[k]$ in this paper.

2.3 Wavelet-domain Independent Gaussian (WIG) model

Wavelets serve as an approximate Karhunen-Loève or decorrelating transform for fBm [14], fGn, and more general LRD signals [23]. Hence, the difficult task of modeling these highly correlated signals in the time domain reduces to a simple one of modeling them approximately by an uncorrelated process in the wavelet domain.

The WIG model synthesizes Gaussian LRD data by generating the parent node of the scaling coefficient tree, $U_{0,0}$, with a required Gaussian distribution and wavelet coefficients as independent (and hence uncorrelated) zero-mean Gaussian random variables, identically distributed within scale according to

$$W_{j,k} \sim N(0, \sigma_j^2), \quad (2.11)$$

with σ_j^2 the wavelet-coefficient variance at scale j [14, 44, 41, 22, 27]. Scaling coefficients at finer scales on the tree are then recursively computed through (2.8) until the finest scale scaling coefficients $U_{n,k}$ and hence the required signal $C_{\text{wig}}[k]$ are obtained. The result is a fast $O(N)$ algorithm for generating a length- N signal (see Figure 2.1(c)).

An attractive feature of the WIG model is its flexibility in matching different correlation structures of LRD processes. A power-law decay for the σ_j^2 's leads to approximate wavelet synthesis of fBm or fGn [14, 44]. However, while network traffic may exhibit LRD consistent with fBm or fGn, it may have short-term correlations

that vary considerably from pure fBm or fGn scaling. Such LRD processes can be modeled by setting σ_j^2 to match the measured or theoretical variances of the wavelet coefficients of the desired process [27]. Thus, for a length- N signal, the WIG is characterized by approximately $\log_2(N)$ parameters.

The WIG is an *additive model* because we can express the signal $C_{\text{WIG}}[k]$ directly as a sum of independent random variables. Decomposing each shift k into a binary expansion $k = \sum_{i=0}^{n-1} k'_i 2^{n-1-i}$ defines the $k'_i \in \{0, 1\}$ uniquely and we can write

$$C_{\text{WIG}}[k] = 2^{-n/2} U_{n,k} = 2^{-n} \left(U_{0,0} + \sum_{i=0}^{n-1} (-1)^{k'_i} 2^{i/2} W_{i,k_i} \right), \quad (2.12)$$

with

$$k_0 := 0, \text{ and } k_i = \sum_{j=0}^{i-1} k'_j 2^{i-1-j}, \quad i = 1, \dots, n-1. \quad (2.13)$$

This result can be derived by iteratively applying (2.8).

The WIG model assumes Gaussianity, but network traffic signals (such as loads and interarrival times) can be highly nonGaussian (Figure 1). Not only are these signals strictly non-negative, but they can exhibit “spiky” behavior corresponding to a marginal distribution whose right-side tail decays much more slowly than that of a Gaussian. We seek a more accurate marginal characterization for these spiky, non-negative LRD processes, yet wish to retain the decorrelating properties of wavelets and the simplicity of the WIG model.

Chapter 3

Multifractal Wavelet Model

In order to model non-negative signals using the wavelet transform, we must develop conditions on the scaling and wavelet coefficient values for $c(t)$ in (2.4) to be non-negative. While cumbersome for a general wavelet system,¹ these conditions are simple for the Haar system.

Since the Haar scaling coefficients $u_{j,k}$ represent the local mean of the signal at different scales and shifts, they are non-negative if and only if the signal itself is non-negative; that is, $c(t) \geq 0 \Leftrightarrow u_{j,k} \geq 0, \forall j, k$. This condition leads us directly to a set of constraints on the Haar wavelet coefficients. Combining (2.8) with the constraint $u_{j,k} \geq 0$, we obtain the condition

$$c(t) \geq 0 \Leftrightarrow |w_{j,k}| \leq u_{j,k}, \quad \forall j, k. \quad (3.1)$$

The positivity constraints (3.1) inspire a very simple multiscale, multiplicative signal model for positive processes. In the *multifractal wavelet model* (MWM) [36] we compute the wavelet coefficients recursively by

$$W_{j,k} = A_{j,k} U_{j,k}, \quad (3.2)$$

with $A_{j,k}$ a random variable supported on the interval $[-1, 1]$. We assume that the $A_{j,k}$'s are independent. Together with (2.8), we obtain (see Figure 3.1)

$$U_{j,2k} = 2^{-1/2}(1 + A_{j+1,k}) U_{j-1,k}, \quad U_{j,2k+1} = 2^{-1/2}(1 - A_{j+1,k}) U_{j-1,k}. \quad (3.3)$$

¹The conditions are straightforward also for certain biorthogonal wavelet systems.

See [42] for a similar model used as an intensity prior for wavelet-based image estimation.

The MWM synthesizes a data trace in a similar manner to the WIG with the difference that independent multipliers $A_{j,k}$ are generated instead of independent wavelet coefficients $W_{j,k}$. After generating the coarsest scale scaling coefficient $U_{0,0}$ and the multipliers $A_{j,k}$, the MWM generates scaling coefficients at finer scales of the scaling coefficient tree recursively using (3.3) until the finest scale has been reached. The total cost for computing N MWM signal samples is $O(N)$. In fact, synthesis of a trace of length 2^{18} data points takes just 8 seconds of workstation cpu time.

The MWM is a *multiplicative model* because we can express the signal $C_{\text{MWM}}[k]$ directly as a product (or cascade) of independent random multipliers $1 \pm A_{j,k}$. Using the notation introduced in (2.13), we have

$$C_{\text{MWM}}[k] = 2^{-n/2} U_{n,k} = 2^{-n} U_{0,0} \prod_{i=0}^{n-1} (1 + (-1)^{k_i} A_{i,k_i}), \quad (3.4)$$

which should be compared with (2.12).

It remains to choose an appropriate distribution for the multipliers $A_{j,k}$ and the scaling coefficient $U_{0,0}$. For simplicity of development, we will assume that the $A_{j,k}$'s are mutually independent and independent of $U_{0,0}$. Consequently, $A_{j,k}$ and $U_{j,k}$ are independent for all j, k . We will also assume that the $A_{j,k}$'s are symmetric about 0 and identically distributed within scale; it is easily shown that these two conditions are necessary for the resulting process to be first-order stationary [36]. Due to its flexible shape (see Figure 3.1(b)), compact support and tractability to closed-form calculations, we choose the *symmetric beta distribution*² [21], $\beta_{-1,1}(p, p)$ (see Figure 3.1(b)) for the $A_{j,k}$'s

$$A_{j,k} \sim \beta_{-1,1}(p_j, p_j), \quad (3.5)$$

²We denote a beta random variable with support $[a, b]$ by $\beta_{a,b}$

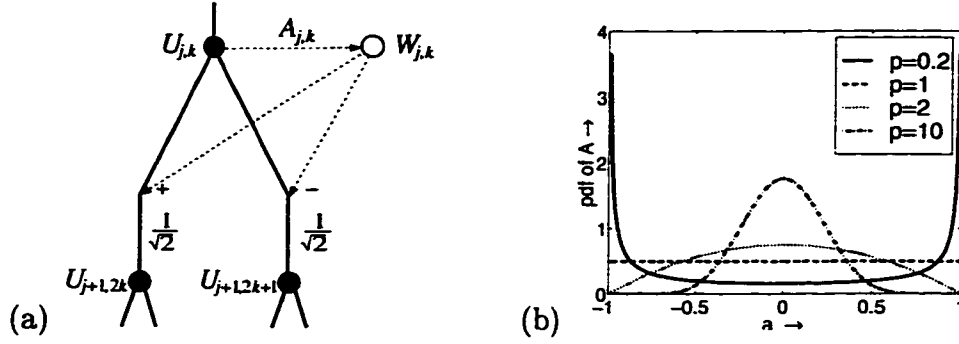


Figure 3.1 : (a) *Multifractal wavelet model (MWM) construction*: At scale j , generate the multiplier $A_{j,k} \sim \beta(p_j, p_j)$, and then form the wavelet coefficient as the product $W_{j,k} = A_{j,k}U_{j,k}$. At scale $j+1$, form the scaling coefficients in the same manner as the WIG model in Figure 2.1(c). (b) *Probability density function of a $\beta_{-1,1}(p, p)$ random variable A* . For $p = 0.2$, $\beta_{-1,1}(p, p)$ resembles a binomial distribution, and for $p = 1$ it has a uniform density. For $p > 1$ the density is close to a truncated Gaussian density with increasing resemblance as p increases.

with p_j the beta parameter at scale j . The variance of a random variable $A \sim \beta(p, p)$ is

$$\text{var}[A] = \frac{1}{2p+1}. \quad (3.6)$$

In the MWM, the p_j play a rôle analogous to the σ_j^2 of the WIG model as they allow us to control the wavelet energy decay across scale through

$$\frac{\text{var}(W_{j-1,k})}{\text{var}(W_{j,k})} = \frac{2 \text{var}[A_{j-1,k}]}{\text{var}[A_{j,k}] (1 + \text{var}[A_{j-1,k}])} = \frac{2p_j + 1}{p_{j-1} + 1}. \quad (3.7)$$

Thus, to model a target process with the MWM, we can select the p_j 's to match the signal's theoretical wavelet-domain energy decay. Or, given training data, we can select the parameters to match the sample variances of the wavelet coefficients as a function of scale. With one beta parameter per wavelet scale, the MWM has approximately $\log_2 N$ parameters for a trace of length N . Distributions with more parameters (e.g., discrete distributions or mixtures of betas) could be used to capture

Table 3.1 : *Comparison of the tree-based WIG and MWM models. For approximating a signal with a strict fGn covariance structure, both the WIG and MWM require only three parameters (mean, variance, and H).*

WIG	MWM
Additive	Multiplicative
Gaussian	Asymptotically Lognormal
LRD matched	LRD matched
$\log_2 N + 2$ parameters	$\log_2 N + 2$ parameters
$O(N)$ synthesis	$O(N)$ synthesis

high-order data moments at a cost of increased model complexity [36]. See Table 3.1 for a comparison of the WIG and MWM properties.

To complete the modeling, we must choose the parameter p_0 of the model and characterize the distribution of the coarsest scaling coefficient $U_{0,0}$. From (3.2) and (3.6) we obtain

$$(2p_0 + 1)\text{var}(W_{0,0}) = \mathbb{E}[U_{0,0}^2], \quad (3.8)$$

which allows us to calculate p_0 from estimates of $\mathbb{E}[U_{0,0}^2]$ and $\text{var}(W_{0,0})$. A precise model of $U_{0,0}$ would require a strictly non-negative probability density function to ensure the non-negativity of the MWM output. In our simulation experiments in Chapter 5, we use a β random variable with positive support, that is, $U_{0,0} \sim \beta_{0,M}(p, q)$ with $M \geq 0$. See Figure 1(c) for a sample realization of the MWM. Clearly, the MWM produces positive “spiky” data akin to the real traffic unlike the WIG model.

Chapter 4

Queuing Analysis of Wavelet-Based Models

The importance of queuing analysis in network design and control cannot be overemphasized. Buffer dimensioning in routers and call admission control are but two of the many crucial areas in networking research that rely on an accurate characterization of the queuing behavior of data traffic.

The discovery of LRD in traffic has created a challenging new area of research in queuing theory. The strong correlations in LRD traffic are responsible for its “bursty” nature that can lead to higher packet loss in queues than that predicted by short-range dependent (SRD) models [12]. This finding has been bolstered by theoretical queuing results for fGn, the pre-eminent LRD traffic model at present. When fGn is input to an infinite-length queue with constant service rate, the tail queue distributions decay asymptotically with a Weibullian law

$$P[Q > x] \simeq \exp(-\delta x^{2-2H}), \quad (4.1)$$

with δ a positive constant that depends on the service rate of the queue [11, 29]. Clearly, (4.1) reveals that the decay of the tail queue distribution for fGn with $H > 1/2$ is much slower than the exponential decay predicted by SRD classical models [12] which correspond to the case $H = 1/2$. In spite of this result, there is still an ongoing discussion on the effect of LRD on queuing, with researchers arguing both for and against its importance [38, 17, 34, 32, 28, 16].

In Chapter 3 we described the versatile MWM model that is capable of capturing the correlation structure of traffic as well generating a positive output. Since the

MWM has a fast $O(N)$ data synthesis algorithm, it is of potential use for simulation experiments. In order to expand the range of applications of the MWM beyond simulation, we provide in this section a theoretical queuing analysis of it. We exploit the inherent binary tree structure of the Haar scaling coefficients of the MWM to derive a simple approximate formula for the tail queue probability. Our analysis is also applicable to other tree-based models like the WIG. We will show later in Chapter 5 that our theoretical result is a good approximation of the empirical tail queue behavior of both the WIG and the MWM.

4.1 Queue size and multiple time scales

Consider a discrete time random process X_i , $i \in \mathbb{Z}$, which we regard as the workload entering an infinite buffer single server queue with constant link capacity c . Let Q_i represent the queue size at time instant i . Denote by K_r the aggregate traffic arriving between time instants $-r + 1$ and 0, that is,

$$K_r = \sum_{i=-r+1}^0 X_i. \quad (4.2)$$

In the sequel, we refer to K_r as representing data at time-scale r . We set $K_0 := 0$. The famous Lindley's equation [26] gives

$$\begin{aligned} Q_0 &= \max(Q_{-1} + X_0 - c, 0) = \max(\max(Q_{-2} + X_{-1} - c, 0) + X_0 - c, 0) \\ &= \max(Q_{-2} + X_{-1} + X_0 - 2c, X_0 - c, 0) \end{aligned} \quad (4.3)$$

$$= \max(Q_{-2} + K_2 - 2c, K_1 - c, K_0). \quad (4.4)$$

Proceeding iteratively we obtain

$$Q_0 = \max(Q_{-r} + K_r - rc, \dots, K_0). \quad (4.5)$$

Since $Q_{-r} \geq 0$ for all r we must have $Q_0 \geq \sup_{r \in \mathbb{Z}_+} (K_r - rc)$. Denoting by $-j$ the last instant the queue was empty before time instant 0 (we set $-j = 0$ if $Q_0 = 0$),

we obtain $Q_0 = K_j - jc \leq \sup_{r \in \mathbb{Z}_+} (K_r - rc)$. Thus if the queue was empty at some time in the past,

$$Q_0 = \sup_{r \in \mathbb{Z}_+} (K_r - rc). \quad (4.6)$$

We will study the quantity \tilde{Q}_0 obtained by taking the supremum in (4.6) over a finite subset of \mathbb{Z}_+ corresponding to powers of 2 values for n (or dyadic time scales), that is,

$$\tilde{Q}_0 := \sup_{m \in \{0, \dots, n\}} (K_{2^m} - c2^m), \quad (4.7)$$

for some fixed $n \in \mathbb{Z}_+$. Clearly, $\tilde{Q}_0 \leq Q_0$.

We will approximate the desired tail queue probability by¹ $\mathbf{P}(\tilde{Q}_0 > b) \approx \mathbf{P}(Q_0 > b)$. Assuming the existence of a *critical time scale* (CTS) [38, 28, 16], that is, r^* such that $\mathbf{P}(K_{r^*} - cr^* > b) \approx \mathbf{P}(Q_0 > b)$, we present the following two heuristic arguments for our approximation:

1. Dyadic time scales, though a small subset of \mathbb{Z}_+ , span the entire range of time scales. This ensures that the nearest dyadic time scale to the CTS captures its effect on $\mathbf{P}(Q_0 > b)$. One argument for considering dyadic time scales in (4.7) (i.e. $r \in \{2^m : m \in \{0, \dots, n_0\}; n_0 \in \mathbb{Z}_+\}$) rather than, say, equally spaced time scales (i.e. $r \in \{1 + \gamma m : m \in \{0, \dots, n_1\}; n_1, \gamma \in \mathbb{Z}_+\}$), is that the analysis does not change significantly if we start analyzing data at a coarser time resolution. As an illustration, let us say we have a real data trace of packets arriving every milli-second and wish to study the tail queue behavior of the underlying real process using (4.7). By using equally spaced time scales with $\gamma = 3$, that is $r \in \{1, 4, 7, 10, \dots\}$, we will consider data at time resolutions 1ms, 4ms, 7ms and so on. If instead we had the same real data set but with a finest resolution

¹Here \approx denotes that two quantities are approximately equal.

corresponding to bytes per 2ms, we would consider time resolutions 2ms, 8ms, 14ms, 20ms etc. Intuitively, we would expect these analyses to give different results. If instead we use dyadic time scales, we will use time resolutions 1ms, 2ms, 4ms, 8ms etc. in the first case and 2ms, 4ms, 8ms etc. in the second. Our results will be the same barring the effect of the finest time scale (1ms) on the queue size.

2. The approximation will capture the effect of the CTS if $n > r^*$ and we are thus justified in neglecting time scales $r > n$.

4.2 Queuing analysis

4.2.1 Queuing formula for tree-based models

Performing an exact queuing analysis of tree-based models like the WIG and MWM is very complicated because their binary tree naturally produces a process that is not *strictly stationary* [36]. We would thus expect the distribution of the queue size to vary with time. As an illustration, in Figure 2.1(b) the neighboring nodes $U_{j+2,4k}$ and $U_{j+2,4k+1}$ have a parent node $U_{j+1,2k}$ at scale $j+1$ while the nodes $U_{j+2,4k+1}$ and $U_{j+2,4k+2}$ do not.

Here, we will concentrate on the tail queue probability of the models at the instant the last data point of the model output, $C[2^n - 1]$, enters the queue. In other words, we choose $X_i = C[2^n - 1 + i]$. The Haar scaling coefficients on the branch linking $U_{0,0}$ and $U_{n,2^n-1}$, or right edge of the tree of Figure 2.1(b), are related to the quantities K_{2^m} (4.2), that is

$$K_{2^{n-i}} = 2^{-i/2} U_{i,2^i-1}, \quad \text{for } i = 0, \dots, n, \quad (4.8)$$

and a queuing analysis is feasible.

Experimental evidence indicates that such an analysis closely approximates the empirically observed tail queue probability obtained by passing several synthetic traces of the models through a queue. Though it is not obvious why this must be true, there is some intuition as to why the result of such an analysis must closely match the queuing behavior of stationary data that we model. Say we are modeling fGn, $G(k)$, with the WIG and are interested in obtaining the tail queuing probability at the time instant $G(2^n - 1)$ enters the queue. Then on taking a Haar wavelet transform of fGn data points between 0 and $2^n - 1$, the quantities K_{2^m} correspond to the nodes on the right edge of the WIG's scaling coefficient tree. Since the Haar wavelet coefficients of fGn are nearly uncorrelated [22], the corresponding WIG scaling coefficients model the quantities K_{2^m} well.

For the ease of notation let us denote \tilde{Q}_0 and Q_0 of Section 4.1 by \tilde{Q} and Q respectively. Let E_i denote the event $\{K_{2^{n-i}} < b + c2^{n-i}\}$. The following Lemma simplifies our analysis.

Lemma: *Assume that the events E_i are of the form $E_i = \{J_i < b_i\}$, where $J_i = J_{i-1} + L_{i-1}$ for $1 \leq i \leq n$ and where J_0, L_0, \dots, L_n are independent. Then for $1 \leq i \leq n$.*

$$P(E_i | E_{i-1}, \dots, E_0) \geq P(E_i).$$

Proof: Given in the appendix.

Assuming the conditions of the Lemma to be true

$$\begin{aligned} P(\tilde{Q} > b) &= 1 - P(\tilde{Q} < b) = 1 - P(\cap_{i=0}^n E_i) \text{ from (4.7)} \\ &= 1 - P(E_0) \prod_{i=1}^n P(E_i | E_{i-1}, \dots, E_0) \\ &\leq 1 - \prod_{i=0}^n P(E_i) =: P_{\text{app}}(\tilde{Q} > b). \end{aligned} \tag{4.9}$$

We thus arrive at an upper bound approximation $P_{\text{app}}(\tilde{Q} > b)$ of $P(\tilde{Q} > b)$ by

assuming the events E_i to be independent. Intuitively, our approximation (4.9) not only assumes that dyadic scales are sufficient for an accurate queuing analysis, but also that they are sparse enough to be considered independently.

Recall from Section 4.1 that $\tilde{Q} \leq Q$. This implies that $\mathbf{P}(\tilde{Q} > b) \leq \mathbf{P}(Q > b)$, which means that $\mathbf{P}_{\text{app}}(\tilde{Q} > b)$ is an upper bound of a lower bound on $\mathbf{P}(Q > b)$. Our queuing approximation is thus

$$\boxed{\mathbf{P}(Q > b) \approx 1 - \prod_{i=0}^n \mathbf{P}(K_{2^{n-i}} < b + c2^{n-i}) =: \mathbf{P}_{\text{app}}(\tilde{Q} > b).} \quad (4.10)$$

4.2.2 Queuing analysis of the WIG

For the WIG, on choosing $J_0 := U_{0,0}$ and $L_i := -2^{i/2}W_{i,2^i-1}$ we obtain from (4.8)

$$K_{2^{n-i}} = 2^{-i} \left(U_{0,0} + \sum_{j=0}^{i-1} L_j \right) = 2^{-i} J_i. \quad (4.11)$$

Setting $b_i = b2^i + c2^n$, we observe that the WIG satisfies the conditions of the Lemma.

We thus use (4.10) to approximate $\mathbf{P}(Q > b)$ for the WIG.

Since for the WIG $K_{2^{n-i}}$ is Gaussian, by estimating the mean and variance of $K_{2^{n-i}}$, the probability $\mathbf{P}(E_i)$ can be computed from the cumulative distribution of a Gaussian distribution for which numerous approximations are available [21].

4.2.3 Queuing analysis of the MWM

Denoting $A_{j,2^j-1}$ by A_j , (4.8) reduces to

$$K_{2^{n-i}} = U_{0,0} \prod_{j=0}^{i-1} (1 - A_j) / 2. \quad (4.12)$$

The event E_i is thus

$$\begin{aligned} E_i &= \{K_{2^{n-i}} < b + c2^{n-i}\} = \{U_0 \prod_{j=0}^{i-1} (1 - A_j) < b2^i + c2^n\} \\ &= \left\{ \log(U_0) + \sum_{j=0}^{i-1} \log(1 - A_j) < \log(b2^i + c2^n) \right\}. \end{aligned} \quad (4.13)$$

By setting $J_0 := \log(U_0)$, $L_i := \log(1 - A_i)$ and $b_i := \log(b2^i + c2^n)$ we see that the MWM satisfies the Lemma. Consequently, we use (4.10) to approximate $\mathbf{P}(Q > b)$ for the MWM.

For the MWM, obtaining $\mathbf{P}(E_i)$ is not as straightforward as for the WIG. If $U_{0,0}$ is equal to a constant M times the random variable $\beta_{0,1}(p_{-1}, q_{-1})$, then from (4.12) K_{2^n-i} is M times several independent $\beta_{0,1}$ random variables. We approximate K_{2^n-i}/M by a beta random variable, $\beta_{0,1}(d_i, e_i)$, using Fan's approximation [13]. Thus, if $(1 - A_j)/2 \sim \beta_{0,1}(p_j, q_j)$ then

$$d_i = S(T - S^2)^{-1}(S - T) \text{ and } e_i = (1 - S)(T - S^2)^{-1}(S - T), \quad (4.14)$$

where

$$S = \prod_{j=-1}^{i-1} \frac{p_j}{p_j + q_j} \text{ and } T = \prod_{j=-1}^{i-1} \frac{p_j(p_j + 1)}{(p_j + q_j)(p_j + q_j + 1)}. \quad (4.15)$$

This approximation matches the mean and variance of the actual distribution of K_{2^n-i} exactly and closely approximates the first 10 moments [13]. We thus use the cumulative distribution of a β random variable to calculate $\mathbf{P}(E_i)$, for which several approximations are available [21].

Thus given model parameters for the WIG and MWM, \mathbf{P}_{app} can be calculated as above. If instead, we model training data with the WIG and MWM we must first estimate the model parameters after taking a fast $O(N)$ wavelet transform of the data.

Chapter 5

Experimental Results

The first goal of this section is to evaluate the capability of the MWM in modeling LAN and WAN traffic and compare it to the WIG. The second goal is to determine the accuracy of our theoretical queuing approximation (4.10). Although the LRD of network traffic was first established in Ethernet LAN traffic [24] and later in WAN traffic [33], few models exist for them, and LRD traffic modeling has been mainly restricted to video traffic [27, 20].

5.1 Real data

We use two well-known real data traces in our experiments. The first (LBL-TCP-3) contains two hours wide-area TCP traffic between the Lawrence Berkeley Laboratory and the rest of the world in 1994 [33]. We form a data trace by counting the number of bytes that arrive in consecutive time intervals of 6ms and use the first 2^{20} data points in our simulation experiments. This trace has a sample mean of 257.5 bytes/(unit time) and sample standard deviation of 562.6 bytes/(unit time).

The second real data set (BC-pAug89) is one of the celebrated Ethernet data traces collected at Bellcore Morristown Research and Engineering facility in 1989 and has been extensively analyzed [24]. We obtain a data trace by summing the bytes of packets that arrive in consecutive time intervals of 2.6ms. We use the first 2^{20} data points of this trace in our experiments. This trace has mean 345.8 bytes/(unit time) and standard deviation 703.4 bytes/(unit time). The BC-pAug89 trace is approxi-

mately a realization of a second-order self similar process with $H = 0.79$ [2].

5.2 Matching of Real Data

In order to study how well the WIG and MWM models match real data, we train them on the the real data traces. To fit the WIG and MWM models to the data, we use the procedure outlined in Section 2.3 and Chapter 3, which involves taking a Haar wavelet transform of the real data and estimating the variances σ_j^2 of the wavelet coefficients at each scale. We estimate these variances only at the 15 finest scales, because at coarser scales there are not a sufficient number of coefficients to obtain good variance estimates. As a result, we synthesize data traces of maximum length 2^{15} data points. For the MWM, we model the coarsest-scale scaling coefficient $U_{0,0}$ as a constant times a symmetric¹ $\beta_{0,1}$ random variable with mean and variance equal to the sample mean and variance of the scaling coefficients of the real data at this scale.

With trained models in hand, we now generate synthetic data traces. Due to space constraints, fitting results for only the LBL-TCP-3 trace are presented here (see [36] for results on BC-pAug89). Recall from Figure 1 that visually the synthetic MWM looks very similar to the real trace while the WIG does not. On comparing the marginals of WIG and MWM traces to that of the LBL-TCP-3 trace at three different aggregation levels (Figure 5.1), we observe that the MWM marginals are similar to that of the real data trace, while the Gaussian WIG marginals differ significantly.

¹ At very coarse time resolutions where the data is approximately Gaussian, a symmetric β random variable can suffice to model the scaling coefficients. If the coarsest scale is chosen to coincide with time scales where a Gaussian approximation may not be appropriate, an asymmetric $\beta(p, q)$ distribution, that is $p \neq q$, would be a more appropriate choice.

To compare the correlation matching abilities of the two models, we plot variance-time plots of the real data, the MWM traces, and the WIG traces in Figure 5.2. The variance-time plot estimates were obtained by averaging the empirical variance-time plots of 32 independent realizations of the models. Observe that, as expected, both the MWM and WIG models do a good job of matching the correlation structure of the real data.

5.3 Queuing experiments

Intuitively, the more traffic characteristics (correlation structure, marginals etc.) a model matches, the better will it match the queuing behavior of real traffic. Hence, it is not surprising that a perfect fitting of second-order correlations *and* marginals as done in [18] leads to a good match of queuing behavior. In contrast to [18] that modeled data using several parameters, the MWM through few parameters, closely approximates the correlation structure of a target data set and naturally produces a marginal distribution resembling that of the real trace. Since both the WIG and MWM capture the correlation structure of real data, we study the effect of marginals on queuing behavior by comparing their ability to capture the queueing behavior of the two real data sets.

In all experiments, the coarsest scale scaling coefficient in the MWM is distributed as a symmetric beta random variable.

In experiments with real traces, data traces are fed as input to an infinite-length single-server queue with link capacity 800 bytes/unit time. We estimate the tail queue probabilities of the various data traces as

$$\hat{P}_i[Q > x] = \frac{\text{number of time instants } Q > x}{\text{total time duration of trace } i}. \quad (5.1)$$

We also provide estimated confidence intervals with confidence level of 95% for the

estimated queue distribution $(1/L)\sum_{i=1}^L \hat{P}_i[Q > x]$, where L is the total number of traces, assuming that it is a Gaussian random variable [19].

With both real traces, we performed the same queuing experiment. After synthesizing 1000 WIG and MWM traces of length 2^{15} , we fed them as input to our theoretical queue and obtained their queuing behavior.

In Figure 5.3 we compare the average queuing behavior of the WIG and MWM traces to that of the real traces. Observe that the MWM traces closely match the queuing behavior of the real data traces while the WIG traces do not. Also observe that our theoretical queuing approximation (P_{app}) of (4.10) tracks the tail queue probability of the models closely.

To confirm the accuracy of (4.10) further experiments were performed with synthetic WIG and MWM traces. We chose model parameters corresponding to an fGn correlation structure and varied the mean, variance and Hurst parameter and fixed the link capacity at $c = 10$ units. Each trace was of length 2^{20} data points. Due to space constraints, we present results for only few cases in Figure 5.4. In all cases we observe that (4.10) is indeed a good approximation to the empirical tail queue behavior and that the MWM exhibits larger tail queue probabilities than the WIG.

These queuing experiments indicate that the correlation structure of traffic is not the only factor with an impact on the queuing behavior of data traffic. Since the MWM outperforms the WIG model in matching queuing behavior, we conclude that the marginals have a substantial effect on the queuing behavior of traffic.

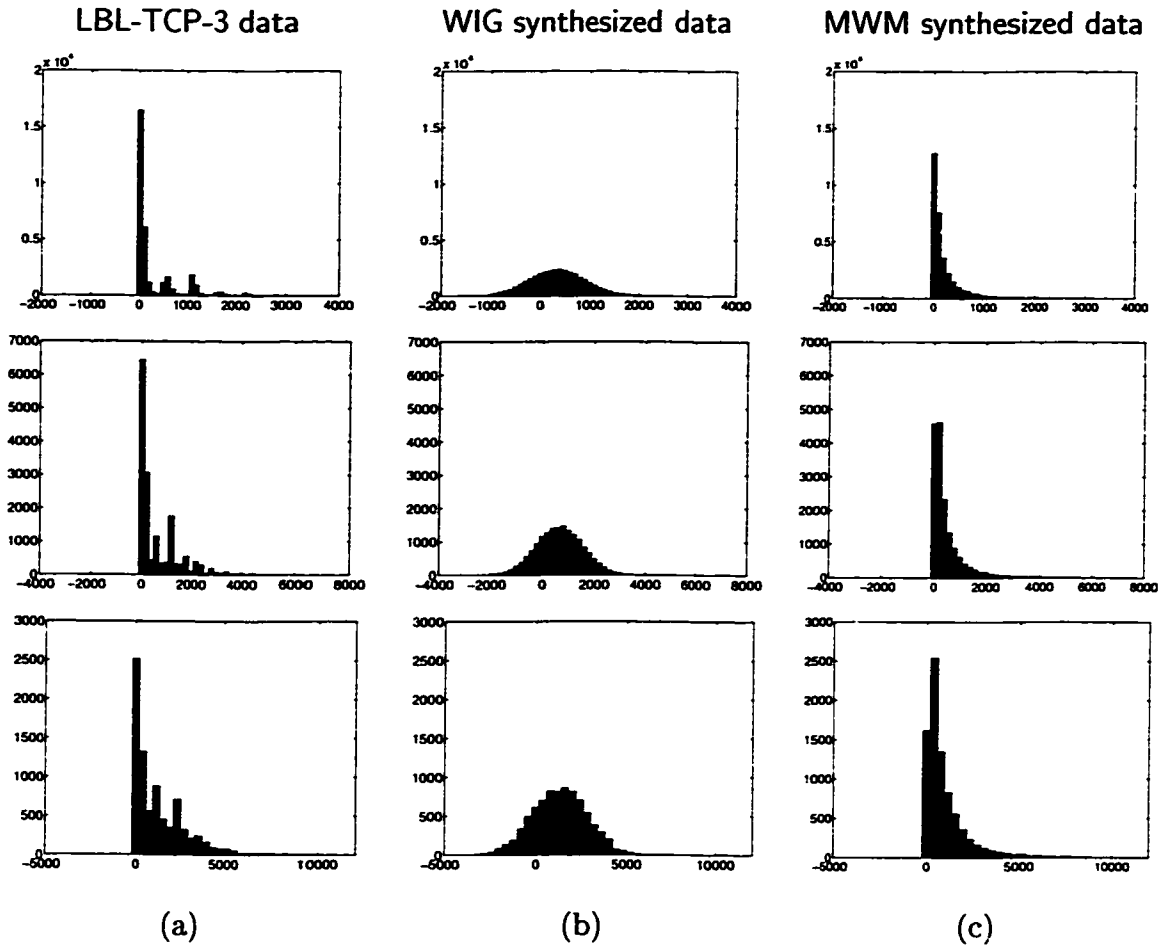


Figure 5.1 : Histograms of the bytes-per-times process at different aggregation levels for (a) wide-area TCP traffic at the Lawrence Berkeley Laboratory (trace LBL-TCP-3) [33], (b) one realization of the WIG model, and (c) one realization of the MWM synthesis. The top, middle and bottom plots correspond to bytes arriving in intervals of 6ms, 12ms and 24ms respectively. Note the large probability mass over negative values for the WIG model.

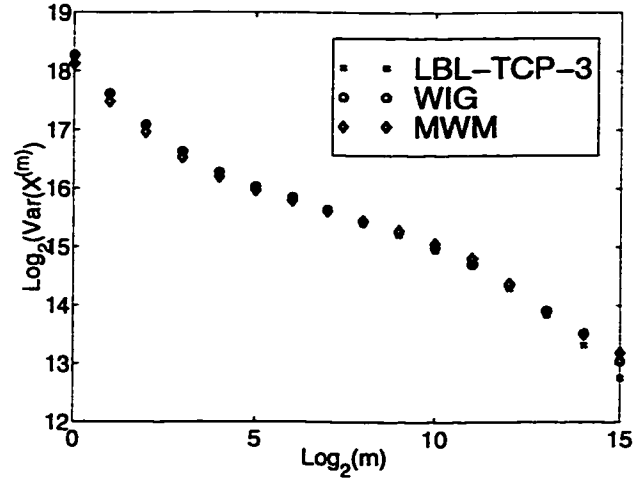


Figure 5.2 : Variance-time plot of the LBL-TCP-3 data “x”, the WIGdata “o”, and one realization of the MWM synthesis “o”. Both the MWM and WIG capture the correlation structure of the real data.

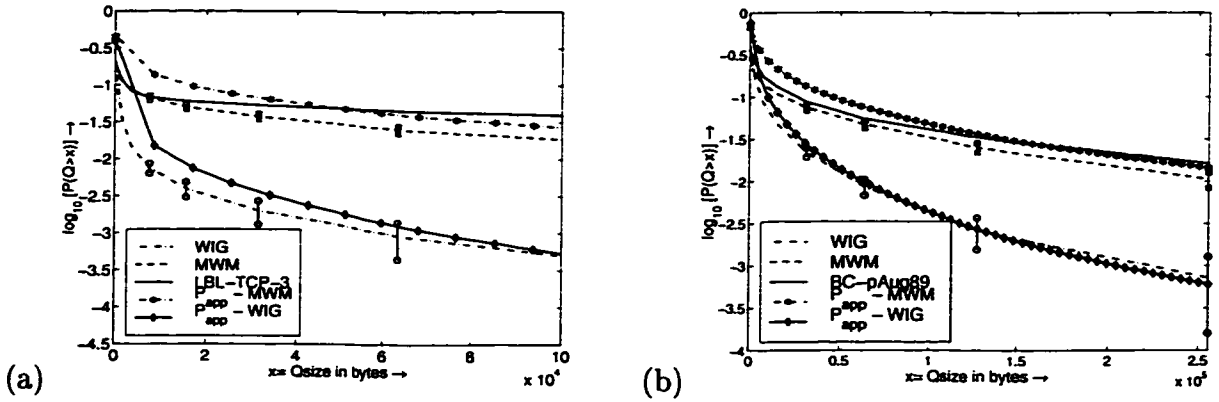


Figure 5.3 : Comparison of the queuing performance of real data traces with those of synthetic WIG and MWM traces. In (a), we observe that the MWM synthesis matches the queuing behavior of the LBL-TCP-3 data closely, while the WIG synthesis does not. In (b), we observe a similar behavior with the BC-pAug89 data. We also observe that our theoretical prediction of queuing behavior for the WIG and MWM matches their empirical queuing behavior.

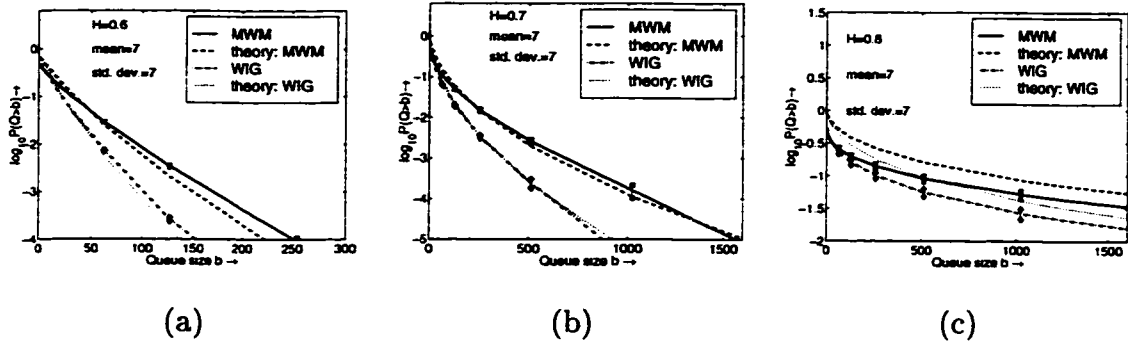


Figure 5.4 : Validation of theoretical formula (4.10) for tail queue probability of the MWM and WIG models. Experiments used synthetic WIG and MWM traces corresponding to an fGn correlation structure for different values of Hurst parameter H . In (a) $H = 0.6$, in (b) $H = 0.7$ and in (c) $H = 0.8$. In all cases, the mean, standard deviation and link capacity were 7, 7 and 10 units respectively. Observe that in all cases the formula gives a good approximation to the empirical queuing behavior. Also observe that the MWM exhibits a higher tail queue probability in all cases.

Chapter 6

The MWM is a Cascade

We now link the MWM with the theory of multiplicative cascades. Cascades provide a natural framework for producing positive “bursty” processes and offer greater flexibility and richer scaling properties than fractal models such as fGn and fBm. Closely related to cascades is the powerful theory of *multifractals*, which provides statistical tools for measuring “burstiness”.

6.1 Cascades

The backbone of a cascade is a construction where one starts at a coarse scale and develops details of the process on finer scales iteratively in a multiplicative fashion. This construction procedure naturally results in a process that “sits” just above the zero line and emits occasional positive jumps or spikes. In contrast, additive self-similar models such as fGn and the WIG “hover” around the mean with occasional outbursts in both positive and negative directions.

The MWM is a multiplicative cascade, as (3.3) and (3.4) reveal (see Figure 6.1(a)). In accordance with the notation for cascades, setting

$$M_0^0 = U_0^0 \quad \text{and} \quad M_{k_i}^i = \frac{(1 + (-1)^{k'_{i-1}} A_{i-1, k_{i-1}})}{2}, \quad 0 < i \leq n, \quad (6.1)$$

and substituting into (3.4) leads us to (see Figure 6.1(a))

$$C_{\text{MWM}}[k] = 2^{-n} M_0^0 \prod_{i=1}^n M_{i, k_i}, \quad (6.2)$$

with the k_i and k'_i defined as in (3.4).

6.2 Multifractal analysis

Intuitively, multifractal analysis measures the frequency with which bursts of different strengths occur in a signal. Consider a positive process $Y(t)$. The strength of the burst of Y at time t , also called the degree of *Hölder continuity*, can be characterized by

$$\alpha(t) = \lim_{k_n 2^{-n} \rightarrow t} \alpha_{k_n}^n \quad \text{where} \quad \alpha_{k_n}^n := -\frac{1}{n} \log_2 |Y((k_n + 1)2^{-n}) - Y(k_n 2^{-n})| \quad (6.3)$$

where $k_n 2^{-n} \rightarrow t$ means that $t \in [k_n 2^{-n}, (k_n + 1)2^{-n})$ and $n \rightarrow \infty$. The smaller the $\alpha(t)$, the larger the increments of Y around time t , and the “burstier” it is at time t . The frequency of occurrence of a given strength α , can be measured by the *multifractal spectrum*:

$$f(\alpha) := \dim\{t : \alpha(t) = \alpha\} \quad (6.4)$$

By definition, f takes values between 0 and 1 and is often shaped like a \cap and concave. The smaller the $f(\alpha)$, the “fewer” points t will exhibit $\alpha(t) \approx \alpha$. If $\bar{\alpha}$ denotes the value $\alpha(t)$ assumed by “most” points t , then $f(\bar{\alpha}) = 1$. See Figure 6.1 for the multifractal spectrum of the LBL-TCP-3 data set and of synthetic MWM data. We observe that the MWM captures the spectrum of the real data except for large values of α . This means that the MWM does not generate as many small values as the signal possesses.

6.3 Multifractal spectrum and higher-order moments

Though (6.4) gives us a simple measure of burstiness in data, in practice it is impossible to compute the right side of (6.4). However, $f(\alpha)$ can be obtained through the use of high and low-order moments of the signal $Y(t)$.

Define the *partition function* that captures the scaling of different moments of Y as

$$T(q) := \lim_{n \rightarrow \infty} \frac{1}{-n} \log_2 \mathbb{E}[S_n(q)], \quad (6.5)$$

with

$$S_n(q) := \sum_{k_n=0}^{2^n-1} |Y((k_n+1)2^{-n}) - Y(k_n 2^{-n})|^q = \sum_{k_n=0}^{2^n-1} 2^{n\alpha_{k_n}^n}. \quad (6.6)$$

The multifractal spectrum $f(\alpha)$ and $T(q)$ are closely related, as the following hand-waving argument shows. Grouping in the sum $S_n(q)$ of (6.6) the terms behaving as $\alpha_{k_n}^n \approx \alpha$, and using (6.4) we get

$$\begin{aligned} S_n(q) &= \sum_{\alpha} \sum_{\alpha_n \sim \alpha} (2^{-n\alpha})^q \approx \sum_{\alpha} 2^{nf(\alpha)} 2^{-nq\alpha} \\ &\approx 2^{-n \inf_{\alpha} (q\alpha - f(\alpha))}. \end{aligned} \quad (6.7)$$

We conclude that we must “expect” $T(q)$ to equal $\inf_{\alpha} (q\alpha - f(\alpha))$, the so-called *Legendre transform* of $f(\alpha)$. For the special case of an MWM process, i.e., $Y = D$ (see Section 2.2 for the definition of D), it can be shown (see [35]) that the inverse relation holds, called the *multifractal formalism*

$$f(\alpha) = T^*(\alpha) := \inf_q (q\alpha - T(q)). \quad (6.8)$$

In order to estimate $T(q)$ from a data set, it is customary to use the approximation $2^{-nT(q)} \approx S_n(q)$. For the MWM this is equivalent to

$$2^{-jT(q)} \approx \sum_{k=0}^{2^j-1} |2^{-j/2} U_{j,k}|^q. \quad (6.9)$$

The slope of a linear fit of $\log S_{(j)}(q)$ against j will give $T(q)$.

For the MWM, assuming the moments of the multipliers M_{i,k_i} converge to a limiting random variable $M \sim \beta_{0,1}(p, p)$, we find

$$T_D(q) = \begin{cases} -1 - \log_2 \mathbb{E}[M^q] = -1 - \log_2 \frac{\Gamma(p+q)\Gamma(2p)}{\Gamma(2p+q)\Gamma(p)} & \text{if } q > -p \\ -\infty & \text{if } q \leq -p. \end{cases} \quad (6.10)$$

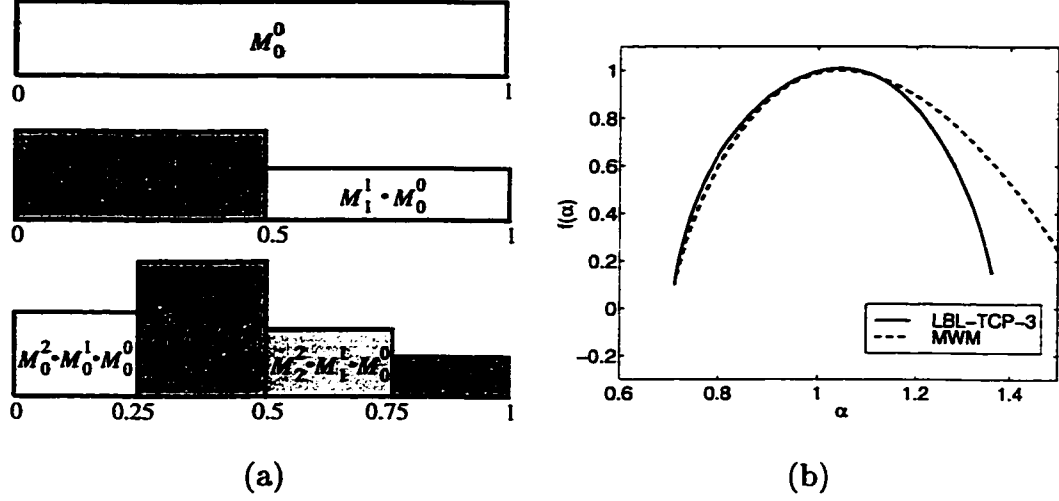


Figure 6.1 : (a): The MWM translates immediately into a multiplicative cascade in the time domain (cf. (6.2)). (b) Multifractal spectra of the LBL-TCP-3 data and one realization of the MWM synthesis. The MWM spectrum matches that of the real data closely except for large values of α or small values of the signal.

For the self-similar fBm,

$$T_{\text{fBm}}(q) = \begin{cases} qH - 1 & \text{for } q > -1, \\ -\infty & \text{for } q \leq -1. \end{cases} \quad (6.11)$$

On taking the Legendre transform of T_{fBm} we observe that fBm possesses only one degree of “burstiness” ($\alpha(t) = H$) which is omnipresent. Consequently, fBm (or its increments process fGn) cannot capture the complicated multifractal behavior or “burstiness” of real data like the LBL-TCP-3 trace (Figure 6.1).

Chapter 7

Conclusions

The MWM provides a new multiscale tool for synthesis of nonGaussian LRD traffic. Computations involving the MWM are extremely efficient — synthesis of a trace of N sample points requires only $O(N)$ computations. In fact, synthesis of even long 2^{18} point data sets takes just seconds of workstation cpu time. The parameters of the MWM, numbering approximately $\log N$, are identical in number to the WIG model and are simple enough to be either inferred from observed data or chosen a priori. We can reduce the number of parameters further by developing a parametric characterization of the wavelet energy decay across scale.

With the MWM, we have been able to fit actual traffic traces, and have developed preliminary queueing results that demonstrate the impact of the nonGaussian nature of traffic on queueing performance.

We derived an approximate queueing formula for the MWM and demonstrated its accuracy through experiments. As a consequence, the versatile MWM model is now viable for numerous applications including call admission control.

Further research could make the MWM practicable for data prediction. The parameters of the MWM could also be used to capture the effect of different protocols on shaping data traffic (e.g., the TCP protocol). In short, the use of the MWM in real-time network protocols and control algorithms seems very promising.

Appendix A

Proof of Lemma

Lemma: Assume that the events E_i are of the form $E_i = \{J_i < b_i\}$, where $J_i = J_{i-1} + L_{i-1}$ for $1 \leq i \leq n$ and L_i are mutually independent and independent of J_0 . Then $\mathbf{P}(E_i | E_{i-1}, \dots, E_0) \geq \mathbf{P}(E_i)$ for $1 \leq i \leq n$.

Proof

Let f_Z and F_Z denote the probability density function and cumulative density function, respectively, of a random variable Z . We will simplify the notation by denoting $F_{Z|E}(z|E)$, where E is an event, by $F_{Z|E}(z)$. Let $J_{i,1} := J_i | E_{i-1}, \dots, E_0$ for $i \geq 1$ and $J_{0,1} = J_0$. Also, let $J_{i,0} := J_i | E_i, \dots, E_0$. Then

$$F_{J_{i,1}}(x) = \begin{cases} F_{J_i | E_{i-1}, \dots, E_0}(x) & \text{for } i \geq 1 \\ F_{J_0}(x), & i = 0, \forall x \in \mathbb{R}, \end{cases} \quad (\text{A.1})$$

and

$$F_{J_{i,0}}(x) = F_{J_i | E_i, \dots, E_0}(x) \quad \forall x \in \mathbb{R}. \quad (\text{A.2})$$

Claim: $F_{J_{i,1}}(x) \geq F_{J_i}(x) \quad \forall x \in \mathbb{R}$ and $\forall i$.

From (A.1) we see that this claim is true for $i = 0$. Let us assume

$$F_{J_{i,1}} \geq F_{J_i}. \quad (\text{A.3})$$

The key fact to note is that $J_{i+1,1} = J_{i,0} + L_i$, since L_i is independent of J_j and hence of the events E_j for $j < i$. Now from Baye's rule [31]

$$\begin{aligned} F_{J_{i,0}}(x) &= \begin{cases} \frac{F_{J_{i,1}}(x)}{F_{J_{i,1}}(b_i)} & \text{if } x \leq b_i \\ 1 & \text{otherwise} \end{cases} \\ &\geq F_{J_{i,1}}(x) \end{aligned} \quad (\text{A.4})$$

$$\begin{aligned}
F_{J_{i+1},1}(x) &= \mathbf{P}(J_{i,0} + L_i < x) \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{x-l_i} f_{J_{i,0}}(j_{i,0}) f_{L_i}(l_i) \, dj_{i,0} \, dl_i \\
&= \int_{-\infty}^{\infty} F_{J_{i,0}}(x - l_i) f_{L_i}(l_i) \, dl_i \\
&\geq \int_{-\infty}^{\infty} F_{J_{i,1}}(x - l_i) f_{L_i}(l_i) \, dl_i && \text{from (A.4)} \\
&\geq \int_{-\infty}^{\infty} F_{J_i}(x - l_i) f_{L_i}(l_i) \, dl_i && \text{from (A.3)} \\
&= \mathbf{P}(J_i + L_i < x) \\
&= F_{J_{i+1}}(x)
\end{aligned} \tag{A.5}$$

Thus, by induction the claim is proved. Since the claim is true $\forall x \in \mathbb{R}$, by setting $x = b_{i+1}$ in (A.5) the lemma is proved. \diamond

Bibliography

- [1] P. Abry, P. Gonçalves, and P. Flandrin. Wavelets, spectrum analysis and $1/f$ processes. *preprint*, 1996.
- [2] P. Abry and D. Veitch. Wavelet analysis of long range dependent traffic. *IEEE Trans. Inform. Theory*, 4(1):2–15, 1998.
- [3] S. Bates and S. McLaughlin. The estimation of stable distribution parameters from teletraffic data. *preprint*, 1998.
- [4] F. Brichet, J. Roberts, A. Simonian, and D. Veitch. Heavy traffic analysis of a fluid queue fed by a superposition of ON/OFF sources. *COST*, 242, 1994.
- [5] C. S. Burrus, R. A. Gopinath, and H. Guo. *Introduction to Wavelets and Wavelet Transforms: A Primer*. Prentice Hall, 1998.
- [6] J. Choe and N.B. Shroff. Supremum distribution of gaussian processes and queueing analysis including long-range dependence and self-similarity. *Stochastic Models* submitted, 1997.
- [7] D. Cox. Long-range dependence: A review. *Statistics: An Appraisal*, pages 55–74, 1984.
- [8] M. Crovella and A. Bestavros. Self-similarity in World Wide Web traffic. Evidence and possible causes. In *Proceedings of SIGMETRICS '96*, May 1996.
- [9] I. Daubechies. *Ten Lectures on Wavelets*. SIAM, New York, 1992.

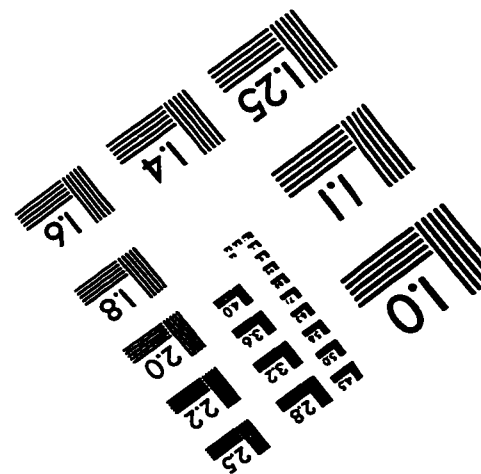
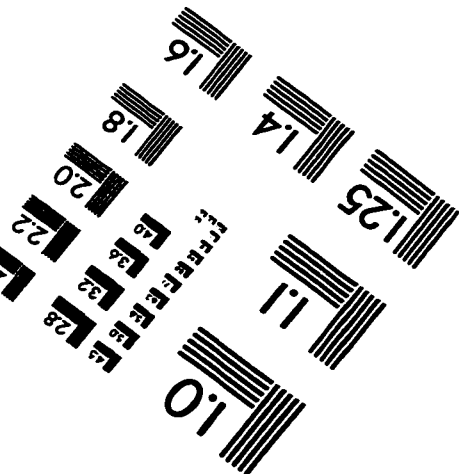
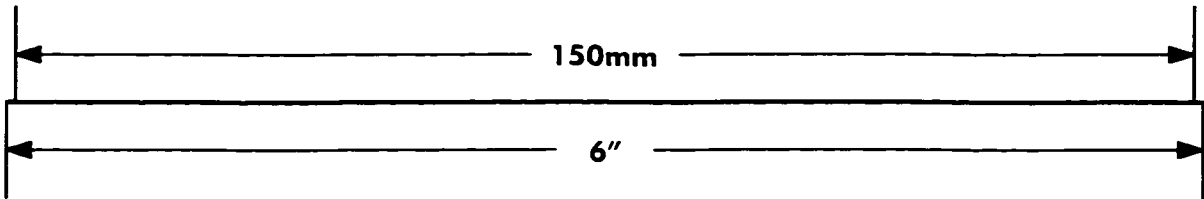
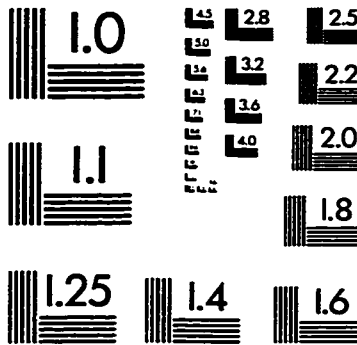
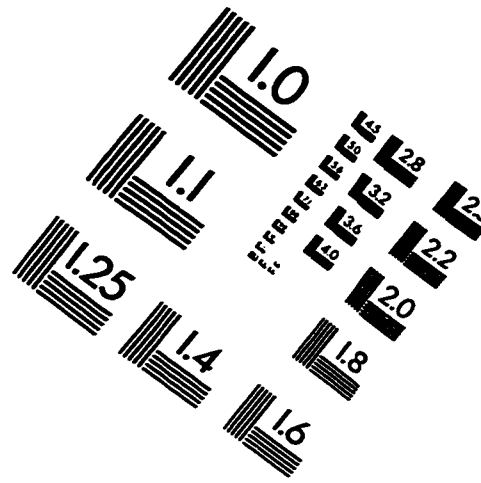
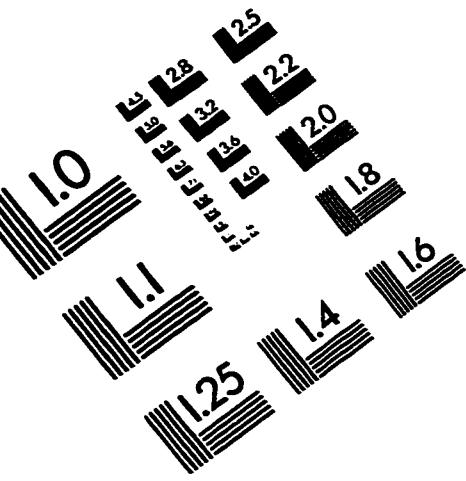
- [10] N. Duffield. Economies of scale for long-range dependent traffic in short buffers. *Telecommunication Systems*, to appear, 1998.
- [11] N. Duffield and N O'Connell. Large deviations and overflow probabilities for the general single-server queue, with applications. *Math. Proc. Cambr. Phil. Soc.*, 118:363–374, 1995.
- [12] A. Erramilli, O. Narayan, and W. Willinger. Experimental queueing analysis with long-range dependent traffic. *IEEE/ACM Transactions on Networking*, 4(2):209–223, April 1996.
- [13] Da-Yin Fan. The distribution of the product of independent beta variables. *Commun. Statist.-Theory Meth.*, 20(12):4043–4052, 1991.
- [14] P. Flandrin. Wavelet analysis and synthesis of fractional Brownian motion. *IEEE Trans. Inform. Theory*, 38(2):910–916, Mar. 1992.
- [15] G. Gripenberg and I. Norros. On the prediction of fractional Brownian motion. *J. Applied Probability*, 33:400–410, 1996.
- [16] M. Grossglauser and J-C. Bolot. On the relevance of long-range dependence in network traffic. *Computer-Communication-Review*, 26(4):15–24, October 1996.
- [17] Daniel P. Heyman and T. V. Lakshman. What are the implications of long-range dependence for VBR-video traffic engineering? *IEEE/ACM Transactions on Networking*, 4(3):301–317, June 1996.
- [18] C. Huang, M. Devetsikiotis, I. Lambadaris, and A. Kaye. Modeling and simulation of self-similar VBR compressed video: a unified approach. *Computer-Communication-Review*, 25(4):114–125, Oct. 1995.

- [19] Raj Jain. *The Art of Computer Systems Performance: Techniques for experimental design, measurement, simulation, and modeling*. John Wiley & Sons, Inc., 1991.
- [20] Predrag R. Jelenkovic and Aurel A. Lazar. The Effect of Multiple Time Scales and Subexponentiality in MPEG Video Streams on Queueing Behavior. *IEEE Journal on Selected Areas in Communications*, 15(6):1052–1071, August 1997.
- [21] N. Johnson, S. Kotz, and N. Balakrishnan. *Continuous Univariate Distributions*, volume 1-2. John Wiley & Sons, New York, 1994.
- [22] L.M. Kaplan and C.-C.J. Kuo. Fractal estimation from noisy data via discrete fractional Gaussian noise (DFGN) and the Haar basis. *IEEE Trans. Signal Proc.*, 41(12):3554–3562, Dec. 1993.
- [23] L.M. Kaplan and C.-C.J. Kuo. Extending self-similarity for fractional Brownian motion. *IEEE Trans. Signal Proc.*, 42(12):3526–3530, Dec. 1994.
- [24] W. Leland, M. Taqqu, W. Willinger, and D. Wilson. On the self-similar nature of Ethernet traffic (extended version). *IEEE/ACM Trans. Networking*, pages 1–15, 1994.
- [25] N. Likhanov, B. Tsybakov, and N. Georganas. Analysis of an ATM buffer with self-similar input traffic. *Proc. IEEE, Info com '95 (Boston 1995)*, pages 985–992, 1995.
- [26] D. V. Lindley. The theory of queues with a single server. *Proceedings of the Cambridge Philosophical Society*, 48:277–289, 1952.
- [27] S. Ma and C. Ji. Modeling video traffic in the wavelet domain. In *Proc. of 17th Annual IEEE Conf. on Comp. Comm., INFOCOM*, pages 201–208, Mar. 1998.

- [28] A. L. Neidhardt and J. L. Wang. The concept of relevant time scales and its application to queuing analysis of self-similar traffic. In *Proc. SIGMETRICS '98/PERFORMANCE '98*, pages 222–232, 1998.
- [29] I. Norros. On the use of fractional Brownian motion in the theory of connection-less networks. *COST*, 242, 1994.
- [30] I. Norros. Four approaches to the fractional Brownian storage. *Fractals in Engineering*, pages 154–169, 1997.
- [31] A. Papoulis. *Probability, Random Variables, and Stochastic Processes*. McGraw-Hill, New York, 1991.
- [32] Minothi Paulekar and Armand M. Makowski. Tail probabilities for a multiplexer with self-similar traffic. *Proc. IEEE INFOCOM*, pages 1452–1459, 1996.
- [33] V. Paxson and S. Floyd. Wide-area traffic: The failure of Poisson modeling. *IEEE/ACM Transactions on Networking*, 3:226–244, 1995.
- [34] B. Venkateshwara Rao, K. R. Krishnan, and Daniel P. Heyman. Performance of Finite-Buffer Queues under Traffic with Long-Range Dependence. *Proc. IEEE GLOBECOM*, 1:607–611, November 1996.
- [35] R. H. Riedi. Multifractal processes. *IEEE Info. Theory*, submitted 1999.
- [36] R. H. Riedi, M. S. Crouse, V. Ribeiro, and R. G. Baraniuk. A multifractal wavelet model with application to network traffic. *IEEE Trans. Info. Theory*, (*Special issue on multiscale statistical signal analysis and its applications*), 45(3):992–1018, April 1999. Available at www.dsp.rice.edu.
- [37] J. Roberts, U. Mocci, and J. Virtamo (eds.). Broadband network teletraffic. In *Lecture Notes in Computer Science, No 1155*. Springer, 1996.

- [38] B. K. Ryu and A. Elwalid. The Importance of Long-range Dependence of VBR Video Traffic in ATM Traffic Engineering: Myths and Realities. *Proc. ACM SIGCOMM Conf.*, 26(4):3–14, 1996.
- [39] M. Taqqu and J. Levy. *Using renewal processes to generate LRD and high variability*. In: Progress in probability and statistics, E. Eberlein and M. Taqqu eds., volume 11. Birkhaeuser, Boston, 1986. pp 73–89.
- [40] M. Taqqu, V. Teverovsky, and W. Willinger. Estimators for long-range dependence: An empirical study. *Fractals.*, 3:785–798, 1995.
- [41] A. Tewfik and M. Kim. Correlation structure of the discrete wavelet coefficients of fraction Brownian motion. *IEEE Trans. Inform. Theory*, 38(2):904–909, Mar. 1992.
- [42] K. E. Timmerman and R. D. Nowak. Multiscale Bayesian estimation of Poisson intensities. In *Proc. 31st Asilomar Conf.*, Pacific Grove, CA, Nov. 1997.
- [43] W. Willinger, M. Taqqu, R. Sherman, and D. Wilson. Self-similarity through high-variability: Statistical analysis of ethernet LAN traffic at the source level. *IEEE/ACM Trans. Networking (Extended Version)*, 5(1):71–86, Feb. 1997.
- [44] G. W. Wornell. A Karhunen-Loève like expansion for $1/f$ processes via wavelets. *IEEE Trans. Inform. Theory*, 36(2):859–861, Mar. 1990.

IMAGE EVALUATION TEST TARGET (QA-3)



APPLIED IMAGE, Inc.
1653 East Main Street
Rochester, NY 14609 USA
Phone: 716/482-0300
Fax: 716/288-5989

© 1993, Applied Image, Inc., All Rights Reserved