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.

ProQuest Information and Learning
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA
800-521-0600

MULTIRESOLUTION ANALYSIS FOR EFFECTIVE DYNAMIC BANDWIDTH MANAGEMENT

by

CLAUDE TURNER

A dissertation submitted to the Graduate Faculty in Engineering in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

The City University of New York

2003

UMI Number: 3074688

ProQuest Information and Learning Company
300 North Zeeb Road
P.O. Box 1346
Ann Arbor, MI 48106-1346

This manuscript has been read and accepted for the Graduate Faculty in Engineering in satisfaction of the dissertation requirement for the degree of Doctor of Philosophy.

10/16/02
Date

DEPUTY PROVOST & PROFESSOR JOSEPH BARBA.
Chair of Examining Committee

10/16/2002
Date

DEAN MUMTAZ KASSIR.
Executive Officer

PROFESSOR MITRA BASU

DOCTOR ROLSTON JEREMIAH
Lucent Technologies

PROFESSOR MYUNG LEE

PROFESSOR MARK AREND

Supervisory Committee

THE CITY UNIVERSITY OF NEW YORK

ABSTRACT

MULTIRESOLUTION ANALYSIS FOR EFFECTIVE DYNAMIC BANDWIDTH MANAGEMENT

by

Claude Turner

Advisor: Professor Joseph Barba

The aim of this dissertation is to develop effective dynamic bandwidth management algorithms to meet the multiple quality of service (QoS) requirements of the various applications utilizing a telecommunication network. In recent years, growing processing power and bandwidth have stimulated a new breed of applications, such as IP-based Private Branch Exchange, video on demand and video conferencing. These new applications require more innovative traffic management algorithms to meet multiple QoS requirements, and future applications promise to be even more challenging. One of the main obstacle to improving QoS is congestion at network nodes, which results in cell loss, delay and traffic fluctuation. It is believe that predicting the bandwidth at the output of each network node is the key to reducing such congestion and, thus, improving QoS. The proposed method is compatible with current technologies such as Voice Over IP (VoIP), Multi-protocol Label Switching (MPLS) and asynchronous transfer mode (ATM). The novelty of this approach hinges on its exploitation of the discrete wavelet transform (DWT) of the network traffic. Recent studies have shown that while network traffic consists of both long- and short- term dependence, its

wavelet coefficients are short term dependent only. Hence, significant advantages may be gained by employing legacy traffic models in the wavelet domain. The DWT decomposes the traffic into a low-frequency component and its high frequency components--the details. The low-frequency component of the traffic, are significant for long-term traffic behavior and, hence, for bandwidth allocation. The high-frequency components are significant in the short-term behavior of the traffic and, hence, for buffer-allocation.

ACKNOWLEDGEMENTS

I would like to thank my advisor, Professor and Deputy Provost Joseph Barba, for his wise guidance and tremendous support over the past few years. Many thanks also to my committee members for their helpful advice. Special thanks to Dr. Rolston M. Jeremiah of Lucent Technologies who contributed generously with his time and offered many helpful ideas throughout the years. It was Dr. Jeremiah who introduced me to the field of wavelets as well as to many of its applications to signal processing. Special thanks also to Dr. Rulei Ting, formerly of Lucent Technologies, who provided very valuable advice on the application of wavelet theory to network traffic bandwidth management.

Finally, I would like to extend my deepest appreciation to my family, relatives and friends for their long-time encouragement and support—especially my son, Ian, who has helped me tremendously in maintaining my sanity with his humor and love.

This dissertation is dedicated to my beloved parents, Sarah and Hewlett Turner who have given so much and have asked so little. I love you both!

My work would not have being possible without the generous support of the following:

- The James Bruce LLewellyn Fellowship.
- The CUNY Writing Fellows Program.
- The NYC Louis Stokes Alliance.

Contents

List of Tables

List of Figures

NOTATIONS

- $\mathbb{N}$ - the set of nonnegative integers $\{0, 1, 2 \ldots\}$.
- $\mathbb{N}^*$ - the set of positive integers $\{1, 2, 3 \ldots\}$.
- $\mathbb{Z}$ - the set of integers $\{\ldots, -1, 0, 1, \ldots\}$.
- $\mathbb{Z}^* = \mathbb{Z} - \{0\}$ - the set of nonzero integers.
- $\mathbb{R}$ - the set of real numbers.
- $\mathbb{R}^*$ - the set of nonzero real numbers.
- x - the traffic sequence $(x_m)_{m \in \mathbb{N}}$.
- $\Lambda_{m,i}$ - the set $\{m, m+1, \cdots, i\}$, $i > m$, i, $m \in \mathbb{N}$.
- Λ_m - the set $\{1, \cdots, m\}$, $m \in \mathbb{N}$.
- B - the bandwidth formula.
- ρ - the prediction interval.
- $D_j x$ - the detail signal of x at level j.
- $A_j x$ - the approximation signal of x at level j.
- n - the traffic length.

Chapter 1

Introduction

1.1 Overview

Congestion at a communication network node plays a major role in the degradation of QoS in the network. Typically, several traffic sources arrive at a network node, and all or a subset of these must utilize one of several different out-going links. (Figure 1.1 shows the simple case of a single output link with several input sources.) The problem is to determine the bandwidth requirement on an out-going link so as to reduce congestion and to, in general, improve *quality of service* (QoS) in the network. A variety of methods have been proposed to deal with this problem, but they all have limitations. The Effective Bandwidth Method [1], which has been one of the more popular of these and is based on assigning a fixed bandwidth to each source, tends to over- and under- estimate the bandwidth at various times. Thus, in recent years, there have been an emergence of a variety of alternative bandwidth management methods [2] [3] [4]. One approach that seems

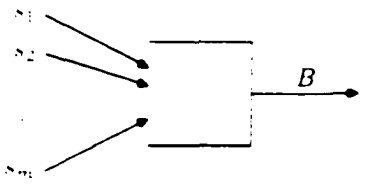


Figure 1.1: Network Node

to hold great promise, is that of modelling the traffic and/or the traffic bandwidth in the wavelet domain [2] [4] [5] [6]. This new approach is driven by recent research showing network traffic to be self-similar (or long range dependent (LRD)) [7] [8] [9], and is thus not adequately modelled by the short-range dependent (SRD) models which have been the mainstay for legacy traffic models. However, research also show that the *discrete wavelet transform* (DWT) – a tool that allows for the analysis of a signal at different resolutions – de-correlates network traffic so that at each scale, wavelet coefficients of the traffic is short-range dependent (SRD) [10] [11]. The first of these findings is so significant that it has prompted one of the world's giant telecommunication companies, Nortel, to begin funding the Nortel Chaotic Telecoms Project (see *http://gridlock.york.ac.uk/nndg/telecoms.html*). The goal of this project is to investigate prediction methods for network behavior so as to achieve better control. The second finding is also significant. Because of the SRD nature of wavelet coefficients, they lend themselves to modelling with legacy traffic models.

There are several approaches to solving the bandwidth allocation problem in the wavelet domain. One of these is to calculate the bandwidth base on the notion that the traffic's future high frequency traffic components are similar to high frequency components of its present and recent past. A second method is based on the idea of first generating syn-

thetic traffic in the wavelet domain ahead of time and then estimating the corresponding bandwidth from this synthetic traffic. The synthetic traffic, in this case, must of course embody similar characteristics to that of the network traffic, that is, the actual network traffic with a synthetic model. A third method involves predicting the next ρ values of the actual network traffic ahead of time in the wavelet domain and then performing some sort of transformation on these values to obtain the bandwidth. Each of these methods share the following in common: (1) a need to have a training set of data—real or synthetic—with similar characteristics to the actual network traffic, (2) the need for some sort of bandwidth model or transformation based on the traffic available in (1) to model the traffic bandwidth, and (3) the need to separate (or decompose) the traffic into its low and high frequency components by way of the discrete wavelet transform. The high frequency components of the network traffic characterize the short range dependency in the traffic, while the low frequency components characterize the traffic's long range dependency.

1.2 Research Objectives

The objectives of this dissertation are as follows:

- To develop a class of bandwidth models based on the discrete wavelet transform to model the traffic bandwidth at a network node output;
- To develop mathematically-proven approaches to estimate (or predict) wavelet coefficients needed by each bandwidth formula;

- To develop a generalized dynamic bandwidth allocation scheme based on the bandwidth models and prediction algorithms developed here;
- To analyze the above algorithms using a simulated ATM switch and determine quality of service parameters to test the performance of each algorithm;
- To show that the above approach leads to improvement in two critical QoS parameters– cell loss ratio and cell delay.

1.3 Methodology

In this section, we discuss in brief our overall strategy for the construction of the integrated dynamic bandwidth management model mentioned in the third objective of the previous section. Assume the traffic to be the number of packet arrivals x_i during the i-th time slot of duration $T_0 \in \mathbb{R}^+$ of the measurement. Let $I_n = (1, 2, \cdots, n)$, for some $n \in \mathbb{N}^*$, then the traffic sequence may be written as $x = (x_i)_{i \in I_n}$. The goal of this integrated approach, illustrated in Figure 1.2, is given the λ past values of the traffic $x_{k-\lambda-\rho+1}, x_{k-\lambda-\rho+2}, \cdots, x_{k-\rho}$, predict (or estimate) the future bandwidth requirement $B_{k-\rho+1}, B_{k-\rho+2}, \cdots, B_k$ for the next ρ traffic samples $x_{k-\rho+1}, x_{k-\rho+2}, \cdots, x_k$ to arrive. It proceeds as follows. As the traf-

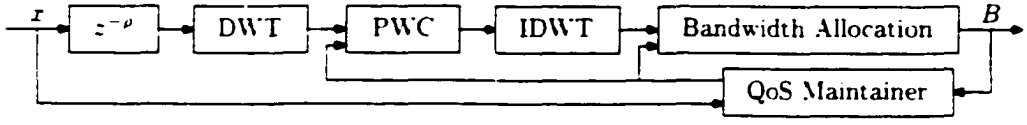


Figure 1.2: Bandwidth Allocation Model

fic arrives, a copy of each x_i is stored into a buffer, which uses a first-in-first-out (FIFO)

discipline (see Figure 1.1). Once λ traffic samples have accumulated into this buffer, their discrete wavelet transform is computed, and the wavelet coefficients obtained are used in a prediction algorithm to predict wavelet coefficients for future traffic $x_{k-\rho+1}, x_{k-\rho+2}, \cdots, x_k$. These predicted coefficients are in turn used to calculate the bandwidth requirement. Note that a certain amount of computational time will be associated with the prediction algorithm. Let this time c be measured in terms of the number of samples that arrive during the computation. Thus, the value of the bandwidth that exists before the computation begins is maintained until at least the arrival of sample $x_{\lambda+c}$; whereupon, it is updated to the newly predicted bandwidth. Let $k = \lambda$. When traffic sample $x_{k+\rho}$ arrives and its value stored, the bandwidth is again updated using the λ most recently stored buffer values, $x_{k-1}, x_{k-2}, \cdots, x_{k-\ell}$, in the prediction algorithm to predict the bandwidth for the next ρ traffic samples to arrive. The bandwidth continues to update in this way until the n^{th} traffic sample arrives. For the sake of convenience, in the sequel, λ, ρ and traffic length n will be assumed to be powers of two.

An idealized pictorial example of the algorithm is given in Figure 1.3 (b) for $\lambda = 32$ and $\rho = 8$. In this example the initial buffer allocation occurs between 1 and 40 and the traffic values $x_1, x_2, \cdots, x_{32}$ are used to predict the bandwidth for x_{41} through x_{48}. In the next update $x_9, \cdots, x_{40}$ are used to predict the bandwidth for $x_{49}, \cdots, x_{56}$. Other updates continue in a similar manner.

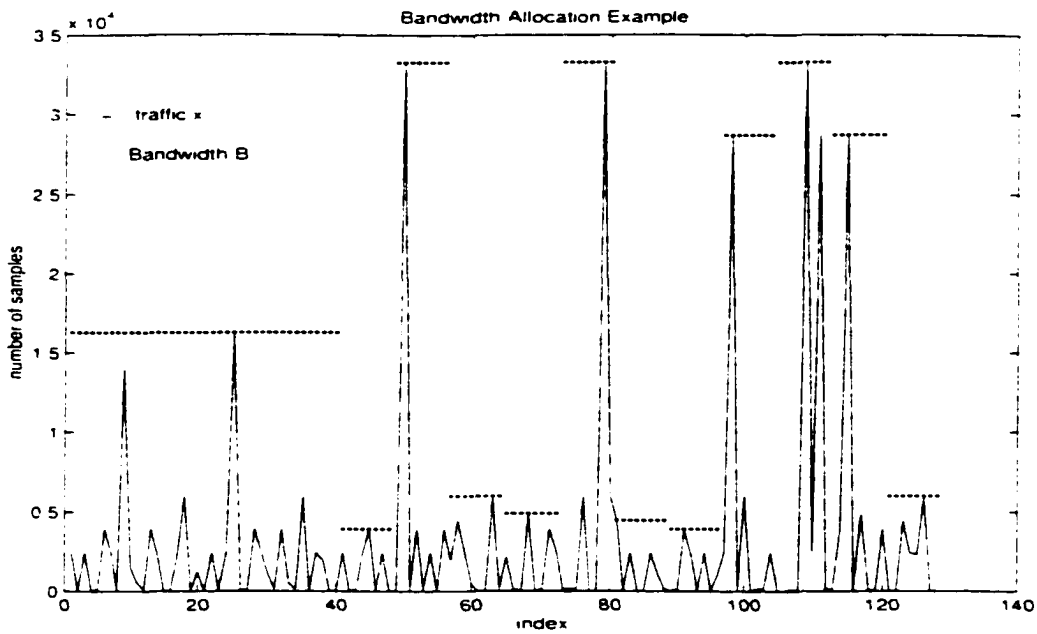


Figure 1.3: Bandwidth Allocation Algorithm

1.4 Dissertation Outline

This dissertation consists of seven chapters. Chapter One provides an introduction, states the research objectives and describe our general method of approach to proposed problem. Chapter Two introduces the necessary mathematical background. Self-similarity, long range dependence and short range dependence are defined and examples are also given. This Chapter also introduces the wavelet transform and multiresolution analysis.

Chapter Three covers the basic and related concepts of ATM Technologies, Quality of Services, Traffic Characterization and resource management in Multimedia Communication. The necessity of traffic prediction in determining bandwidth and resource allocation, the impact of low-frequency components on long term behavior (bandwidth allocation) and

the impact of high frequency components on the short term behavior (buffer allocation) are also discussed here.

Chapter Four introduces bandwidth models.

Chapter Five proposes two bandwidth prediction algorithms. The first is based on the Independent Wavelet Model. The second is a regression-based model.

Chapter Six proposes bandwidth allocation algorithms based on the bandwidth formulas and prediction algorithms of the two previous chapters. Some of the traffic traces use in our experiments, the experimental results for the bandwidth allocation algorithms and their discussions are also provided here. Chapter Seven concludes dissertation and describe future work.

Chapter 2

Mathematical Preliminaries

2.1 Definitions

Definition 2.1 (integrable functions) *The space of measurable integrable functions* $L^1(\mathbb{R})$ *is given by*

$$L^1(\mathbb{R}) = \{f : \|f\| = \int_{\mathbb{R}} |f| < \infty\}. \tag{2.1}$$

Definition 2.2 (square-integrable functions) *The space of measurable square integrable functions* $L^2(\mathbb{R})$ *(the Hilbert Space) is given by*

$$L^2(\mathbb{R}) = \{f : \|f\|^2 = \int_{\mathbb{R}} |f|^2 < \infty\}. \tag{2.2}$$

Definition 2.3 (inner product) *Let* $f, g \in L^2(\mathbb{R})$*, the* $L^2(\mathbb{R})$ *inner product of* f *with* g *is written*

$$\langle f, g \rangle = \int_{\mathbb{R}} fg \tag{2.3}$$

Definition 2.4 (The ℓ^p Spaces) *Let $1 \leq p \leq +\infty$. Then*

$$\ell^p = \begin{cases} \{x \in \Xi | \sum_{n\in\mathbb{Z}} |x(n)|^p < +\infty\} & \text{if } 1 \leq p \leq +\infty. \\ \{x \in \Xi | sup_{n\in\mathbb{Z}} |x(n)| < +\infty & \text{if } p = +\infty. \end{cases} \tag{2.4}$$

Definition 2.5 (norm on ℓ^p) *The ℓ^p norm is defined as follows*

$$(\forall x \in \ell^p) \quad \|x\| = \begin{cases} \left(\sum_{n\in\mathbb{Z}} |x(n)|^p\right)^{1/p} & \text{if } 1 \leq p \leq +\infty \\ \sup_{n\in\mathbb{Z}} |x(n)| & \text{if } p = +\infty. \end{cases} \tag{2.5}$$

Note that if $1 \leq p \leq q \leq +\infty$ then $\ell^p \subset \ell^q$ and that the inclusion is proper if $p < q$.

Definition 2.6 (Hölder's Inequality) *Hölder inequality for $1 \leq p \leq +\infty$ states*

$$(\forall x \in \ell^p)(\forall y \in \ell^{p/(p-1)}) \quad \left|\sum_{n\in\mathbb{Z}} x(n)y(n)\right| \leq \sum_{n\in\mathbb{Z}} |x(n)y(n)| \leq \|x\|_p \|y\|_{p/(p-1)}. \tag{2.6}$$

Definition 2.7 (Finite Energy Signals ℓ^2) *The special case $p = 2$ gives the important space ℓ^2 of finite energy signals with scalar product*

$$(\forall x \in \ell^2)(\forall y \in \ell^2) \quad \langle x, y\rangle = \sum_{n\in\mathbb{Z}} x(n)y(n) \tag{2.7}$$

and norm (2.8)

$$\|x\|_2 \triangleq \|x\| = \sqrt{\langle x, x\rangle} = \sqrt{\sum_{n\in\mathbb{Z}} |x(n)|^2} = \sqrt{E_x}. \tag{2.9}$$

Definition 2.8 (Cauchy-Schwartz Inequality) *The Cauchy-Schwartz inequality is defined as follows*

$$(\forall x \in \ell^2)(\forall y \in \ell^2) \quad \langle x, y\rangle \leq \|x\| \cdot \|y\|. \tag{2.10}$$

Definition 2.9 (finite-dimensional distributions) *The n^{th}-order distribution of a stochastic process x is the joint distribution:*

$$F(x_1, x_2, \ldots, x_n; t_1, t_2, \ldots, t_n) = P\{x(t_1) \leq x_1, x(t_2) \leq x_2, \ldots, x(t_n) \leq x_n\}. \tag{2.11}$$

Its corresponding n^{th}-order density function is given by

$$f(x_1, x_2, \ldots, x_n) = \frac{\partial^n F(x_1, x_2, \ldots, x_n)}{\partial x_1 \partial x_2 \cdots \partial x_n}. \tag{2.12}$$

Definition 2.10 (covariance) *The covariance of random variables x and y with mean $\mu_x = E[x]$ and $\mu_y = E[y]$, respectively, is given by*

$$C_{xy} = E[(x - \mu_x)(y - \mu_y)] = E[xy] - E[x]E[y]. \tag{2.13}$$

Definition 2.11 (correlation coefficient) *The correlation coefficient of random variables x and y is defined by*

$$r = \frac{C_{xy}}{\sigma_x \sigma_y} \tag{2.14}$$

where C_{xy} is the covariance of RVs x and y, and σ_x and σ_y are the standard deviation of x and y, respectively.

Definition 2.12 (autocorrelation) *The autocorrelation of a stochastic process x is the covariance of the product $x(t_1)x(t_2)$ of the RVs $x(t_1)$ and $x(t_2)$ and is given by:*

$$R(t_1, t_2) = E[x(t_1)x(t_2)] = \int_{-\infty}^{\infty} x_1 x_2 f(x_1, x_2; t_1, t_2). \tag{2.15}$$

Definition 2.13 (autocovariance) *The autocovariance of a stochastic process x is the covariance of the RVs $x(t_1)$ and $x(t_2)$:*

$$\gamma(t_1, t_2) = E[(x(t_1) - \mu(t_1))(x(t_2) - \mu(t_2))] = R(t_1, t_2) - \mu(t_1)\mu(t_2). \tag{2.16}$$

Definition 2.14 (strict stationarity) *A process $x(t)$ is strict-sense stationary (or wide-sense stationary or covariance stationary) if $(x(t_1), x(t_2), \ldots, x(t_n))$ and its k-shifted version $x_k = (x(t_1 + k), x(t_2 + k), \ldots, x(t_n + k))$ possess the same joint distribution for all $n \in \mathbb{Z}^+, t_1, t_2, \ldots, t_n, k \in \mathbb{Z}$.*

Definition 2.15 (Second-Order Stationarity) *A process $x(t)$ is second-order stationary if its autocovariance function $\gamma(r, s) = E[(x(r) - \mu)(x(s) - \mu)]$ satisfies translation invariance, i.e., $\gamma(r, s) = \gamma(r + k, s + k)$ for all $r, s, k \in \mathbb{Z}$.*

2.2 Wavelets, Multiresolution Analysis & the DWT

The wavelet transform, which was popularized by Grossman and Morlet [12], has been utilized extensively in the signal processing community [13][14]. This is largely due to the work of Mallat [15], who developed an elegant methodology to link subband decomposition with multiresolution analysis and wavelets, and Daubechies [16], who developed a class of smooth wavelets. In recent years, a significant amount of research has been done on the application of results of wavelets and multiresolution approximation theory to the telecommunication networking area and, in particular, in the area of traffic modelling and analysis [2][10][11]. The versatility of wavelets, in particular, its ability to simultaneously localize in the spatial and frequency domains, makes it better suited for signal representation than many of the classical methods.

A mathematical result, which is quite important in signal processing, states that any

continuous function (signal) can be represented by more elementary functions. Formally, given a signal x in a Hilbert space $L^2(\Re)$ [17] [18], there exists a basis $\{o_k\} \subset L^2(\Re)$ and coefficients $\{c_k \in \mathbb{C}\}$ such that

$$x = \sum_{k \in \mathbb{Z}} c_k o_k. \tag{2.17}$$

The coefficients $\{c_k\}$ are given by the scalar product

$$(\forall k \in \mathbb{Z}) \quad c_k = \langle o_k, x \rangle = \int_R \hat{o}_k x \tag{2.18}$$

where $\{\hat{o}_k\}$ is the dual of $\{o_k\}$. When a signal is represented in this way, it is possible to determine which of the coefficients contribute significantly to its representation. Thus, it is sometimes possible to represent X with fewer and in many cases finitely many coefficients.

2.2.1 Classical Transforms

Several transforms are exist to represent a signal as described by (2.17). However, they all have significant drawbacks and are unsuitable for some applications. The Karhunen-Loève (K-L) transform [19] is a specialized time series representation where the basis functions $\{o_k\}$ yield a set of uncorrelated coefficients $\{c_k\}$. Nevertheless, it has two major disadvantages. It is signal dependent. Further, the K-L transform is computationally complex since there is no fast algorithm available when $\{c_k, \phi_k\}$ is arbitrary.

The Fourier transform (FT) [20] decomposes a signal into sinusoids at various frequencies. That is, it transforms the signal from the spatial to the frequency domain. Unfortunately, the spatial information is lost during such a transformation. If the signal is non-stationary,

its transitory characteristics such as drifts, trends, abrupt changes, the beginning and end of events, are not detected.

The Short-Time Fourier transform (STFT) is an extension of the FT. It enables localization in the spatial domain by applying the FT over windowed signal pieces. However, the window size remain fixed irrespective of the frequencies being analyzed.

The Discrete Cosine Transform (DCT) is often used to approximate the K-L transform [20]. It is used extensively in standards for speech, image and video compression. However, the signal must be partitioned into blocks before the DCT can be applied. Consequently, the correlation across boundaries are not removed and blocking effects result. As an alternative, subband or wavelet schemes are employed.

2.2.2 The Continuous Wavelet Transform

A wavelet is any function $\psi \in L^1(\Re)$ satisfying the following properties:

$$0 = \int_{\Re} \psi(t)\, dt \tag{2.19}$$

$$C_\psi = \int_{\Re} \frac{|\widehat{\psi}(\nu)|^2}{|\nu|} d\nu < \infty. \tag{2.20}$$

Equation (2.19) implies that the wavelet is a bandpass or oscillating function. Indeed, the wavelet function is usually chosen as a small wave defined on a support, which is almost limited in time and having most of its energy within a limited frequency band, i.e., its spread in time and frequency are relatively limited. However, both the time support and frequency band cannot both be finite, and there is an interval on which they are effectively

limited.

The continuous wavelet transform consists of the collection of coefficients

$$\{(W_\psi x)(a,b) = \langle x, \psi_{a,b}\rangle = \int_{\Re} x(t)\psi_{a,b}(t)dt \ : \ a \in \Re^+, b \in \Re\} \tag{2.21}$$

where the dilation parameter a defines the scale of time (or, equivalently, the range of frequencies) over which the signal is to be analyzed, the shift parameter b defines the time instant around which the signal x is to be analyzed and $\{\psi_{a,b}\}$ is the set of analyzing functions given by

$$\{\psi_{a,b}(t) = a^{-1/2}\psi\left(\frac{t-b}{a}\right) \ : \ a \in \Re^+, b \in \Re\}. \tag{2.22}$$

The reciprocal $1/a$ of the scaling parameter is called the *resolution*. An increase in the scale (decrease in resolution) corresponds to a stretching of the wavelet function ψ, while a decrease in scale corresponds to a compression of ψ.

The inverse continuous wavelet transform is given by

$$(\forall t \in \Re) \quad x(t) = C_\psi^{-1} \int\int \frac{da\ db}{a^2} \langle x, \psi_{a,b}\rangle \psi_{a,b}(t). \tag{2.23}$$

2.2.3 The Discrete Wavelet Transform

A discrete wavelet transform can be obtained by critically sampling the time-scale plane, i.e., to keep among the $\{(W_\psi x)(a,b) \ : \ a \in \Re^+, b \in \Re\}$, only a discrete set of coefficients while still retaining the information in x. In this case, the family of wavelets of interest

then becomes, for $j,\ k \in \mathbb{Z}$,

$$\psi_{j,k}(t) = a_0^{-j/2}\psi(a_0^j t - kb_0) \tag{2.24}$$

Thus, $a = a_0^j$ and $b = kb_0a_0^j$. A *discrete wavelet transform (DWT)* can now be defined as

$$\{(Tx)_{jk} = \langle \psi_{jk}, x\rangle = a_0^{-j/2}\int_{\mathbb{R}} \psi(a_0^{-j}x - kb_0)x(t)dt \ :\ (j,k) \in \mathbb{Z}^2\} \tag{2.25}$$

Indeed, it suffices to choose dyadic values $a = 2^j$ and $b = k2^j$ (i.e., $a_0 = 2,\ b_0 = 1$)for the scaling and translation parameters, respectively.

2.2.4 Multiresolution Analysis

The theory of multiresolution analysis makes it possible to derive a wavelet ψ such that

$$(\forall x \in \mathbb{R}) \quad \psi_{j,k}(x) = 2^{-j/2}\psi(2^{-j}x - k) \quad j,k \in \mathbb{Z} \tag{2.26}$$

is a basis of $L^2(\mathbb{R})$. Let V_j be the subspace of $L^2(\mathbb{R})$ spanned by $(\phi_{j,k})_{k\in\mathbb{Z}}$, i.e. $V_j = span\{\phi_{j,k}\}_{k\in\mathbb{Z}}$ where $\phi_{j,k}(t) = 2^{-j/2}\phi(2^{-j}t - k)$ and $\phi \in V_0$ is the so-called *scaling function*. A sequence of nested subspaces $\{V_j\}_{j\in\mathbb{Z}}$ constitute a *multiresolution analysis* for $L^2(\mathbb{R})$ if the following conditions are met:

1. $\cdots \subset V_1 \subset V_0 \subset V_{-1} \subset \cdots$
2. $L^2(\mathbb{R}) = \overline{\bigcup_{j\in\mathbb{Z}} V_j}$ and $\bigcap_{j\in\mathbb{Z}} V_j = \{0\}$
3. $\exists\phi \in V_0$ such that $\{\phi_{0,n}\}$ is an unconditional Riesz basis for V_0.
4. $(\forall j \in \mathbb{Z}) \quad x \in V_j \iff x(2\cdot) \in V_{j-1}$

What is the significance of these conditions? Property 1. shows that the spaces $\{V_j\}$ are nested. Property 2. $L^2(\mathbb{R}) = \overline{\bigcup_{j\in\mathbb{Z}} V_j}$. means that the V_j's are *approximation* subspaces for $L^2(\mathbb{R})$. Property 3 expresses the fact that the set of shifted scaling functions $\{\phi_{0,k} : k \in \mathbb{Z}\}$ forms a Riesz basis for V_0. meaning that they are linearly independent and span the space V_0. but they are not necessarily orthogonal nor do have to be of unit length. Properties 3 & 4 together imply that the set $\{\phi_{j,k} : k \in \mathbb{Z}\}$ is an unconditional Riesz basis for V_j. Thus. multiresolution analysis involves successively projecting the signal x into each of the approximation subspaces V_j:

$$A_j(t) = (P_{V_j}x)(t) = \sum_{k\in\mathbb{Z}} a_x(j,k)\phi_{j,k}(t) \tag{2.27}$$

where $P_{V_j}x$ denotes the orthogonal projection of x onto the space V_j. From Property 4. $V_j \subset V_{j-1}$. thus A_j is a coarser approximation of x than is A_{j-1}. Moreover. Property 2 implies that all information is removed from the signal. as $j \to \infty$. MRA is. therefore. a study of a signal x by examining coarser and coarser approximations of it as the details or high frequencies of the data are extracted.

The information that is removed when going from one approximation to the next is called the detail:

$$D_j(t) = A_{j-1}(t) - A_j(t). \tag{2.28}$$

from which it follows that $A_j = A_L + \sum_{m=j+1}^{L} D_m$. $j = 1,\ldots,L$. Results from MRA theory guarantee that the detail signal can be obtained by projecting the signal x onto a collection of subspaces. $\{W_j = V_{j-1} - V_j\}$. called the wavelet subspaces. Moreover. there

exists a function ψ called the mother wavelet, which can be derived from ϕ, such that its templates $\{\psi_{j,k} : k \in \mathbb{Z}\}$ constitutes a basis for W_j :

$$D_j(t) = (P_{W_j} x)(t) = \sum_{k \in \mathbb{Z}} d_x(j,k)\psi_{jk}(t). \tag{2.29}$$

In theory, the projection procedure can be performed from $j \to -\infty$ up to $j \to +\infty$. In practice, however, the range of j is limited to $j = 0, 1, \ldots, L$. Thus, it's only necessary to consider

$$V_L \subset V_{L-1} \subset \cdots \subset V_0. \tag{2.30}$$

This means that the analysis of x is now restricted to that of its orthogonal projection $A_0(t)$ onto the reference space V_0. The fine scale approximation $A_0(t)$ can now be written as a summation of details at different resolutions together with a final resolution that belongs to V_L.

$$A_0(t) = A_L(t) + \sum_{j=1}^{L} D_j \tag{2.31}$$

$$= \sum_{k} a_x(L,k)\phi_{L,k}(t) + \sum_{j=1}^{L}\sum_{k \in \mathbb{Z}} d_x(j,k)\psi_{j,k}(t) \tag{2.32}$$

If the signal x belongs to the set V_0, then one can replace A_0 by x in (2.31). In this case, Equation 2.31 is refered to as the *Inverse Discrete Wavelet Transform (IDWT)*. Otherwise, unavoidable information loss will occur for the initial projection[21]. Subsequent projections, however, will not incur any further loss. Varying L simply means deciding if more or less information is written in the details as opposed to A_L, the final approximation.

Since the A_js are coarser and coarser approximation of x, ϕ needs to be a lowpass function. The D_js constitute an information differential, therefore, ψ is a bandpass function—a small

wave (a *wavelet*).

Thus, given a scaling function ϕ and a mother wavelet ψ, the *Discrete (or non-redundant) Wavelet Transform* would then consists of the collection of coefficients:

$$x \rightarrow \{\{a_x(L,k) : k \in \mathbb{Z}\}, \{d_x(j,k) \; : \; j = 1,2,\ldots,L, \; k \in \mathbb{Z}\}\}. \tag{2.33}$$

These coefficients are defined through inner products with two sets of functions:

$$a_x(j,k) = \langle x, \phi_{j,k} \rangle \tag{2.34}$$

$$d_x(j,k) = \langle x, \psi_{j,k} \rangle \tag{2.35}$$

where $\psi_{j,k}$ (resp., $\phi_{j,k}$) are shifted and dilated templates of ψ (resp. ϕ). Given coefficients $a_x(j,\cdot)$ and $d_x(j,\cdot)$ at resolution j, it is possible to obtain coefficients at the finer level $j-1$ through the use of a so-called dual mother wavelet $\tilde{\psi}$ (resp. the dual scaling function $\tilde{\phi}$). The definition of the dual is dependent on whether one chooses to use orthogonal, semiorthogonal or biorthogonal DWT wavelets [16] as will be seen later. The role of the wavelet and its dual can arbitrarily be switched; as is the role of the scaling function and its dual. The $d_x(\cdot,\cdot)$ constitute a subsample $\{(W_{\tilde{\psi}}x)(a,b) \; : \; a \in \mathbb{R}^+, b \in \mathbb{R}\}$ located on the so-called dyadic grid, $d_x(j,k) = (W_{\tilde{\psi}}x)(2^j, 2^j k)$. The logarithm of the scale $\log_2(a = 2^j) = j$ is called the *octave*. In many cases, a scale will often be referred to by its corresponding octave.

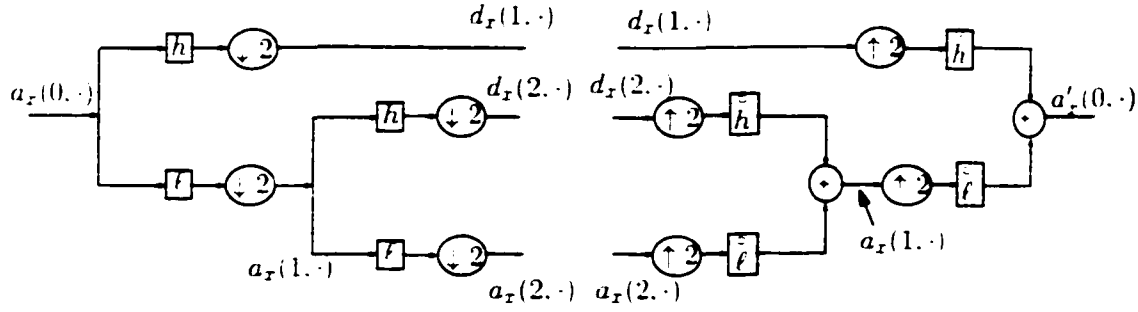


Figure 2.1: Decomposition (left half) and Reconstruction (right half) Algorithms

2.2.5 The Fast Pyramidal Algorithm

This subsection introduces the fast pyramidal algorithm of S. Mallat [15] for the computation of the discrete wavelet transform.

Due to the nested structure of multiresolution analysis, it is possible to obtain the approximation (resp. detail) coefficients $a_x(j-1,\cdot)$ (resp. $d_x(j-1,\cdot)$ from a convolution of the approximation coefficients one octave lower $a_x(j,\cdot)$ with a filter ℓ (resp. h) to be derived from ϕ (resp. ψ) [15] as follows:

$$a_x(j-1,k) = ((a_x(j,\cdot) * \ell) \downarrow 2)(k) = \sum_m \ell_{2k-m} a_x(j,m) \tag{2.36}$$

$$d_x(j-1,k) = ((a_x(j,\cdot) * h) \downarrow 2)(k) = \sum_m h_{2k-m} d_x(j,m). \tag{2.37}$$

This algorithm is quite efficient and is of order n, the length of x, which is significantly less than the order n^2 for the DFT.

The *synthesis*, where the fine approximation at level j is obtained from the coarse approx-

imation and detail coefficients at level $j-1$, is obtained via the converse operation:

$$a_x(j,k) = \left((a_x(j+1,\cdot) \uparrow 2) * \tilde{\ell}\right)(k) + \left((d_x(j+1,\cdot) \uparrow 2) * \tilde{h}\right)(k) \tag{2.38}$$

$$= \sum_m \tilde{\ell}_{2m-k} a_x(j,m) + \sum_m \tilde{h}_{2m-k-1} d_x(j,m). \tag{2.39}$$

Thus, we see that the determination of the filter coefficients ℓ and h and their duals $\tilde{\ell}$ and $\tilde{h}$ is of central importance in computing wavelet coefficients. The next two subsections show how to compute these coefficients in the case of orthogonal and biorthogonal wavelets, respectively.

2.2.6 Orthogonal Wavelet Bases

This section discusses how to obtain the filter coefficients for the fast pyramidal algorithm in the case of orthonormal wavelet bases. Let $\varphi \in V_0$ such that $(\varphi_{0k}(t) = \varphi(t-k))_{k\in\mathbb{Z}}$ constitute an orthonormal basis for V_0 and let $a_x(j,k) = \langle x, \varphi_{jk}\rangle$. From the multiresolution analysis conditions, $(\varphi_{jk}(t))_{k\in\mathbb{Z}} = (2^{-j/2}\varphi(2^{-j}t-k))_{k\in\mathbb{Z}}$ then constitute an orthonormal basis for V_j with projection onto V_j given by

$$A_j(t) = \left(P_{V_j} x\right)(t) = \sum_{k\in\mathbb{Z}} a_x(j,k)\varphi_{jk}(t). \tag{2.40}$$

Since $\varphi \in V_0 \subset V_{-1} = \overline{Span\{\varphi_{-1k} : k \in \mathbb{Z}\}}$, then the function φ necessarily satisfies an equation of type

$$\varphi(t) = \sum_{k\in\mathbb{Z}} h_k \varphi(2t-k), \tag{2.41}$$

where $c_k = \langle \varphi, \varphi_{-1k}$. By the orthogonality of $(\varphi_{0k})_{k\in\mathbb{Z}}$, we have

$$\sum_{n\in\mathbb{Z}} h_n h_{n-2k} = 2\delta_{k0}. \tag{2.42}$$

The associated orthonormal wavelet basis is given by

$$\psi(t) = \sum_{k\in\mathbb{Z}} (-1)^k h_{-k+1}\phi(2t-k). \tag{2.43}$$

where (h_k) is given by (2.41). It follows that $(\psi_{jk})_{(j,k)\in\mathbb{Z}^2}$ constitute an orthonormal basis for $L^2(\mathbb{R})$. Moreover, for every fixed j, $(d_x(j,k) = \langle x, \psi_{j,k}\rangle)_{k\in\mathbb{Z}}$ express the difference between the approximations of x with resolutions 2^j and 2^{j-1}.

$$P_{V_{j-1}}x = P_{V_j}x + \sum_{k\in\mathbb{Z}} d_x(j,k)\psi_{jk}. \tag{2.44}$$

The approximation and wavelet coefficients are obtained by taking the inner product of x with (2.41) and (2.43), respectively, to get

$$a_x(j,k) = \sum_n \ell_{n-2k} a_x(j-1,n) = \left([\ell * a_x(j-1,\cdot)]_{\downarrow 2}\right)(k) \tag{2.45}$$

$$d_x(j,k) = \sum_n h_{n-2k} a_x(j-1,n) = \left([h * a_x(j-1,\cdot)]_{\downarrow 2}\right)(k) \tag{2.46}$$

where $h_m = (-1)^m \ell_{-m+1}$. Equations (2.45) and shows how to obtain a coarser approximation from a finer one, as well as the difference in information between the two successive approximations, respectively. Recovering the finer approximation from the coarser one together with the difference in information is obtained by using Equations (2.41)(2.43) and (2.44) to get

$$a_x(j-1,m) = \sum_{k\in\mathbb{Z}} [a_x(j,k)\langle \phi_{jk}, \phi_{j-1,m}\rangle + d_x(j,k)\langle \psi_{jk}, \phi_{j-1,m}\rangle] \tag{2.47}$$

$$= \sum_k [\ell_{m-2k} a_x(j,k) + h_{m-2k} d_x(j,k)] \tag{2.48}$$

$$= (\ell * [a_x(j,\cdot)_{\downarrow 2}]) + (h * [d_x(j,\cdot)_{\downarrow 2}])(m). \tag{2.49}$$

Thus, the coefficients for the *reconstruction filters*, $(\tilde{h}_m)_{m\in\mathbb{Z}} = (h_{-m})_{m\in\mathbb{Z}}$ and $(\tilde{\ell}_m)_{m\in\mathbb{Z}} = (\ell_{-m})_{m\in\mathbb{Z}}$, are just the mirror images of the decomposition filters. If ϕ and ψ have compact support, filters h and ℓ have finitely many taps.

It can be shown (see [16]) that exact reconstruction, $x = x'$, by the fast pyramidal algorithm as represented in Figure 2.1 is possible only if

$$\sum_k [\ell_{m-2k}\ell_{n-2k} + h_{m-2k}h_{n-2k}] = \delta_{mn} \tag{2.50}$$

Since $h_n = (-1)^n \ell_{-n+1}$, this reduces to

$$\sum_k = \ell_k \ell_{k-2m} = \delta_{m0}. \tag{2.51}$$

2.2.7 Biorthogonal Wavelet Bases

Biorthogonal wavelets bases have two dual bases, $(\psi_{mn})_{(m,n)\in\mathbb{Z}}$ and $(\tilde{\psi}_{mn})_{(m,n)\in\mathbb{Z}}$, each obtain from a single function, ψ or $\tilde{\psi}$. Unlike orthogonal wavelets, which obtain reconstruction filters coefficients by simply taking the mirror images $(h_{-n})_{n\in\mathbb{Z}}$ and $(\ell_{-n})_{n\in\mathbb{Z}}$ of the decomposition filters $(h_n)_{n\in\mathbb{Z}}$ and $(\ell_n)_{n\in\mathbb{Z}}$, biorthogonal wavelets offers tremendous flexibility in the design of synthesis filters. Moreover, they allow for symmetric filters, which is impossible in the orthogonal case, except for the Haar wavelet.

The multiresolution analysis for biorthogonal wavelet bases differs from the orthogonal case in that it has two hierarchies of approximation spaces.

$$\cdots \subset V_1 \subset V_0 \subset V_{-1} \subset \cdots \tag{2.52}$$

$$\cdots \subset \tilde{V}_1 \subset \tilde{V}_0 \subset \tilde{V}_{-1} \subset \cdots .$$

with spaces W_j and $\tilde{W}_j$ satisfying $W_j \perp \tilde{V}_j$ and $\tilde{W}_j \perp V_j$, but V_j not orthogonal to W_j. As a result, (h_n) and (ℓ_n), representing the coefficients for the decomposition filters in Figure 2.1, and $(\tilde{h}_n)$ and $(\tilde{\ell}(n))$ the coefficients for the reconstruction filters, are related to each other quite differently. Let a data sequence $a_x(0,\cdot)$ be given. Then, its decomposition is obtained by

$$a_x(1,n) = \sum_k \ell_{2n-k} a_x(0,k) \tag{2.53}$$

$$d_x(1,n) = \sum_k h_{2n-k-1} a_x(0,k). \tag{2.54}$$

and its reconstruction by

$$a'_x(0,l) = \sum_n [\tilde{\ell}_{2n-l} a_x(1,n) + \tilde{h}_{2n-l-1} d_x(1,n)]. \tag{2.55}$$

The exact reconstruction requirement, $a'_x(0,\cdot) = a_x(0,\cdot)$, (see Figure 2.1), imposes the condition:

$$\sum_n [\tilde{\ell}_{2n-l}\ell_{2n-k} + \tilde{h}_{2n-l-1}h_{2n-k-1}] = \delta_{lk}. \tag{2.56}$$

which in the z domain is equivalent to

$$\frac{1}{2}[\ell(z)\tilde{\ell}(z) + h(z)\tilde{h}(z)] = 1 \tag{2.57}$$

$$\frac{1}{2}[\ell(-z)\tilde{\ell}(z) + h(-z)\tilde{h}(z)] = 0, \tag{2.58}$$

where $h(z) = \sum_{n=-\infty} h_n z^n$ and $\bar{h}(z) = \sum h_{-n} z^n = \sum_n h_n z^{-n}$, which reduces to the following

conditions on the filter coefficients:

$$\sum_n \ell_n \overset{\circ}{\ell}_{n-2k} = \delta_{k0} \tag{2.59}$$

$$h_n = (-1)^{n-1} \overset{\circ}{\ell}_{-n+1}. \qquad \overset{-}{h} = (-1)^{n-1} \ell_{-n+1}. \tag{2.60}$$

The scaling function and wavelet are given by

$$\varphi(t) = \sqrt{2} \sum_{n \in \mathbb{Z}} \ell_n \varphi(2t - n) \tag{2.61}$$

$$\psi(t) = \sqrt{2} \sum_{n \in \mathbb{Z}} h_{n-1} \varphi(2t - n). \tag{2.62}$$

respectively. Their duals are given by

$$\overset{\circ}{\varphi}(t) = \sqrt{2} \sum_{n \in \mathbb{Z}} \overset{\circ}{\ell}_n \varphi(2t - n) \tag{2.63}$$

$$\overset{\circ}{\psi}(t) = \sqrt{2} \sum_{n \in \mathbb{Z}} \overset{\circ}{h}_{n-1} \overset{\circ}{\varphi}(2t - n). \tag{2.64}$$

respectively.

The functions $\psi_{j,k}(t) = 2^{-j/2} \psi(2^{-j}t - k)$, $j, k \in \mathbb{Z}$, constitute a basis in $L^2(\mathbb{R})$. Their dual frame is given by the $\overset{\circ}{\psi}_{j,k}(t) = 2^{-j/2} \overset{\circ}{\psi}(2^{-j}t - k)$, $j, k \in \mathbb{Z}$ and for any $x \in L^2(\mathbb{R})$ we have

$$x = \sum_{(j,k) \in \mathbb{Z}^2} \left\langle x, \overset{\circ}{\psi}_{j,k} \right\rangle \psi_{j,k} = \sum_{(j,k) \in \mathbb{Z}^2} \langle x, \psi_{j,k} \rangle \overset{\circ}{\psi}_{j,k} \tag{2.65}$$

Moreover, the $\psi_{j,k}, \overset{\circ}{\psi}_{j,k}$ constitute two dual Riesz bases, with

$$\left\langle \psi_{j,k}, \overset{\circ}{\psi}_{l,m} \right\rangle = \delta_{j,l} \delta_{k,m} \tag{2.66}$$

if and only if

$$\int \phi(t)\tilde{\phi}(t-k) = \delta_{k,0}. \tag{2.67}$$

Additional details on biorthogonal wavelets may be obtain in [16][22].

2.3 Self-Similarity, LRD & SRD

2.3.1 Self-Similarity

There are several different—not necessarily equivalent—definitions of self-similarity. The standard one states that a *continuous-time* process $y = \{y(t) : t \geq 0\}$ is self-similar (H-ss) if it satisfies the equation:

$$y(t) \stackrel{d}{=} a^{-H} y(at), \quad \forall t \geq 0,\ \forall a \in \mathbb{R}^+,\ H \in (0,1) \tag{2.68}$$

where $\stackrel{d}{=}$ denotes equivalence in the sense of finite dimensional distribution. This definition of self-similarity, however, has a significant drawback. Unless $y(t)$ is degenerate, i.e., $y(t) = 0$ for every $t \geq 0$, $y(t)$ cannot be stationary (stationarity demands that $y(t) \stackrel{d}{=} y(at)$[23]). In many cases, however, the incremental process $x(t) = y(t) - y(t-1)$, is assumed to be stationary. In this case, the process y is then said to be H-sssi. The canonical example of this is *fractional Brownian Motion*, with Hurst parameter $H = 1/2$, where its incremental process is *fractional Gaussian noise*.

Another definition of self-similarity, which is more appropriate in the context of time-series analysis, assumes $y = (y_i)_{i \in \mathbb{N}}$ to be a discrete-time process with stationary increments

$x = (x_t)_{t \in \mathbb{N}}$, where $x_t = y_t - y_{t-1}$. Let

$$x^{(m)}(k) = \frac{1}{m} \sum_{t=(k-1)m+1}^{km} x_t. \tag{2.69}$$

That is, $x^{(m)}$ is obtained by partitioning x into non-overlapping blocks of size m and then averaging each block. k is the block index. If x is the incremental process of the self-similar process y define by (2.68), it follows (see [23]) that for each m

$$x \overset{d}{=} m^{1-H} x^{(m)}. \tag{2.70}$$

Definition 2.16 (Self-Similarity and Second-order Self-Similarity) *The stationary process $x = (x_t)_{t \in \mathbb{N}}$ is said to be* **exactly self-similar** *if it satisfies equation (2.70) for all aggregation levels m and* **asymptotically self-similar** *if it satisfies (2.70) as $m \to \infty$. A covariance-stationary sequence x is said to be* **exactly second-order self-similar** *if the process $m^{1-H} x^{(m)}$ has the same variance and autocorrelation as x and* **asymptotically second-order self-similar** *if the process $m^{1-H} x^{(m)}$ has the same variance and autocorrelation as x, for all m, or as $m \to \infty$.*

The autocorrelation of an exactly second-order self-similar process is given by

$$\gamma_x(k) = \frac{\sigma^2}{2}((k+1)^{2H} - 2k^{2H} + (k-1)^{2H}) \tag{2.71}$$

(see [24]).

2.3.2 Definition of Long- and Short- Range Dependence

Intuitively, a long-range dependent process (LRD) (also known as *persistence phenomenon* [25] or the *Hurst effect* [26]) can be considered as a phenomenon in which a current observation (sample) is significantly correlated to observations far away from it in time. This phenomenon is of particular interest to traffic modeling, since it has been discovered recently that both Ethernet traffic [7] and video sources [8][27] possess long-range dependence.

A more formal definition of long-range dependence is as follows:

Definition 2.17 (long-range dependence) *Let $x(t)$ be a covariance-stationary process with mean μ and variance σ^2 and autocorrelation $r(k)$, $k \in \mathbb{N}$. $x(t)$ is said to exhibit long-range dependence if*

$$r(k) \sim |k|^{-D} L_1(k), \text{ as } k \to \infty \tag{2.72}$$

where $0 < D < 1$ and $L_1(\cdot)$ is a slowly varying function, that is, $\lim_{t\to\infty} L_1(tw)/L_1(t) = 1$, for all $w > 0$. (Examples of slowly varying function include constants and logarithms).

Under weak regularity conditions on $L_1(\cdot)$, [28] [29][8] show that $x(t)$ is a long-range dependent process if its spectral density is given by

$$f(\lambda) = \lambda^{-\alpha} L_2(\lambda). \tag{2.73}$$

where $0 < \alpha < 1$ and $L_2(\lambda)$ is a slowly varying function. The spectral density function of x is formerly defined as $f(\lambda) = \sum_k r(k) e^{ik\lambda}$. Thus, a covariance-stationary, long-range de-

pendent. stochastic process is also a $1/f$ process. i.e., it is stationary and satisfies Equation (2.73) for some $\alpha \in (0, 1)$.

From the definition of long-range dependence. it can be inferred that a long-range dependent process is characterized by a hyperbolically decaying autocorrelation function. Moreover. the summation of the autocorrelation is infinity. i.e.,

$$\sum_{k=0}^{\infty} r(k) = \infty. \tag{2.74}$$

Thus. although individual correlations at large lags may be small. their summation will have a drastic effect on the overall sum. Therefore. large-lag correlation should not be ignored in the traffic modelling problem.

A *short range dependent* (SRD) process is defined as follows:

Definition 2.18 (short-range dependence) *Let $x(t)$ be a covariance-stationary stochastic process with mean μ. variance σ^2 and autocorrelation function $r(k)$. The process $x(t)$ is said to exhibit short-range dependence if*

$$r(k) \sim |\rho|^k L_1(k), \ k \to \infty. \tag{2.75}$$

where $-1 < \rho < 1$ and $L_1(\cdot)$ is a slowly varying function.

Thus. a short-range dependent process is characterize by an exponentially-decaying autocorrelation. i.e., $r(k) \sim \rho^k$, $0 < \rho < 1$. resulting in a summable autocorrelation function $\sum_k r(k) < \infty$. From the spectral perspective. short-range dependent processes are characterized by a spectral density $f(\lambda)$ which is finite for $\lambda = 0$. On the other. (2.72) or (2.73)

implies that $\sum_k r(k) = \infty$, i.e., a long-range dependent processes has a spectral density that increases without bound as the frequency λ approaches zero.

2.4 Examples of Long- and Short- Range Dependent Gaussian Processes

This section gives some concrete examples of short- and long- range dependent Gaussian processes. A large class of the traditional Markov models are characterized by exponentially decreasing autocorrelation functions and are therefore *short-range dependent*. One such example is the *Auto-regressive Moving Average* (ARMA) model [30]. An ARMA process $(x_k)_{(k\in\mathcal{Z})}$ is defined by

$$\Phi(B)x_k = \Theta(B)\epsilon_k. \tag{2.76}$$

where, for each integer k, ϵ_k is an independent Gaussian random variable with zero mean and unit variance. B represents the backward shift operation, i.e., $Bx_k = x_{k-1}$. $\Phi(B)$ and $\Theta(B)$ are polynomials of degree p and q, respectively. When $q = 0$, the ARMA process degenerates to an auto-regressive (AR) process. The autocorrelation function of an ARMA process decays exponentially at large lags.

Long-range dependent Gaussian processes include the (asymptotically) second-order self-similar processes. Fractional Auto-regressive Integrated Moving Average (FARIMA) processes [25][30], and the (exactly) second-order self-similar processes, Fractional Gaussian Noise (FGN) processes [31], which is the incremental process of the self-similar process, Fractional Brownian Motion (FBM).

FARIMA(p.d.q) [25] is a fractional differentiation of the Auto-regressive Moving Average (ARMA(p.q)) process, where p and q represent the orders of the ARIMA(p.q) process, and $0 < d < 0.5$ is the differentiation degree. It is the generalization of ARIMA(p.d.q) [30] models for integer d. A FARIMA process x_k is defined by

$$\Phi(B)\triangle^d x_k = \Theta(B)\epsilon_k. \tag{2.77}$$

where B represents the backward-shift operation, i.e., $Bx_k = x_{k-1}$. $\Theta(B)$ and $\Phi(B)$ are polynomial functions of orders p and q, respectively. $\triangle = 1 - B$ is the differential operator. $\triangle^d$ is a fractional differencing operator.

$$\triangle^d x_k = \sum_{i=0}^{\infty} \binom{d}{i} (-1)^i x_{k-i} \tag{2.78}$$

where the fractional binomial coefficient is

$$\binom{d}{i} (-1)^i = \frac{\Gamma(-d+i)}{\Gamma(-d)\Gamma(i+1)} \tag{2.79}$$

and $\Gamma(\cdot)$ represents the Gamma function. It has been shown [25] that for a FARIMA(p.d.q) process, the Hurst parameter H equals $0.5 + d$. For positive p or q, FARIMA(p.d.q) is a mixture of both long- and short-range dependence. Because FARIMA(p.d.q) has $p + q + 3$ parameters, it is much more flexible than FGN in terms of simultaneously modeling SRD and LRD.

The interesting range for d is $-1/2 < d < 1/2$. It has been shown in [32] that a $FARIMA(p, d, q)$ process has a spectral density and autocorrelation satisfying $f(\lambda) \sim \lambda^{-2d}$, as $\lambda \to 0$

and $r(k) \sim k^{2d-1}$, as $k \to \infty$. Thus, for $d = 0$, we obtain the short range dependent proses ARMA(p,q), long range dependence or persistence occurs for $d \in (0, 1/2)$, and anti-persistence for $d \in (-1/2, 0)$.

2.5 Estimating LRD Parameters

This section summarizes several of the methods used to estimate the parameters of long range dependence (LRD) in the time, frequency and wavelet domains. The time- and frequency- domain estimators have concentrated mainly on the estimation of the parameter α or equivalently the Hurst parameter $H = (1 + \alpha)$. However, the recent findings of Abry and Veitch [33] claims that there is not one, but *two* parameters, (α, c_r) or equivalently (α, c_f), which characterize long range dependence. Although α is the most important since it defines the existence of LRD itself, and governs the characteristic scaling behavior of a LRD process as well as statistics derived from it, the second parameter plays a significant role in fixing the absolute size of LRD generated effects.

2.5.1 Time Domain Estimators

The classical estimator for H is the *variogram* or *R/S estimator* [34]. This estimator yields poor statistical performance in the form of high bias and suboptimal variance. The Allan variance [35] is a better estimator than the variogram. It measures the expectation of

squared difference of averages of the data x within a window of length T as follows:

$$V_1(T) = \sum_{k=1}^{K} \left(\int_{t_k - T}^{t_k} x(u)du - \int_{t_k}^{t_k + T} x(u)du \right) \tag{2.80}$$

where K is the number of segments of size T in the data, which also behaves as a power law of T when LRD is present and allows an unbiased estimation of H.

The best estimator for H, as well as for other parameters, is the Gaussian Maximum Likelihood Estimator [34]. However, its exact algorithm is computational expensive. Thus, alternative algorithms have been developed; among these, the *Whittle Estimator* [34][33] offers the best performance yielding an asymptotically unbiased and efficient estimator, just as in the exact case.

Time- and frequency- based estimators of H suffer a variety of drawbacks [33], which are unacceptable for many real-time applications in networking. They are mainly parametric in nature and, thus, require a parameterized family of model processes to be chosen a priori. Many—due to their high computational complexities and memory requirements—cannot be implemented exactly for large datasets. Moreover, they are not naturally matched to the essentially simple scale behavior of LRD processes.

2.5.2 Frequency-based Estimators

Since the spectrum of an LRD process x behaves like a power law $(1/|\nu|^{\alpha})$ for ν close to zero, it is natural to think of spectral estimators for the parameters of LRD. One standard

frequency-domain estimator is the following:

$$f_2(\nu) = \sum_{k=1}^{P} \int x(t - kL) w_L(t) \exp(i2\pi\nu t) dt|^2 \tag{2.81}$$

where P denotes the number of data pieces, each of length L, w_L a weighting window. Applying $s^2(\nu)$ to $1/|\nu|^\alpha$ signals, results in an estimator for H, based on linear fit in $\log(\nu)$ vs. $\log(s_2(\nu))$ plot, which has the undesirable effect of being strongly bias.

2.5.3 Wavelet-based Estimators

The unbiased semi-parametric wavelet-based estimator of Abry & Veitch, first proposed in [33] to be later improved in [36], seems to provide a solid alternative. It studies the scale-dependent properties of data directly via the coefficients of a joint scale-time wavelet decomposition. Thus, very little needs to be assumed about the underlying process. It can easily be implemented using techniques from non-redundant multiresolution analysis. It is robust against an important class of non-stationarity—namely, the addition of non-deterministic trends. This is an important advantage in the LRD context where it is difficult in theory and in practice to distinguish between real trends and long-term sample path variations due to LRD (see [33] and references within).

2.5.3.1 The Abry-Veitch Estimator

The Abry-Veitch Estimator starts with the basic definition of an LRD process—namely, the autocorrelation function of a wide-sense stationary LRD process takes the form $\gamma_x(k) \sim$

$c_r k^{-2-2H}$, $H > 0.5$, corresponding to divergence of the autocorrelation sum, $\sum_{k\in\mathbb{Z}} r_x(k) = \infty$, or equivalently, the spectral density satisfies:

$$f(\nu) \sim c_f |\nu|^{1-2H}, \quad \nu \to 0. \tag{2.82}$$

Recall that an LRD process belongs to the class of $1/|\nu|^\alpha$ processes for ν close to 0 where $\alpha = 2H - 1$. Define the following statistical estimator for s:

$$\hat{f}_x(2^{-j}\nu_0) = \frac{1}{n_j} \sum_{k\in\mathbb{Z}} |d_x(j,k)|^2 \tag{2.83}$$

where $|d_x(j,k)|^2$ measures the amount of energy in the analyzed signal about the time instant $2^j k$ and frequency $2^{-j}\nu_0$, with ν_0 an arbitrary reference frequency selected based on choice of ψ. The integer, n_j, is the available number of wavelet coefficients at octave j and is usually given by $n_j = 2^{-j}n$ where n is the length of the data. $\hat{f}(\nu)$ may be interpreted as a measure of the amount of energy that lies within a given bandwidth around the frequency ν.

If x is wide-sense stationary we also have the following estimator for s:

$$E\{\hat{s}(2^{-j}\nu_0)\} = \int f(\nu) 2^j |\Psi(2^j\nu)|^2 d\nu. \tag{2.84}$$

This estimator, however, suffers from the so-called convolutive bias: The spectrum to be estimated is mixed within a frequency range corresponding to the frequency width of the analyzing window at scale j. However, recalling the asymptotic behavior of the spectrum of an LRD process and assuming, momentarily, that this behavior holds for all frequencies,

(2.84) can be rewritten as

$$E\{s(2^{-j}\nu_0)\} = c_f(2^{-j}\nu_0)^{1-2H} \int |\nu|^{1-2H} |\Psi(\nu)|^2 d\nu \tag{2.85}$$

$$= s(2^{-j}\nu_0)^{1-2H} \int |\nu|^{1-2H} |\Psi(\nu)|^2 d\nu. \tag{2.86}$$

For $\frac{1}{\nu^{2H}}$ processes, (2.85) shows that the convolutive bias is converted to a multiplicative one and that this multiplicative constant is independent of j. Therefore, an unbiased estimator $\hat{H}$ of H can be designed by a linear regression of $\log_2(s_x(2^{-j}\nu_0)$ on j as follows:

$$\log_2 s_x(2^{-j}\nu_0) = log_2 \left(\frac{1}{n_j} \sum_{k \in \mathbb{Z}} |d_x(j,k)|^2 \right) = (2\hat{H} - 1)j + c \tag{2.87}$$

where $\hat{c}$ estimates $\log_2(c_f \int |\nu|^{1-2H} |\psi(\nu)|^2 \, d\nu)$ provided that

$$\int |\nu|^{1-2H} |\psi(\nu)|^2 \, d\nu \tag{2.88}$$

converges. Performing a weighted least-square fit between octaves j_1 and j_2 yields the following formula for the estimator $\hat{H}$ of the Hurst parameter H:

$$\hat{H}(j_1, j_2) = \frac{1}{2} \left[\frac{\sum_{j=j_1}^{j_2} C_j j \eta_j - \sum_{j=j_1}^{j_2} C_j \eta_j}{\sum_{j=j_1}^{j_2} C_j \sum_{j=j_1}^{j_2} C_j j^2 - \left(\sum_{j=j_1}^{j_2} C_j j \right)^2} \right] + 1 \tag{2.89}$$

where $\eta_j = log_2 \left(\frac{1}{n_j} \sum_{k \in \mathbb{Z}} |d_x(j,k)|^2 \right)$ is theoretical asymptotic variance and the weight $C_j = \frac{n \ln^2 2}{2^{j+1}}$ is the inverse of η_j.

For $\hat{H}$ to be an unbiased estimator of H, a few conditions must be met. First, Equation (2.82) must be true for each j. This, however, can be relaxed since the range of scales $\{j_1, \ldots, j_2\}$ is chosen by the user. Second, the number of vanishing moments N must satisfy $N > H - 1$.

2.5.3.2 The Abry-Veitch Joint Estimator

The main difference between the Abry-Veitch Joint Estimator and the original Abry-Veitch (AV) estimator is that the former provides a means of computing the important, yet long neglected, second parameter of LRD c_γ or, equivalently, c_f. Thus, instead of the single Hurst parameter H, there are now two LRD parameters to be considered-namely (α, c_γ) or equivalently (α, s). Like the original AV, the joint estimator is also based on linear regression and thus uses a similar approach as used there to estimate α and c_f. Thus, only the main results are summarized here (see [36]). Recall that the wavelet coefficients are given by $\{d_x(j,k) = \langle x, w \rangle (j,k) \in \mathbb{Z}^2\}$ and consider the following definitions:

$$\mu_j = 1/n_j \sum_{k \in \mathbb{Z}} (d_x(j,k))^2 \tag{2.90}$$

$$\sigma^2 = var(log_2(\mu_j)) = \zeta(2, n_j/2)/\ln^2 2 \simeq 2/(n_j \ln^2 2) \tag{2.91}$$

where ζ is a generalized Riemann Zeta function. Also, consider the following definitions:

$$S = \sum 1/\sigma_j^2, \qquad S_x = \sum x_j/\sigma^2, \qquad S_{xx} = \sum x_j^2/\sigma_j^2, \tag{2.92}$$

$$w_j = \frac{(Sj - S_x)/\sigma_j^2}{SS_{xx} - S_x^2}, \quad v_j = \frac{(S_{xx} - S_x j)/\sigma_j^2}{SS_{xx} - S_x^2}, \quad g_j = \psi(n_j/2)/\ln 2 - \log_2(n_j/2). \tag{2.93}$$

$$y_j = \log_2(\mu_j) - g_j, \qquad \hat{b} = \sum w_j y_j, \qquad \hat{a} = \sum v_j y_j. \tag{2.94}$$

Then, the estimators are given by:

$$\hat{\alpha} = \hat{b} \tag{2.95}$$

$$\widehat{c_f C} = p \cdot 2^{\hat{a}}, \text{ where } p = \prod \frac{\Gamma(n_j/2)\exp(\psi(n_j/2)v_j)}{\Gamma(v_j + n_j/2)} \tag{2.96}$$

$$\hat{C} = \begin{cases} \int \Psi^2 d\nu 0 \leq 0. \\ \int |\nu|^{-\hat{\alpha}} |\Psi(\nu)|^2 d\nu 0 < \hat{\alpha} < 1. \\ \int |\nu|^{-1} |\Psi(\nu)|^2 d\nu \hat{\alpha} \geq 0 \end{cases} \tag{2.97}$$

$$\hat{c}_f = \widehat{cC}/\hat{C}. \tag{2.98}$$

2.6 Wavelet Transform of Self-similar and LRD Processes

The wavelet transform of an LRD (or self-similar) process is a short-range dependent process (SRD); or is at least a good approximation of it [10][11][33][34]. Since there are already a variety of efficient SRD models in existence [37][38][39][40][41], it might be more advantageous to analyze and model LRD processes in the wavelet domain. This section summarizes some of the important results of the wavelet transform of LRD and self-similar processes.

2.6.1 Some Key Results from Wavelet Transform

In [11] it has been shown that the following two results from MRA and wavelets are very significant for not only self-similar and LRD processes but for scaling processes in general:

- **W1**: The wavelet basis is constructed from the dilation (change of scale) of the wavelet function, that is, the wavelet analyzing functions themselves $\psi_{j,k}$ exhibit a scale invariance feature.

- **W2**: The wavelet function ψ has $N \geq 1$ *vanishing moments*:

$$\int_{\mathbb{R}} t^k \psi = 0, \quad k = 0, 1, \cdots, N-1. \tag{2.99}$$

The choice of N is dependent on the choice of ψ. The Fourier transform Ψ satisfies the condition $|\Psi(\nu)| \approx |\nu|^N$, $\nu \to 0$ [16].

2.6.2 Wavelet Transform (WT) of H-ss and H-sssi Processes

Let x be a self-similar process (also called an H-ss process), then, for each octave j the wavelet coefficients $d_x(j, \cdot)$ of x exactly reproduces the self-similarity through the following central scaling property (see [11] and references therein):

- **P0 SS**: In the case of the DWT, $d_x(j,k) = \langle x, \psi_{j,k} \rangle$, so that

$$(d_x(j,0), d_x(j,1), \ldots, d_x(j,n_j-1)) \stackrel{d}{=} 2^{j(H-1/2)} (d_x(0,0), d_x(0,1), \ldots, d_x(0,n_j-1)). \tag{2.100}$$

For the CWT, $(W_\psi f)(a,b) = \langle x, \psi_{a,b} \rangle$, hence

$$(W_\psi x)(ca, cb_1), \cdots, (W_\psi x)(ca, cb_n)) \stackrel{d}{=} c^{H-1/2}((W_\psi x)(a,b_1), \ldots, (W_\psi x)(a,b_n)) \quad \forall c > 0. \tag{2.101}$$

and if x is second-order stationary, then from (2.100), we have

$$E[d_x(j,k)^2] = 2^j(2H-1)E[(d_x(0,k))^2] \tag{2.102}$$

(a consequence of **W1**).

If in addition x is H-sssi (see Section 2.3.1), then properties **W1** and **W2** combine to produce the following salient properties:

- **P1 SS**: The Wavelet coefficients with a fixed scale index $\{d_x(j,k) \mid k \in \mathbb{Z}\}$ form a stationary process (a consequence of **W2**). Thus, Equation (2.102) now becomes

$$E\{(d_x(j,k))^2\} = 2^{j(2H+1)}C(H,\psi)\sigma^2, \quad \forall k. \tag{2.103}$$

with $C(H,\psi) = \int |t|^{2H}(\int \psi_0(u)\psi(u-t)du)dt$ and $\sigma^2 Ex(1)^2$.

- **P2 SS**: From the specific correlation structure of the H-sssi process:

$$E[x(t)x(s)] = \frac{\sigma^2}{2}\{|t|^{2H} + |s|^{2H} - |t-s|^{2H}\}. \tag{2.104}$$

it can be shown that for $N \geq H + \frac{1}{2}$, the correlation between wavelet coefficients located at different positions is very small and that their decay can be controlled by increasing N [42][43][44][45]:

$$E[d_x(m,k)d_x(n,l)] \approx |2^m k - 2^n l|^{2H-2N}, \quad |2^m k - 2^n l| \to +\infty. \tag{2.105}$$

2.6.3 WT of LRD Processes

Let x be a second-order stationary process. Its wavelet coefficients $\{d_x(j,k)\}$ satisfy the following properties:

- P0 LRD:

$$E[d_x(j,k)^2] = \int S_x(\nu) 2^j \Psi(2^j \nu)^2 d\nu \tag{2.106}$$

where $S_x(\nu)$ and $\Psi(2^j\nu)$ stands for the power spectrum and the Fourier transform of x, respectively. This has interpretation of classical *inference* formula of linear filter theory: $E[d_x(j,k)^2]$ is a measure of $S_x(\cdot)$ at frequency $\nu_j = 2^{-j}\nu_0$ (ν_0 depends on ψ) through the constant relative bandwidth wavelet filter (see [11] and references there in).

If the process x satisfy the additional properties **W1** and **W2**, we have the following two additional properties:

- **P1 LRD**: Using $S_x(\nu) \sim c_s |\nu|^{-\alpha}$, $\nu \to 0$ (2.106), we obtain

$$E[d_x(j,k)^2] \sim 2^{j\alpha} c_s C(\alpha,\psi), \quad j \to +\infty. \tag{2.107}$$

where $C(\alpha,\psi) = \int |\nu|^{-\alpha} |\Psi(\nu)|^2 \, d\nu$, $\alpha \in (0,1)$. The case $\alpha = 0$ corresponds to trivial scaling at large scales, leaving only short-range dependence at small scales. This result is due to **W1**.

- **P2 LRD:** The covariance function of any two wavelet coefficients is controlled by N and therefore can decay much faster than the correlation of the LRD process x itself and, in fact, is no longer LRD for $N > \alpha/2$. Since $\alpha \in (0,1)$, this is always satisfied.

$$E[d_x(m,k) d_x(n,l)] \approx |2^m k - 2^n l|^{\alpha-1-2N}, \quad |2^m k - 2^n l| \to +\infty. \tag{2.108}$$

Note that the exponents in **P1 SS** and **P2 SS** are different from those of **P1 LRD** and **P2 LRD**.

2.7 Empirical Studies on the Correlation Structure of Wavelet Coefficients for LRD and SRD

Although the theorectical results in the previous section provide important results on the correlation structure of LRD processes, those results are useful only for $k2^m - (l+1)2^n$ large, where m and n are scale indices and k and l are the shift indices. However, this begs the question as to how does one obtain useful information when $k2^m - (l+1)2^n$ is not large. The empirical approach of Ma [10] has provided useful results that gives deeper insight as to determine the following important questions:

1. how do wavelets distinguish SRD from LRD?
2. what correlations need to be captured by wavelet coefficients?

The next two subsection gives a brief summary of the answer to these questions.

2.7.1 Variance of Wavelet Coefficients

It has been shown empirically [10] that one tool that can be utilized to determine how wavelets distinguish LRD from SRD is a variance-scale diagram. The variance-scale diagram is a plot of the scale j versus the logarithm of the variance of wavelet coefficients at

scale j. $\log_2(var(d_x(j,m)))$. . For example, take the following three well known processes: the short-range dependent process AR(1) with parameter $p = 0.9$, the long-range dependent process FARIMA(0,0.4,0) and, finally, the mixed short- and long- range dependent process FARIMA(1,0.4,0). Their plots are shown in Figure 2.2. As can be seen, the LRD

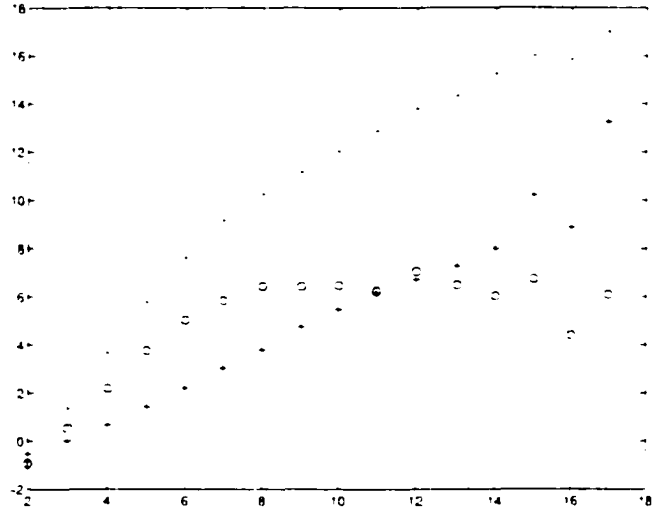


Figure 2.2: log of variance $d_x(j,m)$ versus time scale j. ".": FARIMA(0,0.4,0); "o": AR(0.9); "+": FARIMA(1,0.4,0)

process FARIMA(0,0.4,0) increases exponentially. The SRD process AR(1), $p = 0.9$, increases exponentially for small j but saturates for large j. The variance-scale plot of the mixed SRD-LRD process FARIMA(1,0.4,0) displays the joint characteristics of both the AR(1) and FARIMA(0,0.4,0) processes.

2.7.2 Correlation Structure of Wavelet Coefficients

This section is concerned with what correlations need to be captured by wavelet coefficients. To facilitate this discussion, the notion of wavelets on trees and the single index wavelet coefficient, as develop by Basseville et. al. [46], are first introduced. Figure 2.3 shows the

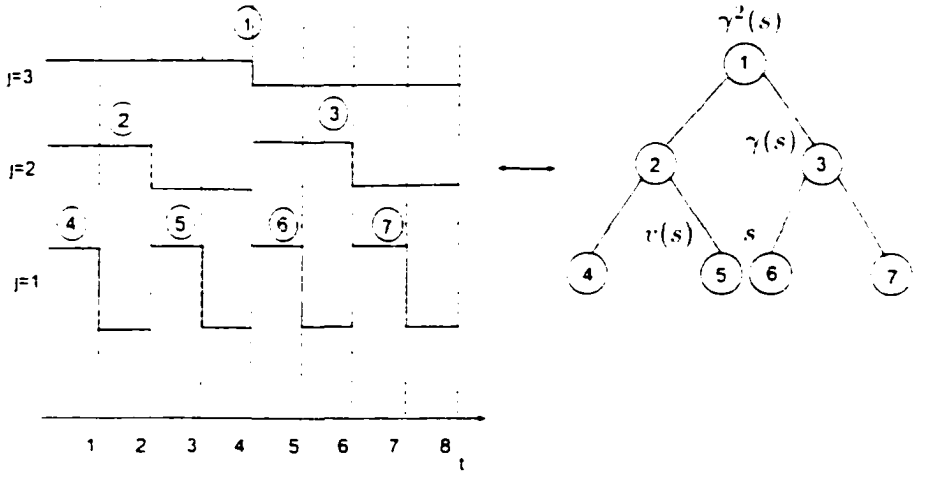


Figure 2.3: Tree representation of the Haar wavelet basis

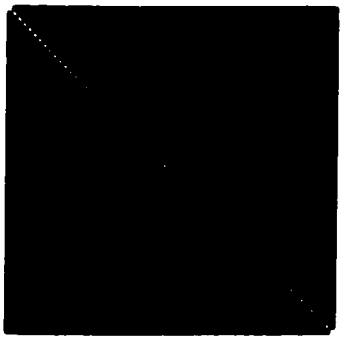
Haar wavelet basis at level $L = 3$ (left) and its corresponding tree diagram (right). The number in the circle represents the one-dimensional index of the corresponding wavelet basis function. In the case of orthogonal wavelets, the single index coefficient s is related to the double index (j, m) by the equation $s = 2^{K-j} + m$, where K represents the limited resolution (or the maximum decomposition level). The functions γ and ν of Figure 2.3 evaluated at node s are defined to be the index of the parent node of s and the index of the left side neighbor of s, respectively.

Now returning to question of the correlation structure of wavelet coefficients. Figure 2.4 (a)-(c) are plots of correlation matrices for the wavelet transform of the SRD process AR(1), $p = 0.9$, the LRD process FARIMA(0,0.4,0) and mixed LRD-SRD process FARIMA(1,0.4,0) with AR parameter $p = 0.9$, respectively. A point (i, j) in the correlation matrix represents the normalized correlation between the i-th and j-th wavelet coefficients, i.e., $\frac{E[d_i d_j]}{\sigma_i \sigma_j}$, where i and j represents the one-dimensional indices (nodes) in the tree diagram. The grayscale is proportional to the magnitude of the correlation matrix. The higher the magnitude of the correlation, the whiter the pixel in the image. Figure 2.4 shows that beside the main diagonal, there are four pairs of lines having "visible" correlation. They correspond to the correlation between $\gamma^k(s)$ and s, where $\gamma(s)$ denotes the parent of node s and $\gamma^k(s)$ denotes the parent of node $\gamma^{k-1}(s)$ with k being 1, 2, 3 and 4 counting from the main diagonal.

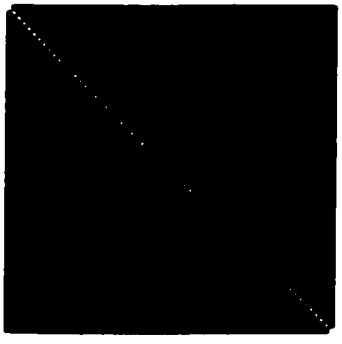
From these figures, it seems reasonable to conclude, as has been done in [10], that the most significant correlation is due to the parent-child correlation. Since the complicated correlations in the time domain concentrates on these types of correlation, it is possible to use a parsimonious model in the wavelet domain to faithfully represent the original traffic.

2.8 Markov Models for Wavelet-Domain Correlation Modeling

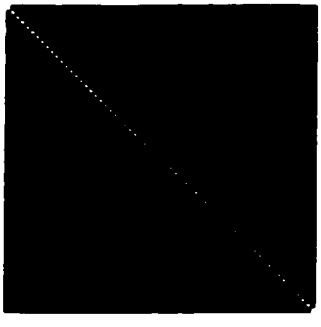
As has been seen in the last section, correlation among wavelet coefficients for LRD and SRD is concentrated in certain types of correlation. This section provides a means of



(a) AR(1), p=0.9



(b) FARIMA(0,0.4,0)



(c) FARIMA(1,0.4,0)

Figure 2.4: Correlation Structures

modelling wavelet coefficients by exploiting the characteristics of the correlation structure of wavelet coefficients. The basic idea is that a parsimonious model can be built using only the strongest correlations, while ignoring the insignificant ones. The wavelet coefficient d_s may be modelled via the following generalized linear equation [10]:

$$d_s = \sum_{t=1, t \neq s}^{N-1} a_s(t) d_t + b_s w_s \tag{2.109}$$

where $\{a_s(t),\ t = 1, 2, \ldots, N-1\}$ and b_s are parameters to be estimated, which are dependent on the one-dimensional index s (see Figure 2.3). The (w_s) are i.i.d random variables.

In order to model a wavelet coefficient in a causal fashion, Equation (2.109) is written

$$d_s = \sum_{t=1}^{s-1} a_s(t) d_t + b_s w_s. \tag{2.110}$$

The estimation of parameters in a linear model can be obtained through estimation theory for Gaussian processes as follows.

Let the random variable z be a linear combination of k Gaussian random variables, denoted by the column vector $\mathbf{y}$, plus a Gaussian noise term. That is,

$$z = \mathbf{y}^T \mathbf{a} + bw. \tag{2.111}$$

where the column vector $\mathbf{a}$ and scalar b are parameters to be determined and w is independent Gaussian noise with unit variance. Note that Equation (2.111) simply rewrites the sum in Equation (2.110) in vector form. The parameter vector a can be estimated by multiplying both sides of Equation (2.111) by the vector $\mathbf{y}$ and taking expectations of both

side of the resulting equation:

$$E[z\mathbf{y}] = E[\mathbf{y}\mathbf{y}^T]\mathbf{a}. \tag{2.112}$$

where $E[z\mathbf{y}]$ represents the cross-correlation between z and $\mathbf{y}$. $E[\mathbf{y}\mathbf{y}^T]$ represents the covariance matrix of $\mathbf{y}$. If $E[\mathbf{y}\mathbf{y}^T]$ is invertible, $\mathbf{a}$ can be determined by:

$$\mathbf{a} = (E[\mathbf{y}\mathbf{y}^T])^{-1}E[z\mathbf{y}]. \tag{2.113}$$

Similarly, multiplying Equation (2.111) by z and taking expectations of both sides of the resulting equation yields

$$b = \sqrt{E[z^2] - E[z\mathbf{y}^T](E[\mathbf{y}\mathbf{y}^T])^{-1}E[z\mathbf{y}]}. \tag{2.114}$$

Therefore the parameters $\mathbf{a}$ and b are determined by the covariance matrix $E[\mathbf{y}\mathbf{y}^T]$, the cross correlation $E[z\mathbf{y}]$ and the variance of z through Equations (2.113) and (2.114). The remainder of this subsection summarizes five models based on Equation (2.110). These models represent a culmination of the works of [47] and [10].

Among these models, there lies an inherent trade-off between complexity and performance. On the one hand, the more complex models lead to better performance. On the other hand, there is a desire for a parsimonious model, i.e., a model that uses as few parameters as possible to model the most significant correlations. In order to study empirically the trade-off between performance and complexity, in the following, five models are presented ranging from the simplest to the most complicated based on the observation in the previous section.

Model 1: The *Independent Wavelet Model* is the simplest of the models. It models only the mean and variance of wavelet coefficients, i.e., $d_x(j,m)$'s are chosen as independent

random variables with zero mean and variance σ_j^2 to be estimated at each octave j. Thus, this model completely neglects dependence among wavelet coefficients and its correlation in the wavelet domain is given by:

$$E[d_{j_1}^{m_1} d_{j_2}^{m_2}] = \begin{cases} \sigma_j^2 & m_1 = m_2 \text{ and } j_1 = j_2. \\ 0 & \text{otherwise.} \end{cases} \tag{2.115}$$

A wavelet coefficient d_s in this case is estimated by $d_s = b_s w$, where (w_s) are iid random variable with zero mean. The parameter, b_s, is the standard deviation of d_s and can be estimated from the square root of the sample variance of wavelet coefficients of a training sequence at the corresponding octave, i.e., $b_s = \sqrt{\frac{1}{n_j}\sum_{m=1}^{n_j} d_x(j,m)^2}$, where n_j is the number of wavelet coefficients at octave j. It has being suggested [10] that this simple model—with appropriate adjustments—suffices for many practical applications.

Model 2: This is a first-order Markov model, which models the parent-child relationship. Results in the last section seem to suggest that this is the most significant correlation. A model which includes the parent-child relationship can be implemented through a first-order Markov model as

$$d_{s_j} = a_j d_{\gamma(s_j)} + b_j w_j. \tag{2.116}$$

where a_j and b_j are parameters to be determined from data. w_j is Gaussian noise with zero mean and unit variance. $\gamma(s_j)$ represents the parent of s_j, where s_j is the one-dimensional index of a wavelet coefficient (node) in the j-th time scale (See Figure 2.3 for an illustration of the one-dimensional index of wavelet coefficients).

Thus, the objective of this model is to model the parent-child correlation in addition to

what is modelled in Model 1. For the sake of convenience, we drop the explicit dependence on j in Equation 2.116 and rewrite it as:

$$d_s = a_s d_{\gamma(s)} + b_s w_s, \tag{2.117}$$

where, again, $a_{\gamma(s)}$ and b_s are parameters to be determined. It can be shown that the parent-child correlation is invariant for wavelet coefficients in the same scale. Therefore, the parent-child correlation $E[d_s d_{\gamma(s)}]$ as well as the targeted variance σ_{d_s} and $\sigma_{d_{\gamma(s)}}$ can be estimated through those of a training sequence. Furthermore, through Equations (2.113) and (2.114), $a_{\gamma(s)}$ and b_s are found to be

$$a_{\gamma(s)} = \frac{E(d_s d_{\gamma(s)})}{\sigma_{d^2_{\gamma(s)}}}, \tag{2.118}$$

and

$$b_s = \sqrt{\sigma^2_{d_s} - a^2_{\gamma(s)} \sigma^2_{d_{\gamma(s)}}}, \tag{2.119}$$

respectively.

Model 3: Model 3 is a third-order Markov model, which models the correlations among s, $\gamma(s)$ and $\nu(s)$. It has been proposed as an improvement to Model 1 in [47] based on the intuition that neighboring wavelets contribute more significantly to the correlation than other wavelets. However, this has shown not to be the case in [10].

The parameters of Model 3 can be calculated in a similar manner to that of Model 2.

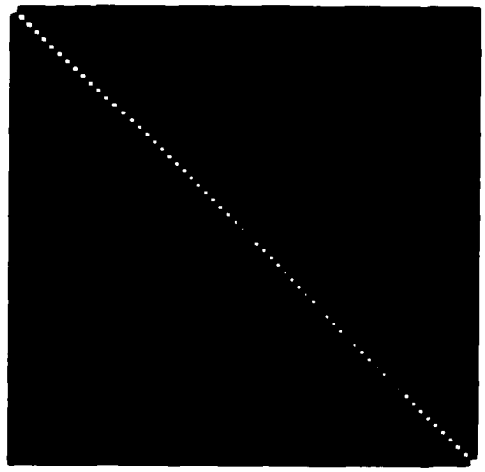
Model 4: Model 4 is an alternative third-order Markov model, which models the correlation among $\gamma(s)$, $\gamma^2(s)$ and s. This model incorporates correlations between node s

and its grandparent node $\gamma^2(s)$ in addition to the parent-child relationship in Model 2. and is equivalent to a third-order Markov model. In the correlation matrix. this model is equivalent to matching the first two strongest line.

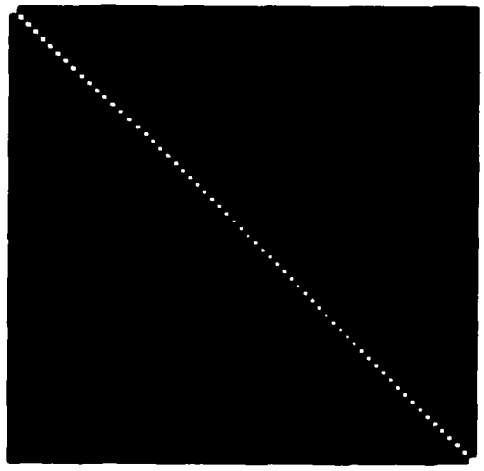
Parameter estimation for Model 4 is slightly different from the previous three models because the invariant property does not hold for all wavelet coefficients in the same scale. The approach used here groups together wavelet coefficients at the same scale and with the same statistics. Model 4 has four such groups. whose shift indices are $\{4k + k_0\}_{k=0}^{2^{k-j-1}-1}$ for $k_0 = 0, 1, 2, 3$ at each j. The correlation to be modelled can be estimated for each group. and Equations (2.113) and (2.114) can be used for each group in the same way as that for the previous three models.

Model 5: Model 5 is a fourth-order Markov model. which models the correlation among $\gamma(s)$, $\gamma^2(s)$, $\gamma^3(s)$ and s. In the graph of the correlation matrix. this is equivalent to matching the first two strongest lines as well as the (barely visible) line near the diagonal. which represents the neighboring relationship.

Parameter estimation for Model 5 is similar to that of Model 4. with the exception that the number of groups increase to 8. The important question now is: how well do Models 1 through 5 capture the correlation structure in the wavelet domain? We plot the correlation matrices of Models 1 and 2 in Figures 2.5 (a) and (b). As expected. Model 1 leaves out all the correlations. Model 2 captures the parent-child relationship. However. its grandparent-child relationship as compared with the original process is significantly diminished. Plots



(a) Model 1



(b) Model 2

Figure 2.5: Wavelet Correlation Matrices for Model 1 and Model 2

of the remaining models may be found in [10]. In terms of complexity of these models. Model 1 requires only one parameter—the variance σ_j^2 at each level j. Models 2. 3. 4. and 5 have two. four. four and five parameters at each scale. respectively.

Chapter 3

Network Preliminaries

3.1 Asynchronous Transfer Mode

Asynchronous transfer mode (ATM) is a technology capable of integrating different types of network into a single consolidated, broadband (high-bandwidth) network. The need for such a technology has arisen because of the conflicting goals of many of the new applications, such as the Internet, the World Wide Web (WWW), video conferencing, video-on-demand, etc., for high bandwidth, without sacrificing the quality of a set of pre-defined parameters. The level of quality an application desires is called the application's *quality of service* (QoS). The key to resolving the conflicting golds of high bandwidth with stringent QoS requirement is *traffic management* (TM). For this reason, ATM, because of its sophisticated traffic management capabilities, has been chosen by the ITU-T as the technology of choice to launch Broadband Integrated Service Digital Network.

3.2 ATM QoS Requirements

QoS requirements are usually specified in terms of information loss, information delay and information variability. The basic QoS parameters for an ATM connection are cell loss ratio, cell transfer delay and cell delay variation (*jitter*). Other QoS parameters comprise of cell misinsertion rate, cell error ratio, cell transfer capacity (throughput), and skew (difference in representation time of two related objects). The major ATM cell performance parameters are *cell error ratio (CER), cell loss ratio (CLR), cell misinsertion rate (CMR), severely errored cell block ratio (SECBR), cell transfer delay (CTD), mean cell transfer delay, cell delay variation (CDV)*. Concise definitions are given in Table 3.1, and a more exhaustive treatment can be found in [48].

Parameter	Acronym	Meaning
Peak cell rate	PCR	Maximum rate at which cells will be send
Sustained cell rate	SCR	The long-term average cell rate
Minimum cell rate	MCR	The minimum acceptable cell rate
Cell delay variance tolerance	CDVT	The maximum acceptible cell jitter
Cell loss ratio	CLR	Fraction of cells lost or delivered too late
Cell transfer delay	CTD	How long delivery takes (mean and maximum)
Cell delay variation	CDV	The variance in cell delivery times
Cell error rate	CER	Fraction of cells delivered without error
Severely-errored cell block ratio	SECBR	Fraction of blocks garbled
Cell misinsertion rate	CMR	Fraction of cells delivered to wrong destination

Table 3.1: Some QoS parameters

3.3 Traffic Characterization and Resource Management

ATM technology is based on the statistical multiplexing of variable rate sources. Due to the burstiness of traffic sources intelligent control mechanisms are required to avoid

degradation in QoS in an ATM network. A significant amount of studies ranging from source characterization to congestion control, solely for the purpose of improving QoS and increasing the network utilization, have already been done. This section summarizes some of these recent studies.

3.3.1 Types of Services

An ATM communication network is comprised of a variety of services such as remote local area networks (LANs), high-resolution still image transfer, high quality interactive videotext, videophone, video conference, distributive TV, voice and others. All of these services have different characteristics [49] in terms of traffic burstiness, bit rate, and call duration as illustrated in Table 3.2.

Table 3.2: Broadband Services and Their Characteristics

Services	$E[s(t)]$	B
Voice	32 Kb/s	2
Interactive Data	1 - 100 Kb/s	10
Bulk Data	1 - 10 1 Mb/s	1 - 10
Standard Quality Video	1 5 - 15 Mb/s	2 - 3
High Definition TV (HDTV)	15 - 150 Mb/s	1 - 2
High Quality Video Telephony	0.2 - 2 Mb/s	5

$s(t)$ is the natural information rate
$E[s(t)]$ is the average information rate of $s(t)$
$B = (max\ s(t))/E[s(t)]$ is the burstiness

Further, each of these imposes different service requirements as shown in Table 3.3. The ATM services are categorized as *constant bit rate* (CBR), *variable bit rate* (VBR) and *available bit rate* (ABR). CBR is used for circuit emulation where a constant bandwidth

Table 3.3: Service Attributes for an ATM Network

Services	BER	PLR	PIR	Delay
Telephony	10^{-7}	10^{-3}	10^{-3}	25 ms / 500 ms
Data Transmission	10^{-7}	10^{-6}	10^{-6}	1000 ms (50 ms)
Broadcast video	10^{-6}	10^{-8}	10^{-8}	1000 ms
Hifi Sound	10^{-5}	10^{-7}	10^{-7}	1000 ms
Remote Process Control	10^{-5}	10^{-4}	10^{-3}	1000 ms

is desired. ABR is subscribed for bulk rate transmission. LAN or TCP/IP traffic. VBR is the most challenging service type with respect to resource management, primarily because heterogeneous traffic sources are multiplexed together and traffic volume fluctuates with time.

Although defined and supported in varying degrees in the network. ABR has enjoyed only modest usage. This is due, in part, by a lack of endpoint support and considerable concern about interaction of ABR with TCP control mechanisms.

On the other hand. Multimedia VBR traffic, such as JPEG and MPEG encoded data [50][51][52], has drawn considerable interest in telecommunication traffic research and applications. Its potentially vast amount of real time services such as image and video applications, is the driving force for its popularity.

The image compression scheme supported by the Joint Photography Experts Group (JPEG) is used primarily for encoding and transmission of still images. The compression techniques employed by JPEG is based on block transforms such as the Discrete Cosine Transform (DCT).

The Moving Pictures Expert Group (MPEG) is a working group of the standardization organization ISO/IEC in charge of the development of international standards for compression, decompression, processing and coded representation of moving pictures, audio and their combination. Thus far, MPEG has created:

- MPEG-1, the standard for storage and retrieval of moving pictures and audio on storage media (approved Nov. 1992)
- MPEG-2, the standard for digital television (approved Nov. 1994)
- MPEG-4 version 1, the standard for multimedia applications (approved Oct. 1998)

and is now developing:

- MPEG-4 version 2 (to be approved Dec. 1999)
- MPEG-7, the content representation standard for multimedia information search, filtering, management and processing (to be approved July 2001).

The most up-to-date information on MPEG is to be found at http://www.cselt.it/mpeg/.

3.3.2 Model-Based Traffic Source Characteristics

Most of the model-based studies rely on statistical characteristics with the imposition of special assumptions and using suitable approximation methods such as fluid models,

phase type processes. Markov Modulated Poisson Process (MMPP). Intercepted Poisson Process (IPP). Fractal Brownian Motion (FBM). among others. Nomura et. al. [53] used three measures: distribution. autocorerrelation and coefficient of variation. to evaluate burstiness. An autoregressive (AR) process or coefficient of variation method can be use to model and characterize video sources. The Autoregressive Moving Average Process (ARMA) was used by Grunenfelder et. al. [54]. A parametric model of encoding algorithms was introduced by Rodriguez. which generalized video traffic characteristics regardless of the actual encoding scheme employed [55].

3.3.3 Non-Model Based Traffic Source Characteristics

Non-model based traffic characterization studies bypass the modeling procedure. Instead. it uses methods such as entropy and energy functions. traffic components decomposition. to mention a few. in order to carry out its analysis.

3.3.4 Bandwidth Analysis and Approximation

One of the major advantages of ATM is its capability to support a wide variety of connections with different bandwidth requirements and traffic characteristics. While this environment provides an increase flexibility in supporting various services. its dynamic nature poses difficult traffic control problems when trying to achieve efficient use of network resources. especially for bandwidth management and allocation of real-time traffic. Providing deterministic guarantees require that the network allocate resources. according to the worst

case scenario. This in turn requires a peak-rate resource allocation scheme, which would result in significant under-utilization of the network for VBR traffic where the peak average ratios is high.

Stochastic fluid models and the theory of large deviations have been used to define effective bandwidth or equivalent capacity of each source. The effective bandwidth is dependent on source characteristics, QoS requirements of the source, and buffer size [56]. An approximation expression for the "equivalent capacity" of both individual and multiplexed connections was proposed by Guerin [57]. Resource management using effective bandwidth has been demonstrated by De Veciana, which is based on the general classification of the ATM connections as deterministic, statistical and best-effort.

However, a study by Choudhury [1], revealed that the effective bandwidth approximation can over-estimate the target small blocking probabilities by several orders of magnitude when there are many sources that are more bursty than Poisson. The effective bandwidth based connection admission control algorithm, which depends solely on the tail asymptotic decay rates, may not be as effective as thought. Choudhury further demonstrated that the effective bandwidth approximation is not always conservative. For sources less bursty than Poisson, the asymptotic constant grows exponentially in the number of sources. Thus, the effective bandwidth approximation can greatly under-estimate the target blocking probabilities. Hence, the need arises for new approximation mechanisms.

Call Admission Control, which determines whether the network has sufficient resources to

support a new connection without degrading the service of existing connections, has also drawn significant attention. Gibbens [58] presented a decision-theoretic approach which employs Bayesian decision theory with time-scale decomposition. Elwalid [59] and others developed QoS assurances, especially the protection of one QoS connection from another, on the basis of leaky bucket adherence by all connections, either at the source and or through network leaking bucket policing with strict drop policy. They demonstrated an approach for determining the admissibility of variable bit rate traffic in terms of allocating buffers and bandwidth to heterogeneous regulated traffic at an ATM node.

3.3.5 Traffic and Congestion Control in ATM Networks

In order to ensure an acceptable quality of service for all coexisting calls sharing the same network resources within the entire duration of their connection, two functions have been defined: Connection Admission Control (CAC) and Usage Parameter Control (UPC).

CAC represents the set of actions taken at call set-up phase to accept or reject an ATM connection. A connection request for a given call is accepted only when sufficient resources are available to carry the new connection through the whole network at its requested QoS while maintaining the QoS of already accepted connection in the network.

Usage Parameter Control or "policing" represents the set of actions taken by the network to monitor and control traffic on an ATM connection in terms of cell traffic volume and cell routing validity. Its main purpose is to enforce the compliance of every ATM connection to

its negotiated traffic contract. Some of the existing policing functions include *Cell Tagging, Leaky Bucket (LB), Jumping Window (JW), Triggered Jumping Windo (TJW), Moving Window (MW)* and *Exponential Weighting Moving Average (EWMA)*[60].

Leaky Bucket, the most popular among the policing functions, has been analyzed as a G/D/1/N queue with finite waiting room N and an appropriate arrival process in work by Butto [61].

3.4 The Effective Dynamic Bandwidth Formula

The Effective Dynamic Bandwidth (EDB) Formula, [2], is a wavelet based bandwidth formula, which has shown very promising results. This section gives a brief discussion of this formula-- highlighting its significance, as well as its limitations.

The EDB formula is given by:

$$B = \alpha_0 \cdot A_m x + A_m \sum_{i=1}^{m} \alpha_i |D_i x|, \tag{3.1}$$

where

- x is the incoming traffic
- $A_j x$ is the j^{th} coarse approximation signal of signal x
- $D_j x$ is the j^{th} detailed approximation for some signal X
- α_i are coefficients to be determined.

The term $A_m x$, which represents the low frequency component of the traffic, reflects the major bandwidth requirement; while the term $A_m \sum_{i=1}^{m} \alpha_i |D_i x|$, which consists basically of an averaging of the high frequency components $\{D_1 x, D_2 x, \cdots, D_m x\}$ represents the traffic fluctuation and has been included to reduce the delay that would have been present were the bandwidth to be based solely on the first term $\alpha_0 \cdot A_m x$. The set of real-valued coefficients $\{\alpha_0, \alpha_1, \cdots, \alpha_m\}$ allow greater adaptation to stringent QoS requirements. If traffic signal x is sampled at T_s seconds, then B is updated every $2^m T_s$ seconds; so that m also determines the update interval. A special case of (3.1) is

$$B = B_{k,m} = A_k x + A_k \sum_{i=m}^{k} |D_i x|. \tag{3.2}$$

where the following assignments have been made: $\alpha_0 = 1$, $\alpha_i = 0$ for $i = 1, 2, \cdots, m$ and $\alpha_i = 1$ for $i = m, m+1, \cdots, k$. The parameter m is related to the bandwidth utilization rate. Figure 3.1 is an illustration of $B_{6,2}$.

First and second order statistical analysis (density function and autocorrelation) of x sampled at $T_s = .5$ seconds and the EDB given in (3.1) with $k = 4$, which corresponds to a bandwidth update interval of 8 seconds, have been computed in [2] for VBR/MPEG and ABR/LAN traffic sources; and comparisons have shown to be quite favorable. One of the important results show that the autocorrelation of the bandwidth representation is similar to the original data, which is a requirement if the formula is to serve as an efficient bandwidth allocator. A similar analysis for VBR/MPEG multiplexed sources at a simulated ATM switch were also shown to be favorable. Finally, the *tail distribution function* (TDF) has been obtained in simulations for a single VBR/MPEG source, mul-

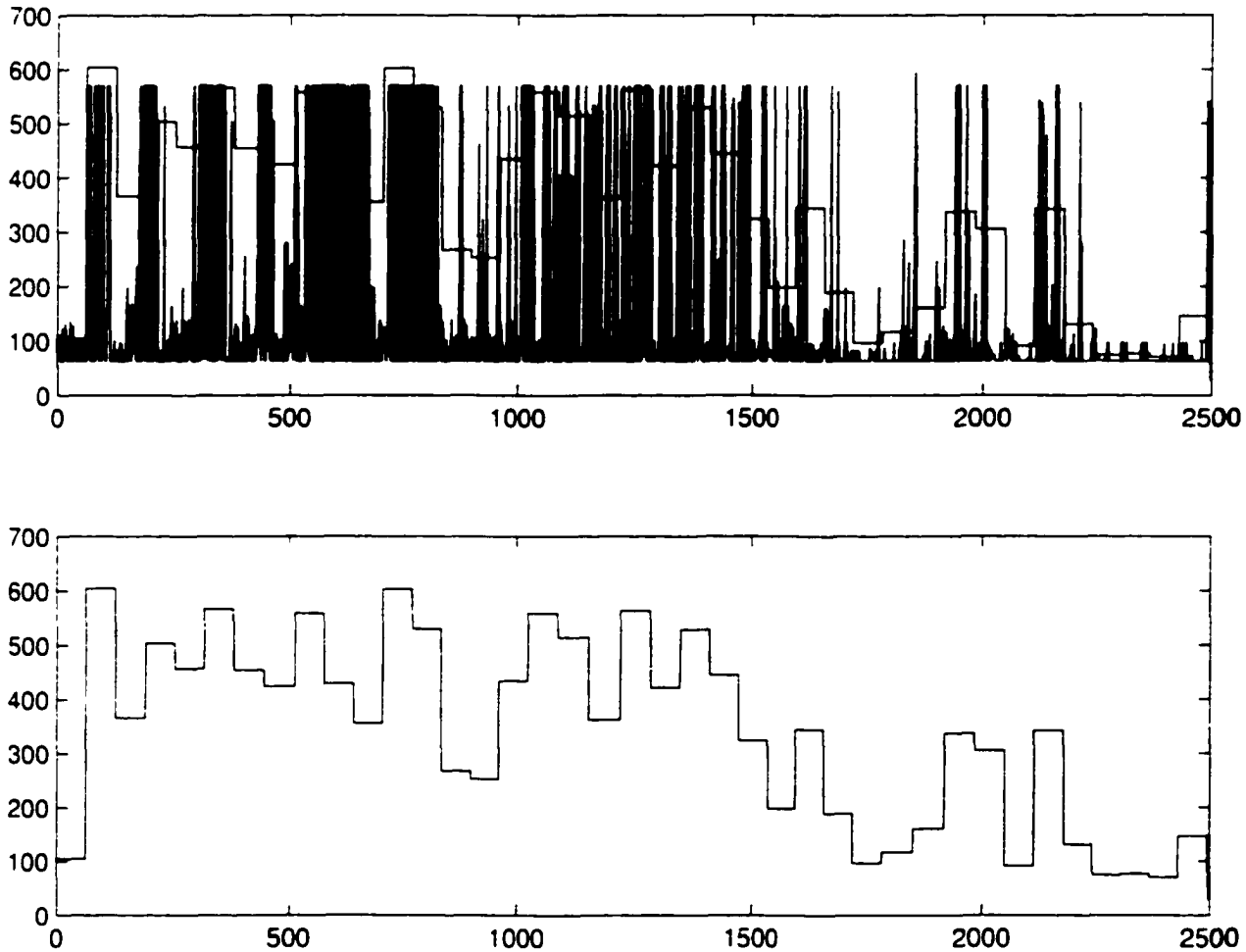


Figure 3.1: EDB: (a) x overlapped with B_{62} (b) B_{62}

tiplexed VBR/MPEG sources and ABR/LAN sources to study the queueing effect at an ATM switch and the results proved to be satisfactory.

Although the EDB formula has shown to be well correlated with the data, it does have some significant drawbacks. From 3.1, it can be seen that significant loss do occur; such loss may not be acceptable for certain applications. Also, no method has been proposed for the determination of coefficients $(\alpha_i)_{i\in\lambda_k}$. Given these drawbacks, we propose a new dynamic bandwidth formula in the next section, which will overcome these limitations.

3.5 A Simple ATM Simulator

In [2], the EDB formula 3.1 has been utilized in an ATM simulator similar to the simplified version of Figure 3.2 to measure the queuing effect of the switch. The Tail Distribution Function (TDF), which is defined , has been obtained for a variety of traffic traces. In this

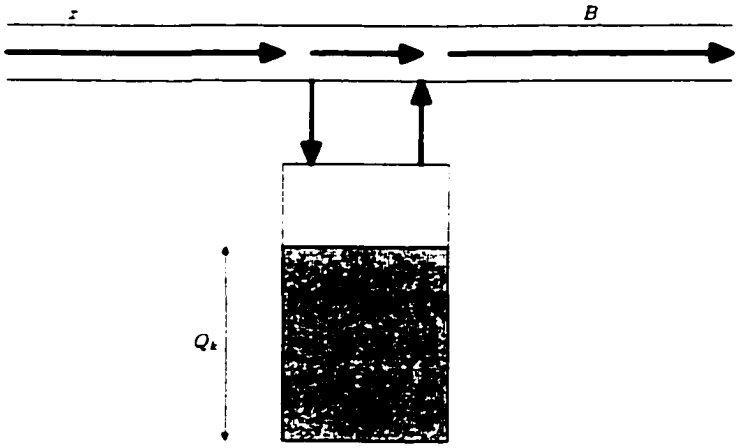


Figure 3.2: ATM Simulator

scenario, when an ABR or VBR traffic stream arrives at the ATM simulator, the signal x is generated as a traffic cell number count in a unit time interval, and the EDB B is calculated using 3.1 and compared with x. If $x(t) < B(t)$, all of the traffic cells are passed to the downstream switching node. If $x(t) > B(t)$, since the traffic cells are bounded by $B(t)$, only $B(t)$ traffic cells are passed to the downstream switching node, and $x(t) - B(t)$ traffic cells remain in the local switch buffer when the switch buffer is unbounded. In subsequent operations, the residual traffic cells are transmitted whenever $x(t) < B(t)$, and only a maximum of $B(t) - x(t)$ is released from the local switch buffer at time t. The Tail distribution function is then obtained as the probability density function of the traffic cell occupancy in the local switch buffer.

In practical scenarios, bounded buffers are assumed. And, since the EDB cannot be pre-generated, it predicted. The prediction scheme used in [2] is an adaptive neural network approach. In Chapter 6, we developed a different formula to represent the bandwidth and prediction schemes for the prediction of this bandwidth.

3.6 Voice Over IP

This section provides a brief summary of Voice Over IP (VoIP). VoIP is the transport of voice traffic using the Internet Protocol (IP). IP is a packet-based protocol, which divides traffic into small packets that are then sent individually to their destination. IP also happens to be the most popular of the network protocols in use today. However IP by itself cannot guarantee any QoS nor can it even guarantee that packets will arrive in the order

in which they were sent. Therefore, other protocols such as Transmission Control Protocol (TCP) and User Datagram Protocol (UDP) have been developed to work in conjunction with IP to deal with these issues. For real time traffic, however, TCP introduces unacceptable delay. UDP on the other hand, when combined with other protocols, such as the Real-Time Transport Protocol (RTP), have yielded much better results for VoIP.

In recent years, VoIP has been generating tremendous excitement in the telecommunications industry. This excitement is driven primarily by the potentially huge profit gains implicit in providing high quality voice over the Internet. Voice have been transmitted successfully for years over the Public-Switched Telephone Network (PSTN). However, the circuit switch technology used by the PSTN is ill-suited for non-voice applications, such as e-mail, WWW, file-transfer, etc., for which IP is so well-suited for. The Internet, which is a vast interconnection of computer networks that communicate using IP, has seen tremendous growth in terms of the number of businesses and the number of users utilizing it. However, thus far, although the Internet has been very successful in delivering non-real time services, it has been unsuccessful in delivering high quality real-time applications to its customers. Indeed, for the most part, it has only been successful in delivering un-paid, low quality, real-time voice and video to its users. This is one of the main reasons that revenues generated by the e-commerce industry are minuscule (millions of dollars) in comparison to those generated by the telephone companies (billions of dollars). For managed networks, however, VoIP has shown great success in transporting high-quality voice. The hope is that the next generation of the Internet might be able to duplicate such high-quality voice levels, and thus enabling the opportunity to offer high-quality, integrated voice-data

services to its users. Thus, one of the greatest challenges facing VoIP is improving QoS. And, one of the key elements in improving QoS is bandwidth management. To ensure that sufficient bandwidth is available to enable high-quality voice necessitates proper control and prioritization of access to the available bandwidth. At present, however, no regulations exist on the Internet to control or prioritize bandwidth. Thus, one person transferring a huge file might cause other user transactions to proceed more slowly and may thereby degrade QoS for these users.

The International Engineering Task Force (IETF) has provided several methods that are designed to enhance or improve QoS. Although these methods by themselves cannot guarantee high-quality voice over the Internet, they have shown to be successful for managed networks. What is needed on the Internet is for the various Internet Service Providers (ISPs) to implement service level agreements so that certain standards are maintained in terms of QoS for the various ISPs on the Internet. The QoS methods of the IETF may be divided into three basic categories resource reservation, prioritization and by simply increasing the available bandwidth. Resource Reservation Protocol (RSVP) is an example of a resource reservation method and is described in RFC 2205. It operates by ensuring resources are available before the transfer of any media from the source to the destination. The process initiates when the sender sends out a PATH message to its targeted destination via a number of routers. The PATH message contains information about the kind of data the sender intends to send, in terms of bandwidth requirement and packet size. When the receiver receives the PATH message, it sends a Reservation Request (RESV) back to the sender along the same route. At each router, resources are allocated based on availability

and upon the receiver's authority to make the request. Finally, when the RESV message reaches the sender with a confirmation, the data (or media) is actually transmitted.

Differentiated Service (DiffServ) is a relatively simple means of prioritizing traffic and is discussed in detail in RFC 2475. This method exploits a field value in the header of an IP version 4 or 6 packet. Depending on the value of this field, a particular traffic stream is marked as requiring a certain type of forwarding. DiffServ defines two types of forwarding: Expedited Forwarding (EF) and Assured Forwarding (AF). EF is specified in RFC 2598. It is a type of service in which packets from a given traffic stream is assigned a minimum departure rate that is greater than the arrival rate at the same node, provided that the arrival rate does not exceed a pre-agreed maximum. The process ensures that queuing delays a removed. Since queuing delays are major causes of end-to-end delay and jitter, the process by extension ensures that delay and jitter are minimized. AF is defined in RFC 2597. It is a type of service in which packets from a given traffic source are forwarded with a high probability, provided that the traffic from that source does not exceed some pre-assigned maximum. AF defines four classes. Each class is allocated a certain amount of resources—bandwidth and buffer space. Within each class, a given traffic can have one of three drop rates. In times of congestion within the resources allocated to a particular AF class at a given router, the packets with the highest drop-rate values are discarded first so that packets with lower drop-rate values will receive some protection. In order for the DiffServ method to work well, the incoming traffic must not contain a high percentage of packets with low drop-rate values. Otherwise, this will defeat the main purpose of the method, which is to ensure that high-priority packets do indeed get through.

Label switching is yet another prioritization method that has garnered quite a bit of interest in the Internet community lately: so much so, that a significant amount of effort has been applied in defining the protocol called Multi-Protocol Label Switching (MPLS). MPLS is similar to DiffServ in the sense that it marks traffic at the entrance to the network. However, the primary function of the marking is not to allocate priority within a router, but to determine the next router in the path from source to destination. MPLS involves the attachment of a short label to a packet in front of the IP header, which enables faster routing decisions to be made at each router. The label identifies a Forward Equivalence Class (FEC), which means that all packets with a given FEC will be treated the same way as far as forwarding is concerned. The packets for a voice stream for example would all belong to the same FEC and would, thus, receive the same forwarding treatment. Therefore, we can ensure that the forwarding treatment applied to a given stream can be set up in such a way so that all packets from points A to B, for example, follow exactly the same path. If the stream requires a specific bandwidth, then that bandwidth can be allocated at the start of the session. This feature can ensure that a given stream receives the bandwidth it requires, and that the packets of this stream arrive to their destination in the order in which they were transmitted.

Chapter 4

Dynamic Bandwidth Representation for Network Traffic

4.1 Introduction

This chapter deals with the first of our objectives as outlined in Chapter 1, namely, the construction of DWT based bandwidth signals for the effective representation of network traffic bandwidth at a network node output. In particular, two new classes of bandwidth signals are presented. The first class, called the Local Maximum Bandwidth Models (LMBM) and denoted $\mathcal{L}$, consists of piecewise constant signals, where each of these piecewise constant signal is constructed from more elementary constant signals with fixed support of length ρ. The value of a particular elementary constant signal is obtained by taking the maximum of the ρ most recent values of the traffic plus some scalar multiple α of the traffic's long-term mean μ_k. In this class, each signal is uniquely determined by α, ρ and $(\mu_k)_{k\in K}$, where K is defined in the following section. The second class of signals, called the Detail Maximum

Bandwidth Models (DLMBM) and denoted $\mathcal{D}$, is similar in construction to $\mathcal{L}$. However, here, the value of each elementary constant signal is obtained by taking some constant multiple of the maximum of the ρ most recent values of the detail signal obtained through the DWT of the traffic, plus the traffic's long-term mean up to some point.

4.2 The Class of Local Maximum Bandwidth Signals

This section presents the class of Local Maximum Bandwidth Models (LMBM), which we denote $\mathcal{L}$. Like the EDB in the last chapter, the goal of an LMBM signal is to provide a means of representing the traffic bandwidth on the output link of a network node. Let the traffic be denoted by $x = (x_1, x_2, \cdots, x_n)$. Define the sets $F_k = \{k-\rho+1, k-\rho-2, \cdots, k\}$ and $K = \rho\mathbb{N}^* \cap \Lambda_n$, and the long term mean of x by $\mu_k = \sum_{i=l}^{k} x_i/(k-l+1)$, $k \in K$, $l \in \mathbb{N}^*$. Let $\tilde{x} = (x_i)_{i \in F_k}$ and $b_k = \max \tilde{x}$. The objective is to determine the bandwidth for the ρ traffic samples $\tilde{x}$ for each $k \in K$. Ideally, if $\tilde{x}$ were available, the best bandwidth would have been $\tilde{x}$ itself for each $k \in K$. However, in a real-life situation this might not be the case. Thus, a means of approximating the bandwidth becomes necessary. One approximation we use is to represent the traffic bandwidth as a function of the maximum of $\tilde{x}$ for each $k \in \rho\mathbb{N}$. More specifically, define the operator $B_{\rho,\alpha,k} : \mathbb{Z}^\rho \to \mathbb{Z}^\rho$ by

$$(\forall k \in K) \quad B_{\rho,\alpha,k}\tilde{x} = (b_k + \alpha\mu_k) \cdot 1_\rho, \tag{4.1}$$

for some $\rho \in K$ and $\alpha \in \mathbb{R}$. Then a signal $B \in \mathcal{L}$ is given by $B = (B_{\rho,\alpha,k}\tilde{x})_{k \in K}$, for some $\rho \in 2^{\mathbb{N}^*}$ and $\alpha \in \mathbb{R}$. Since the sequence, $B_{\rho,\alpha,k}\tilde{x}$, is constant on each F_k, then B is piecewise constant on Λ_n. Note also that the parameter ρ is the bandwidth update

parameter, i.e., it is the number of traffic samples for which each $B_{\rho,\alpha,k}\tilde{r}$ is held constant. Thus, as ρ increases, so to does the time-duration for which $B_{\rho,\alpha,k}$ is held, that is, the traffic bandwidth is updated less often. This is illustrate in Figure 4.1 for various values of ρ with $\alpha = 0$. In a real-life situation, the value of ρ is dependent on the traffic arrival rate.

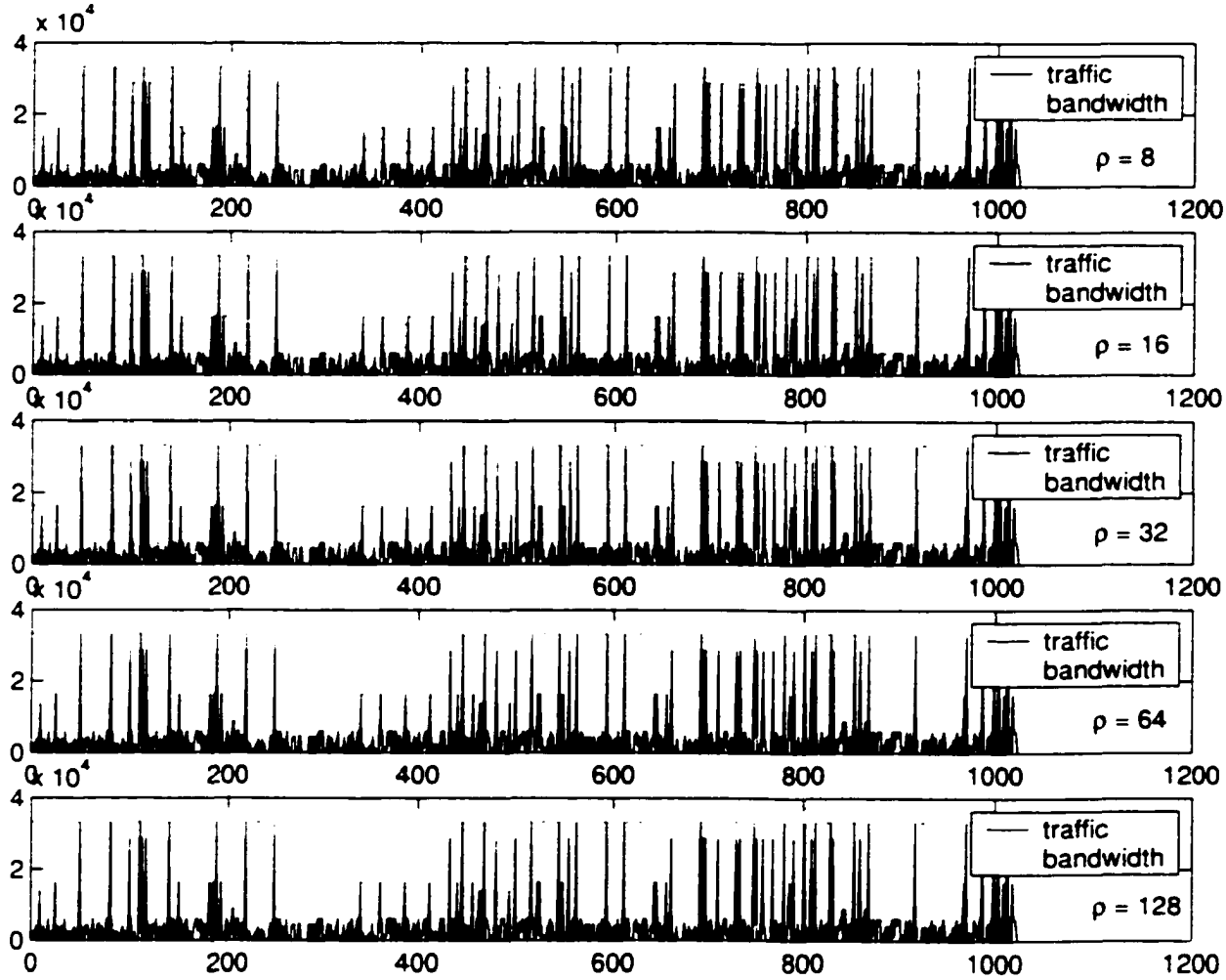


Figure 4.1: LMBM

processor speed, desired QoS and utilization. Since communication speed has in the large

exceeded processor speed [62]. ρ should be chosen sufficiently large to compensate for this difference, especially during times of high traffic volume.

The role of α is to control the average value of the bandwidth signal. As α increases, the average value of $B_{\alpha,\rho}$ also increases and vice versa. Figure 4.2 is an illustration for the case

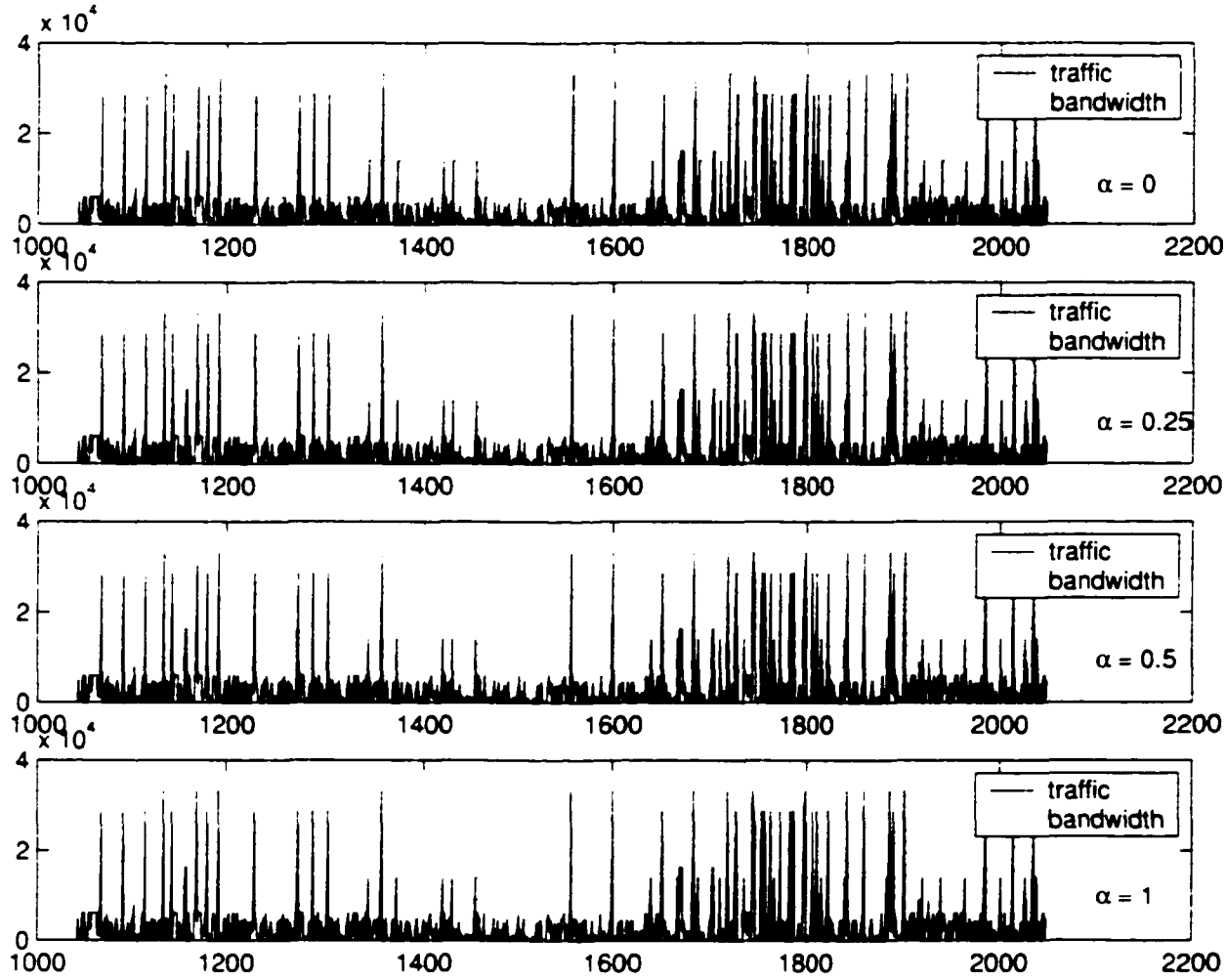


Figure 4.2: LMBM

when ρ is held constant at eight and α allowed to vary. Here we see that as α increases, both cell loss and utilization decrease. Although this is desirable for the cell loss, it is not for the utilization. This implies that there will be a need for a tradeoff between QoS and utilization.

Note that for the mean μ_k to be truly long-term, then k must be sufficiently large and l appropriately chosen. For this reason, we start with $k = 1025$ and $l = 1$. Future values of k are chosen such that $k - l + 1 = 1024$.

Although 4.1 is not explicitly connected to the DWT, it will in the real-life situation. There, it will become necessary to predict $\tilde{x}$. However, the approach we will use is to predict the wavelet coefficients of $\tilde{x}$, and from these get a prediction of $\tilde{x}$. There lies the connection!

4.3 Results

Results for the LMBM and EDB are illustrated in Figures 4.3 and 4.4, respectively. Figure 4.5 shows the resulting cell loss for each formula with $\rho = 32$. It is clear that for the EDB, the cell loss increases rapidly as the traffic load increases. While for the LMBM, the cell loss is zero for each load.

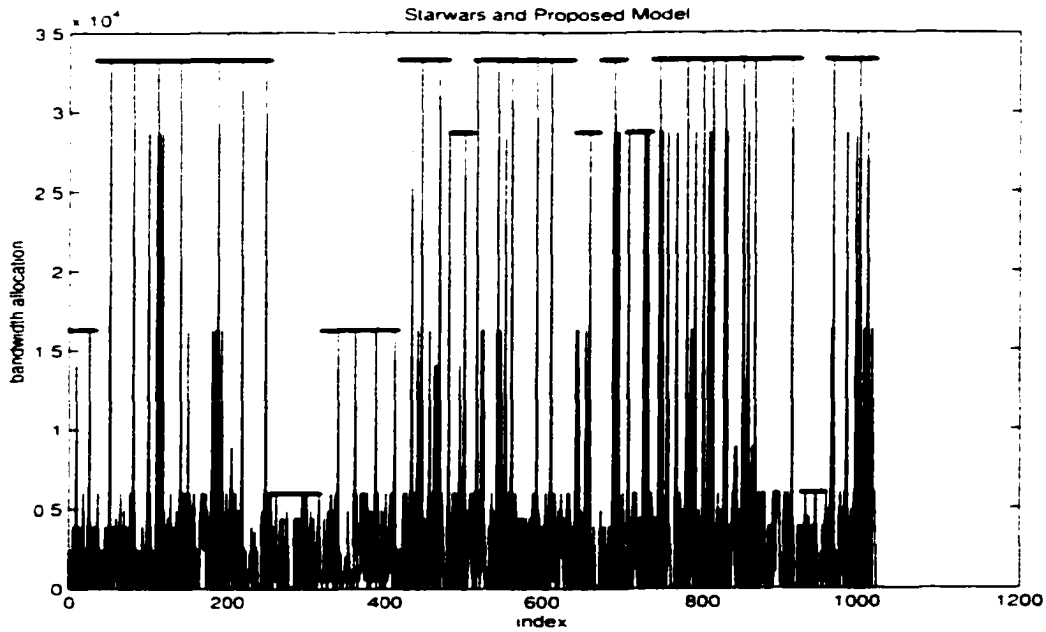


Figure 4.3: LMBM

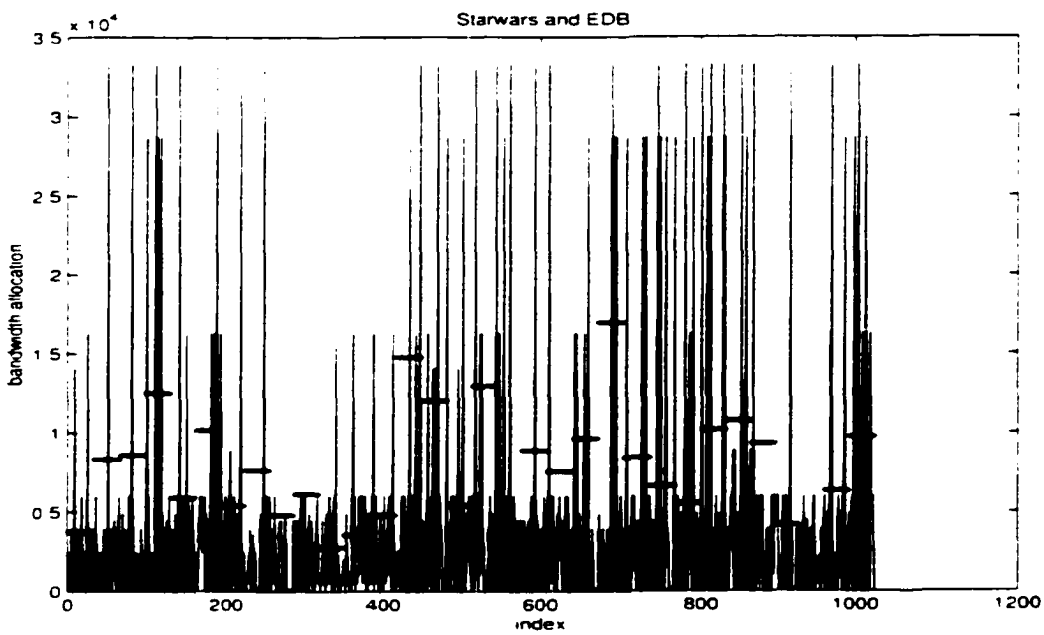


Figure 4.4: EDB with LBL-TCP-3

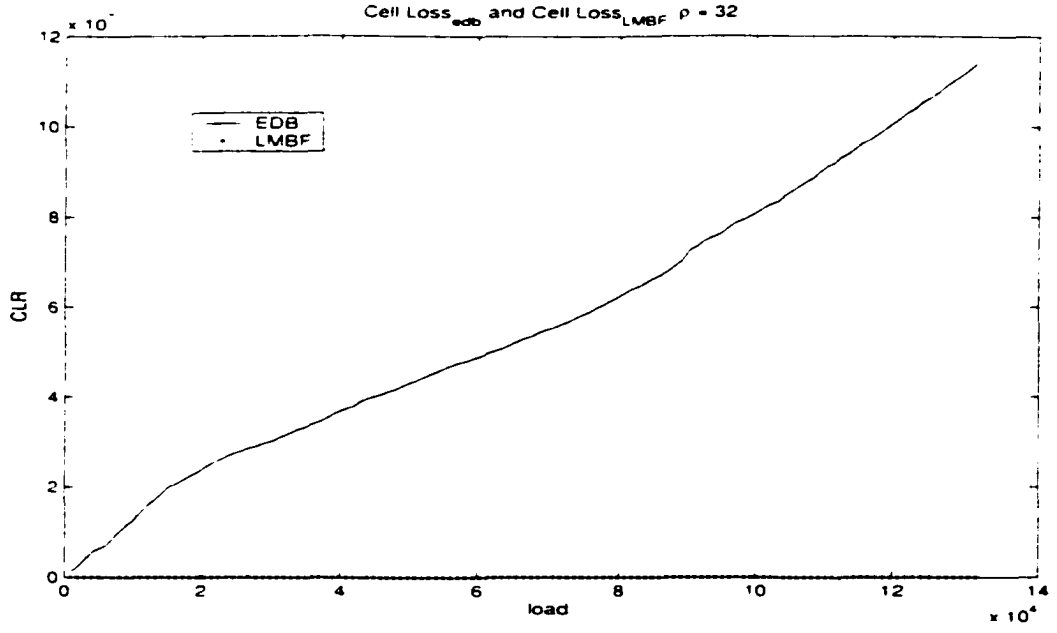


Figure 4.5: CLR: EDB and LMBM

4.4 The Class of Detail Local Maximum Bandwidth Models

We now introduce a second class of bandwidth models—the class of Detail Local Maximum Bandwidth Models (DLMBM)— needed in the first step of the design of our overall bandwidth allocation scheme. Assume the traffic to be the number of packet arrivals x_i during the i-th time slot of duration $T_0 \in \mathbb{R}^+$ of the measurement. Let $I_n = (1, 2, \cdots, n)$, for some $n \in \mathbb{N}^*$, then the traffic sequence may be written as $x = (x_i)_{i \in I_n}$. Let $\dot{x} = (x_{k-\lambda-1}, x_{k-\lambda-2}, \cdots, x_k)$, for some $\lambda < k$, $\lambda \in \mathbb{N}^*$. The class, $\mathcal{D}$, to be constructed shall be based on the detail signal, $D_1\dot{x}(t) = \sum_l d_{\dot{x}}(1, l)\psi_{1,l}(t)$, where $d_{\dot{x}}(1, \cdot)$ are the detail wavelet coefficients of $\dot{x}$ at level 1. Recall that the signal $D_1\dot{x}$ consists of the high frequency components of the traffic, i.e., the SRD in the traffic. Future values of such a

signal are therefore likely to be similar to present and recent past values. The DLMBM will exploit this concept.

Let $K = \rho\mathbb{N} = \{\rho, 2\rho, 3\rho, \cdots, n\}$ for some $\rho \in \mathbb{N}$. For every $k \in K$, each bandwidth model $B = (B_i)_{i \in I_n}$ in $\mathcal{D}$ shall consists of piecewise constant sequences $B_k^\rho = (B_{k-\rho+1}, B_{k-\rho+2}, \cdots, B_k)$, where for each $i, j \in \{k-\rho+1, k-\rho+2, \cdots, k\}$, $B_i = B_j$ for some $\rho \in \mathbb{N}$. Thus, we also have $B = (B_k^\rho)_{k \in K}$. The *bandwidth update parameter* ρ is simply the number of time slots for which B_k^ρ persists, i.e., ρT_0 is the time-duration for which $b^{(k)}$ is maintained. Thus, to obtain $(B_i)_{i \in I_n}$, it is necessary and sufficient to obtain $(B_k)_{k \in K}$. The B_k's may be realized through the functional $\mathcal{B}_{\gamma,k} : E^\lambda \to \mathbb{R}$ given by

$$(\forall k \in K) \quad \mathcal{B}_{\gamma,k}\tilde{x} = \mu_k + \gamma \max D_1 \tilde{x} \tag{4.2}$$

for some $\gamma \in \mathbb{R}$. Indeed, B_k is defined formally as $B_k = \mathcal{B}_{\gamma,k}\tilde{x}$. The space $\mathcal{D}$ is then given by

$$\mathcal{D} = \{B = (B_k)_{k \in K} = (\mathcal{B}_{\gamma,k}\tilde{x})_{k \in K} \mid \gamma \in \mathbb{R}, \rho \in \mathbb{N}\}. \tag{4.3}$$

Figures 4.6 and 4.7 illustrate the behavior of B with ρ and γ, respectively, when $\mu_k = 0$. In Figure 3, γ is held constant so as to study the behavior of ρ on the bandwidth B. Note that as ρ increases, the bandwidth update interval, $T = \rho T_0$, increases—the bandwidth updates less frequently. Conversely, decreasing the value of ρ causes more frequent bandwidth updates to occur. The role of γ is to control the average value of the bandwidth signal, B. As γ increases, the average value of B also increases and vice versa. Figure 4.2 is an illustration for the case when ρ is held constant at 32, while γ varies. Here we see that as γ increases, both cell loss and utilization decrease. Although this is desirable for the cell

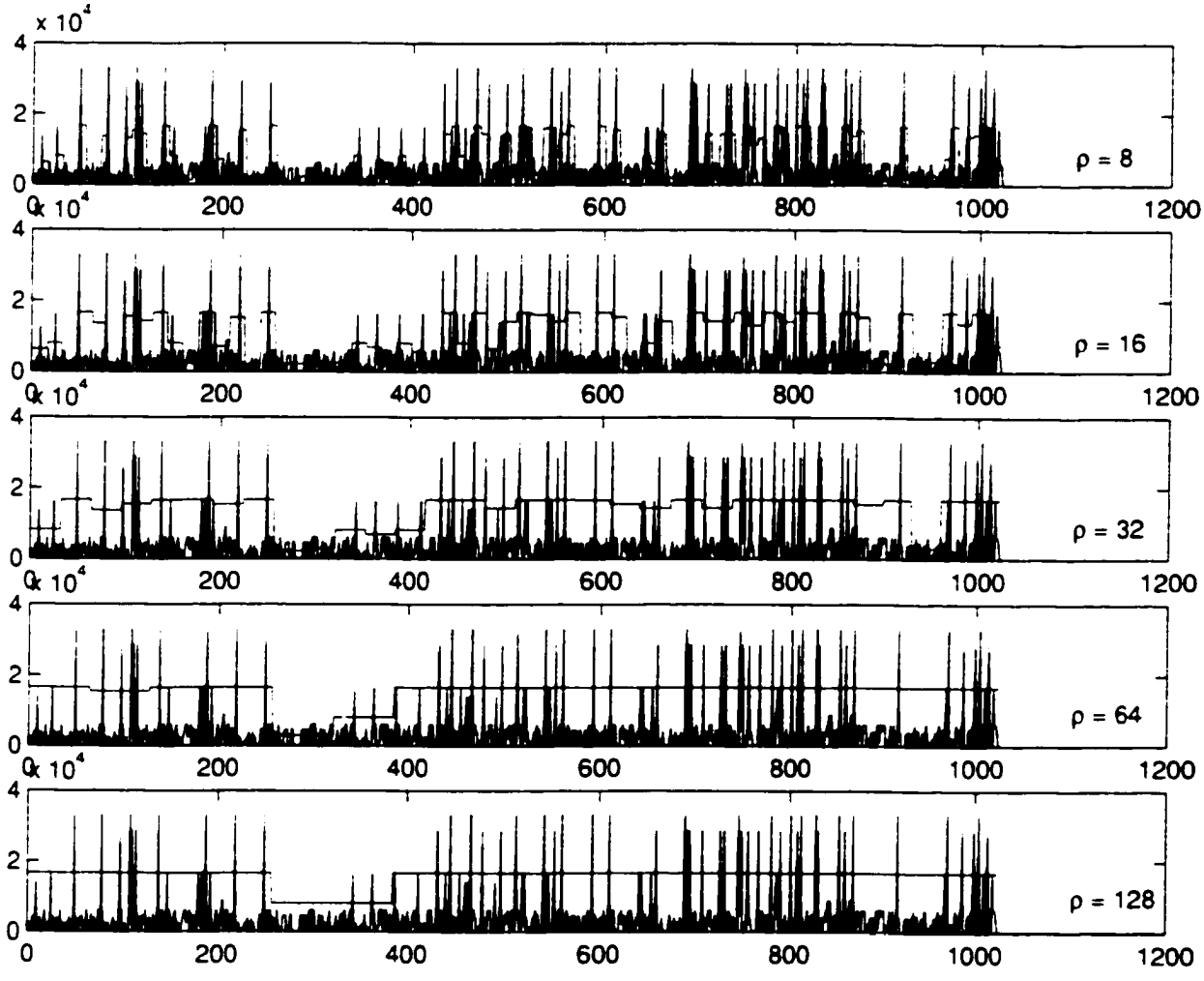


Figure 4.6: Effect of parameter ρ on B

loss, it is not for the utilization. This implies that there will be a need for tradeoff between QoS and utilization.

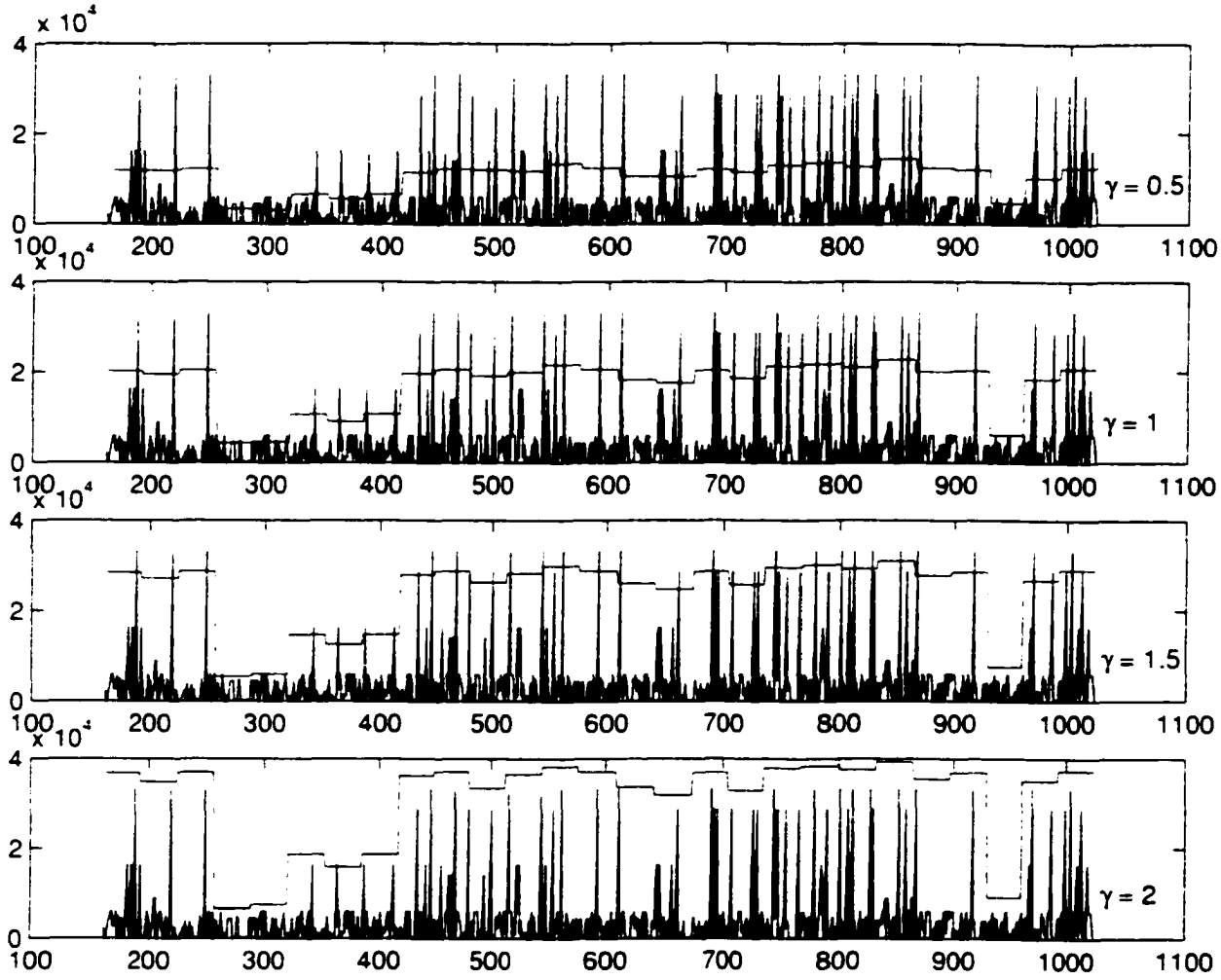


Figure 4.7: Effect of parameter γ on B

The DLMBM will be our bandwidth model of choice in our simulations.

4.5 Bandwidth Modelling for Voice Over IP

Section 3.6 discussed several QoS methods for VoIP. These methods are the Resource-Reservation Protocol (RSVP), the Differentiated Service (DiffServ) protocol and Muli-Protocol Label Switching (MPLS). In this section, we provide a high-level discussion of how the bandwidth B_k may be determine using each of the methods. We start with the DiffServ method and in particular the Expedited Forwarding (EF) service. Recall that with EF, packets from a given traffic stream is assigned a minimum departure rate that is greater than the arrival rate at the same node, provided that the arrival rate does not exceed a pre-agreed upon maximum. Thus, each bandwidth value B_k may be obtained by choosing α such that $B_k > x_k$ for each k, provided that $x_k < W$ for some pre-agreed maximum value, W if that bandwidth is available. Otherwise, non-real time traffic may be buffered until this can be accomplished. For AF, the procedure would be similar. Recall that with AF packets from a given traffic source are forwarded with a high probability, provided that the traffic from that source does not exceed some pre-assigned maximum. Thus, B_k may be chosen so that x_k is forwarded with probability one if there are no other traffic at the node. Otherwise, let B_k be chosen such that x_k is forwarded with the highest possible probability among the various traffic sources.

For the Resource-Reservation Protocol (RSVP), since the PATH message provides information such as bandwidth and packet length, we may obtain bandwidth values B_k in advance from the PATH message itself, which would then override the need to calculate these values.

For the MPLS method we may choose α based on the FEC type, i.e., the kind of traffic being transmitted. Voice traffic, for example, should be allocated sufficient bandwidth so that the need for buffering should be virtually non-existent.

4.6 Summary

This chapter introduces two classes of DWT based bandwidth signals for the representation of traffic bandwidth at a network node. The first is the class of Local Maximum Bandwidth Signals, $\mathcal{D}$. In this class, each signal is obtained by $B = (B_{\rho,\alpha,k}\vec{x})_{k\in K}$, where the operator $B_{\rho,\alpha,k}$ is given by $B_{\rho,\alpha,k}\vec{x} = (\max \vec{x} - \alpha\mu_k) \cdot 1_\rho \, \rho \in \rho\mathbb{N}^*$. The parameter ρ is the bandwidth update interval and α is a parameter which controls the mean of the bandwidth signal. The second class of signals is the class of Detail Local Maximum Bandwidth Signals. It is defined by $B = (T_{\rho,\alpha,k}\vec{x})_{k\in K}$ where the operator $T_{\rho,\alpha,k}$ is given by $T_{\rho,\gamma,k}\vec{x} = (\mu_k - \gamma \max_{t\in F_k} D_1\vec{x}_t) \cdot 1_\rho$. The role of parameter γ is similar to that of α.

Chapter 5

Bandwidth Prediction Algorithms

5.1 Introduction

The previous chapter presented the first of our two-part solution: namely, signals for the modelling of the traffic bandwidth. The current chapter takes up the issue of how to predict the wavelet coefficients needed by each bandwidth signal in the case of a practical situation. In particular, two prediction algorithms are presented. The first is a wavelet-based regression model and the second an Independent Wavelet Model [10] based prediction algorithm. Recall that in the practical situation, we wish to assign a bandwidth to the traffic $(x_t)_{t\in F_k}$ before it arrives at the network node. However, the bandwidth signals proposed in the last chapter are based on $(x_t)_{t\in F_k}$ themselves. Since the $(x_t)_{t\in F_k}$ are not immediately available, we propose to estimate them through prediction. Let the prediction of $(x_t)_{t\in F_k}$ be denoted $\hat{x}_{t\in F_k}$. Thus, what we are attempting to do is to predict $\rho \in \mathbb{N}^*$ values of the traffic given $\lambda \in \mathbb{N}^*$ of its past values, with $\lambda \gg \rho$. Suppose we are given the

traffic $\tilde{x} = (x_1, x_2, \cdots, x_{k-2\rho-\lambda+1})$. Since the bandwidth must be predicted in advance, then it must be assumed that the bandwidth for $(x_{k-2\rho+1}, x_{k-2\rho+2}, \cdots, x_{k-\rho})$ has already been predicted (see Figure ??), and that the goal now is to predict the bandwidth for $(x_t)_{t\in F_k} = (x_{k-\rho+1}, x_{k-\rho+2}, \cdots, x_k)$ before $x_{k-\rho}$ is received. Let the long term mean be given by $\mu_{k-2\rho-\lambda+1} = \sum_{t=l}^{k-2\rho-\lambda+1} x_t$. Then, from the form of the bandwidth signals of the last chapter, we only need to predict the maximum of $(x_t)_{t\in F_k}$. That is, we only need to predict b_k.

Recalling also the long-range dependent nature of network traffic and the short-range dependent nature of its wavelet coefficients, we propose to predict b_k by first predicting the wavelet coefficients for the traffic $(x_t)_{t\in F_k}$, performing an inverse wavelet transform on these predicted coefficients, to get predicted traffic $(\hat{x}_k)_{k\in F_k}$, and then taking the maximum of the $\hat{x}_k$'s. In other words we wish to find a predictor $P : \Xi^\lambda \to \mathbb{R}$ such that $P\tilde{x} = \hat{b}_k$, where $\hat{b}_k$ is as close to b_k as possible in the ℓ^2 sense without unduly sacrificing QoS.

5.2 The ρ-Step Linear Predictor

This section introduces the ρ-step Wavelet Domain Linear Predictor (ρSWD-LP) to predict wavelet coefficients needed by the LMBM. This algorithm is basically an application of the autoregressive (AR) model on each level of a ρ-level discrete wavelet decomposition. Recall that a linear predictor of order p aims to predict the next value of process of length n in

terms of its p most recent values:

$$\hat{x}_{n+1} = \sum_{i=0}^{p-1} \phi_i x_{n-i} = \phi_0 x_n + \phi_1 x_{n-1} + \cdots + \phi_m x_{n-p+1}. \tag{5.1}$$

where the coefficients $(\phi_i)_{i \in \Lambda_p}$ are obtained via the classical Levingston-Durbin Algorithm [30]. Let $L = log_2(\rho)$ and

$$\hat{x} = (x_{k-\lambda-\rho+1}, x_{k-\lambda-\rho+2}, \cdots, x_{k-\rho}) \tag{5.2}$$

be the current buffered traffic with discrete scaling coefficient at level L and wavelet coefficients up to level L given by

$$\{a_{\hat{x}}(L, \cdot), \{d_{\hat{x}}(1, \cdot), d_{\hat{x}}(2, \cdot), \cdots, d_{\hat{x}}(L, \cdot)\}. \tag{5.3}$$

For each level $j \in \Lambda_L$, let the order of the AR process be $\lambda/2^j$. We predict future wavelet coefficients for each j by approximating Equation 5.1 by a $\rho/2^j$ predictor on $d_{\hat{x}}(j, \cdot)$ as follows:

$$(\forall j \in \Lambda_L)(\forall i \in \Lambda_{\rho\ 2^j}) \quad d_{\hat{x}}(j, i) = \sum_{l=0}^{\lambda\ 2^j} \phi_l d_{\hat{x}}(j, (\lambda + \rho)/2^j - l - i). \tag{5.4}$$

Estimating $a_{\hat{x}}(\log_2(\rho), \cdot)$ by $a_{\hat{x}}(\log_2(\rho), \cdot)$, the inverse wavelet transform may be performed on $\{a_{\hat{x}}(\log_2(\rho), \cdot), \{d_{\hat{x}}(1, \cdot), d_{\hat{x}}(2, \cdot), \cdots, d_{\hat{x}}(\log_2(\rho), \cdot)\}$ to obtain predicted traffic $\hat{x}_{k-\rho-1}, \hat{x}_{k-\rho+2}, \cdots, \hat{x}_k$. The $\hat{b}_k$ is then obtained as $\hat{b}_k = \max_{i \in \Lambda_{\rho,k}} \{\hat{x}_i\}$. The procedure is illustrated in Figure 5.1.

More formally, the ρSWD-LP is given as follows:

Algorithm:

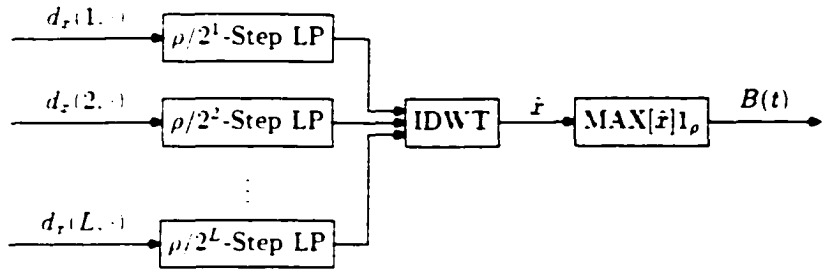


Figure 5.1: ρ-step Wavelet-Domain Linear Predictor

Step 1: Let $\hat{x}$ be given:

Step 2: Compute an L-level wavelet decomposition on $\hat{x}$ to get scaling coefficients $a_{\hat{x}}(L,\cdot)$ and wavelet coefficients $\{d_{\hat{x}}(j,\cdot) : j = 1,\cdots,L\}$:

Step 3: Set $j = 1$ and $i = 1$:

Step 4: Obtain $d_{\hat{x}}(j,i)$ via Equation 5.4:

Step 5: $i = i + 1$:

Step 6: If $i \in \Lambda_{\rho,j}$, goto Step 4:

Step 7: $j = j + 1$:

Step 8: If $j \in \Lambda_L$, set $i = 1$, goto Step 4:

Step 9: Get $(\hat{x}_t)_{t\in I_{k,\rho}}$:

Step 10: Get $(B_t)_{t\in I_{k,\rho}}$ via Equation 4.1:

Step 11: While $k \leq n$, goto Step 1.

5.3 The ρ-Step IWM Predictor

From the previous chapter, we learned that wavelet coefficients needed by the LMBM and DLMBM are not readily available in a practical situation; and must therefore be predicted. It is for this purpose that we present the $\rho-$Step IWM Prediction (ρSIP) algorithm. Let $s = 2^{\log_2(\rho)-j}+i$, $j \in \mathbb{N}^*, i \in \mathbb{N}$ and let $L = \log_2(\rho)$. Let $\tilde{x} = (x_{k-(\lambda+\rho)+1}, x_{k-(\lambda+\rho)+2}, \cdots, x_{k-\rho})$ be given, for some $k \in K$ with $k > \lambda + \rho$. Then the ρSIP algorithm is given as follows:

Step 1: Compute the $\log_2(\rho)$ multi-level wavelet decomposition of $\tilde{x}$ to get $\{a_{\tilde{x}}(L,1), \{d_{\tilde{x}}(j,\cdot) : j = 1,2,\cdots,L\}\}$:

Step 2: Set j=1:

Step 3: Compute the sample variance v_j of $d_x(j,\cdot)$:

Step 4: For $i = 0,1,\cdots,n_j - 1$, predict wavelet coefficients at level j via the IWM equation $d_{\tilde{x}}(j,i) = v_j w_s$, where w is a *white noise* process:

Step 5: While j not equal to L, set $j = j + 1$, goto *Step 3*:

Step 6: Set $\hat{d} = \{a_x(L,1), \{d_{\tilde{x}}(j,\cdot) : j = 1,2,\cdots,L\}\}$

5.4 Summary

This chapter introduced the ρ-step Linear Predictor—Algorithm 1. Section 5.2—and the ρ-step IWM predictor—Algorithm 2. Section 5.3. Each algorithm serves the same purpose, namely, to predict the wavelet coefficients of the next ρ traffic samples to arrive given its λ most recent past values.

Chapter 6

Bandwidth Allocation Algorithms

6.1 Introduction

This chapter proposes two bandwidth allocation algorithms for real-time network traffic. The first is the Consecutive Max Bandwidth Allocation Algorithm, which uses the Class of Detail Maximum bandwidth Models, Subsection 4.4, and is based on the traffic's past behavior. The second is the Independent Wavelet Bandwidth Allocation Algorithm, which is also based on the Class of Detail Maximum bandwidth Models, but uses the prediction algorithm ρSIP, Subsection 5.3. *These bandwidth allocation algorithms are examples of the general algorithm discussed in Chapter 1, Subsection 1.3.*

The aim of a bandwidth algorithm is to estimating the network traffic bandwidth ahead of time before the traffic arrives at the network node output. More specifically, given λ past

values

$$(x_{k-\lambda-\rho+1}, x_{k-\lambda-\rho+2}, \cdots, x_{k-\rho}) \tag{6.1}$$

of the traffic x estimate the bandwidth, B, for the ρ future traffic samples $(x_k, x_{k-1}, \cdots, x_{k-\rho})$ of x ahead of time, where $\rho, \lambda \in \mathbb{N}$. Without lost of generality, ρ and λ will be assumed to be powers of 2 with $\lambda > 2\rho$ a multiple of ρ. Recall that to compute bandwidth $B = (B_1, B_2, \cdots, B_n)$ it is necessary and sufficient to compute only the subsequence $(B_\rho, B_{2\rho}, \cdots, B_n)$.

6.2 The Consecutive Max Bandwidth Allocation Algorithm

This section introduces the first of our bandwidth allocation algorithms—the *Consecutive Max Bandwidth Allocation Algorithm* (CMBAA). The algorithm is shown in block diagram form in Figure 6.1. As mentioned earlier, the CMBAA is a special case of the integrated

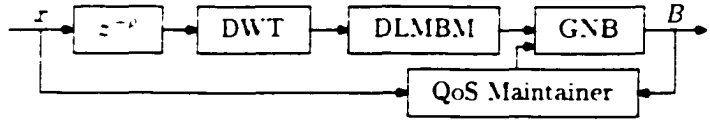


Figure 6.1: The Consecutive Max Bandwidth Formula

bandwidth allocation algorithm given in Subsection 1.3. The Get New Bandwidth (GNB) function is defined by Step 5 to 7 in formal statement of the CMBAA given at the end of this section. We now discuss this CMBAA in detail. Let us assume that the first $c = l/\rho$ values $(B_\rho, B_{2\rho}, \cdots, B_\lambda)$ of this subsequence are given. That is, an initial bandwidth, $B_1, B_2, \cdots, B_\lambda$ is pre-allocated on the output link of the network node for the anticipated

traffic, $x_1, x_2, \cdots, x_\lambda$, at time $t = 0$ say (see Figure 6.2). The choice of these initial B_i's

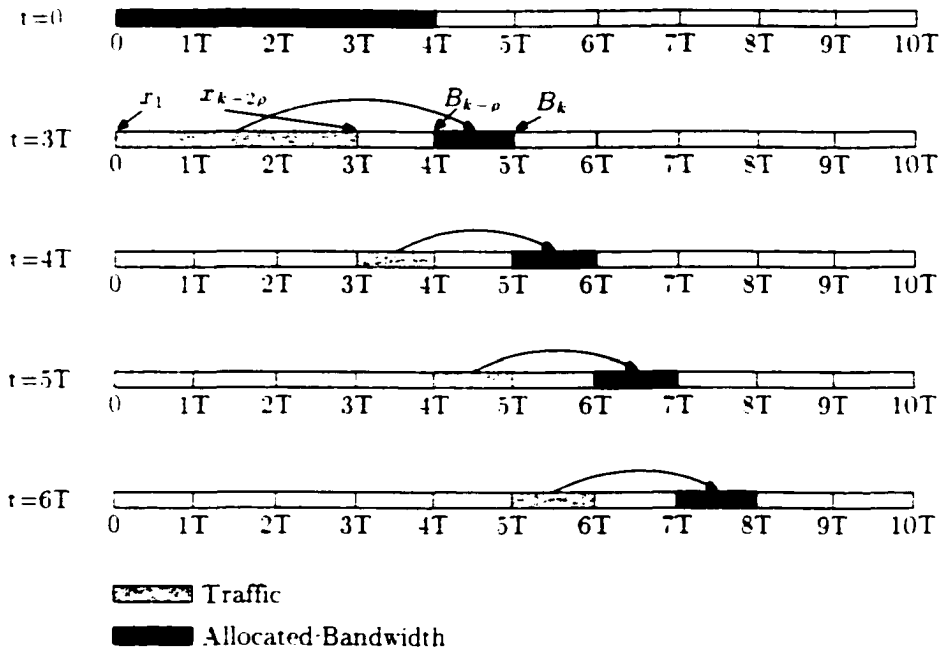


Figure 6.2: Bandwidth Allocation

may be based on some preliminary estimate of traffic or on *a priori* traffic information. As the traffic samples are arriving their values are copied into a buffer which uses a first-in first-out (FIFO) discipline. (*We wish to emphasize here that the traffic samples themselves are not held in this buffer — only a copy of each is*). Let $k = \lambda + \rho$, and make the following assignments at time λT_0:

$$w = (x_i)_{i \in \Lambda_{t-\rho}} \tag{6.2}$$

$$B_{temp} = B_{\gamma,k} w \tag{6.3}$$

$$B_k = B_{temp} \tag{6.4}$$

In the interim while these computations are being made, additional traffic samples,

$$(x_{k-2\rho+1}, x_{k-2\rho+2}, \cdots, x_{k-\rho}), \tag{6.5}$$

or at least a subset of them, are likely to arrive. These would also be copied to the buffer. Note that B_k is the bandwidth for this traffic. At this point, only ρ samples of the traffic will be needed to make the next bandwidth projection B_{k+1}. Thus the FIFO buffer may now be assumed to be of length ρ, so that an equivalent amount of traffic values are dropped from the buffer as the amount copied to it. This means that only ρ traffic samples are maintained in the buffer at any given time. To ensure that the bandwidth is assigned for the next ρT_0 time slots before the traffic arrives, the computation times for the assignments in Equations (6.2) through (6.4) will be assumed to be less than the value of ρT_0.

The bandwidth value $B_{k+\rho}$ for the next time slot is obtained by splitting the sequence $x^{(\rho)} = (x_{k-2\rho+1}, x_{k-2\rho+2}, \cdots, x_{k-\rho})$ into two separate sequences each of length $\rho/2$. The first sequence $u^{(1)} = (x_{k-2\rho+1}, x_{k-2\rho+2}, \cdots, x_{(2k-3\rho)/2})$ consists of the first $\rho/2$ values of $x^{(\rho)}$. The second sequence consists of the remaining $\rho/2$ values of $x^{(\rho)}$, i.e. the latter part of $x^{(\rho)}$, and is given by $u^{(2)} = (x_{(2k-3\rho)/2+1}, x_{(2k-3\rho)/2+2}, \cdots, x_{k-\rho})$. Thus, we have $x^{(\rho)} = (u^{(1)}, u^{(2)})$. Let $b^{(1)} = \mathcal{B}_{\gamma,k} u^{(1)}$ and $b^{(2)} = \mathcal{B}_{\gamma,k} u^{(2)}$. If $(b^{(1)} - B_{temp}) > \tau$ and $(b^{(2)} - B_{temp}) > \tau$, then set $B_{temp} = \max\{b^{(1)}, b^{(2)}\}$ and $B_{k+\rho} = B_{temp}$. Otherwise, set $B_{k+\rho} = B_{temp}$. Set $k = k + \rho$ and update the bandwidth using the updated buffer starting with the computation of the $u^{(1)}$ and $u^{(2)}$. Continue to update the bandwidth in this way as new traffic arrives. The algorithm is illustrated pictorially in Figure 6.1.

The algorithm is given formally as follows: *Algorithm*:

Algorithm:

Step 1: Let tolerance τ be given and assume that $B_1, B_2, \cdots, B_{k-\rho}$ have been pre-allocated for the first $\lambda = l\rho$ anticipated traffic values $(x_1, x_2, \cdots, x_\lambda)$ of x.

Step 2: Let $k = \lambda - \rho$ and $w = (x_1, x_2, \cdots, x_{k-2\rho})$;

Step 3: Let $B_{temp} = \mathcal{B}_{\gamma,k} w$;

Step 4: Set $B_k = B_{temp}$;

Step 5: Let

$$w^{(1)} = (x_{k-2\rho+1}, x_{k-2\rho+2}, \cdots, x_{(2k-3\rho)/2}) \quad \text{and}$$

$$w^{(2)} = (x_{(2k-3\rho)/2+1}, x_{(2k-3\rho)/2+2}, \cdots, x_{k-\rho})$$

Step 6: Set $b^{(1)} = \mathcal{B}_{\gamma,k} w^{(1)}$ and $b^{(2)} = \mathcal{B}_{\gamma,k} w^{(2)}$

Step 7: If $[(b^{(1)} - B_{temp}) > \tau]$ && $[(b^{(2)} - B_{temp}) > \tau]$

$B_{temp} = max\{b^{(1)}, b^{(2)}\}$

$B_{k-\rho} = B_{temp}$

Else

$B_{k-\rho} = B_{temp}$

Step 8: While $k < n$

$k = k + \rho$

Goto Step 5

6.3 Results

In this section, we present the results of our algorithms for the wide area TCP trace, LBL-TCP-3[63] - a portion of which is illustrated in Figure 6.4 with its allocated bandwidth. This trace contains two hours of all wide-area TCP traffic between Lawrence Berkeley Laboratory and the rest of the world. It is freely available on the Internet from the Internet Traffic Archive (*http://ita.ee.lbl.gov/html/traces.html*). We note that although the trace is actually given in bytes/sec, the subsequent results also holds true for traffic given in cells/sec or packets/sec. This is because the number of packets or cell per second is directly proportional to the number of bytes per second. We use the simplified ATM

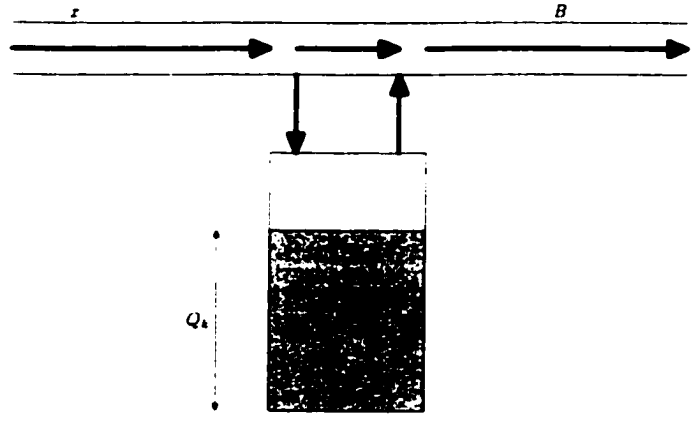


Figure 6.3: ATM Simulator

simulator [2] illustrated in Figure 6.3 to model the traffic at a network node. Here the traffic r arriving at a network node continues to the node output provided that the bandwidth is large enough to accommodate it. $B_i \geq x_i$, $i \in \{k-\rho+1, k-\rho+2, \cdots, k\}$ for some $k \in \rho\mathbb{N} \cap \Lambda_n$. If the traffic exceeds the bandwidth, then traffic samples $D_i = x_i - B_i$, $i \in$

$\{k-\rho+1, k-\rho+2, \cdots, k\}$ are dropped into the FIFO buffer of size M shown. Let Q_{k-1} be the number of cells in the buffer at time $k-1, k \in K$, then the buffer state at time k is given by $Q_k = \min\{M, \max\{0, D_k + Q_{k-1}\}\}$. We test the performance of the proposed

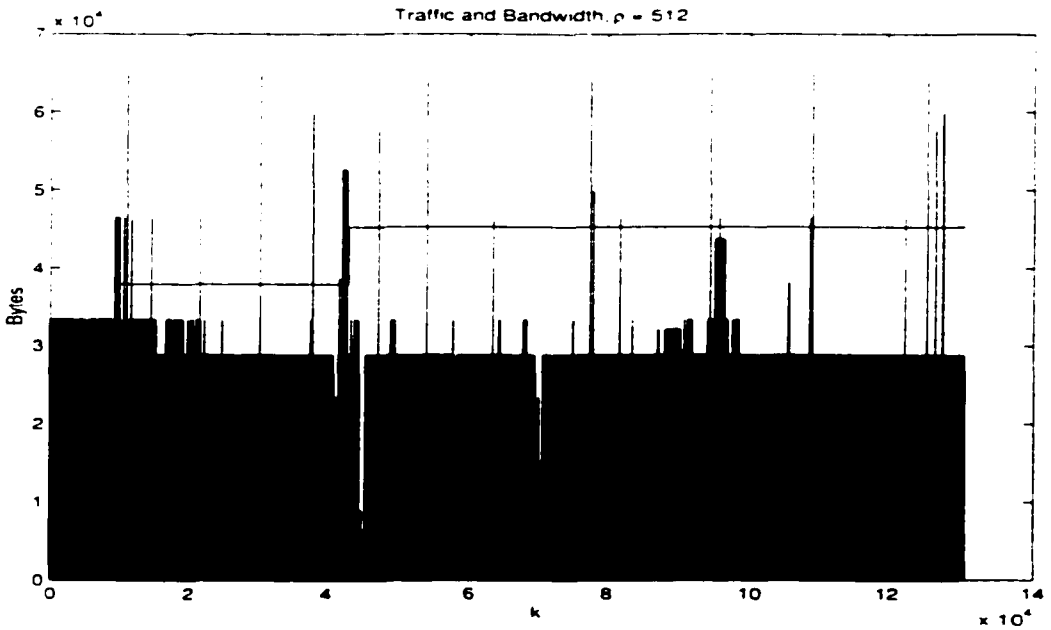


Figure 6.4: Traffic and Bandwidth

algorithm by applying it to the ATM simulator shown in Figure 6.4. The following two QoS parameters are then computed in order to test the performance of the algorithm: the *cell loss ratio* (CLR), defined as the *number of loss cells divided by the number of transmitted cells*. The *maximum cell delay* (MCD), defined as the maximum number of time units for which a cell is held in the buffer. We also compute the utilization, $u = \sum_{i=1}^{n} x_i / \sum_{i=1}^{n} B_i$. A utilization of one when cell loss is zero means that all the allocated bandwidth is actually used; a utilization less than one, means that the bandwidth is actually under-utilized; while a utilization greater than one means that some cell/packet loss have occurred because of

the traffic exceeding the bandwidth at various times.

Results for the proposed algorithm with the ATM simulator are as follows. Figure 6.4 shows the traffic LBL-TCP-3 overlapped with the allocated bandwidth, B. Figure 6.5 displays CLR versus the buffer length, M, with $\gamma = 1.4$, $\rho = \lambda = 512$ and $\tau = 50$. Clearly, CLR decreases as the buffer length increases as one would expect. Figure 6.6 illustrates the

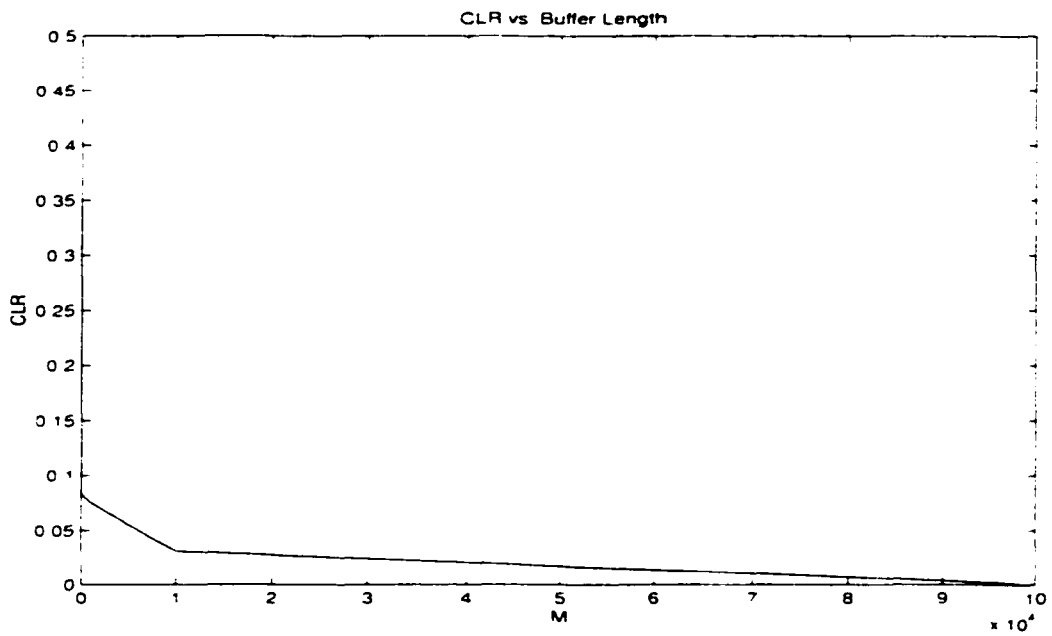


Figure 6.5: CLR vs. M

MCD versus M, again with $\gamma = 1.4$. In contrast to the CLR, the MCD increases with the buffer length M. This is consistent with previous results [2] acknowledging that increasing buffer length by itself may not necessarily improve overall quality of service and may in fact increase delay in the network.

Figures 6.7 and 6.8 show the effect of the parameter γ on CLR and MCD, respectively.

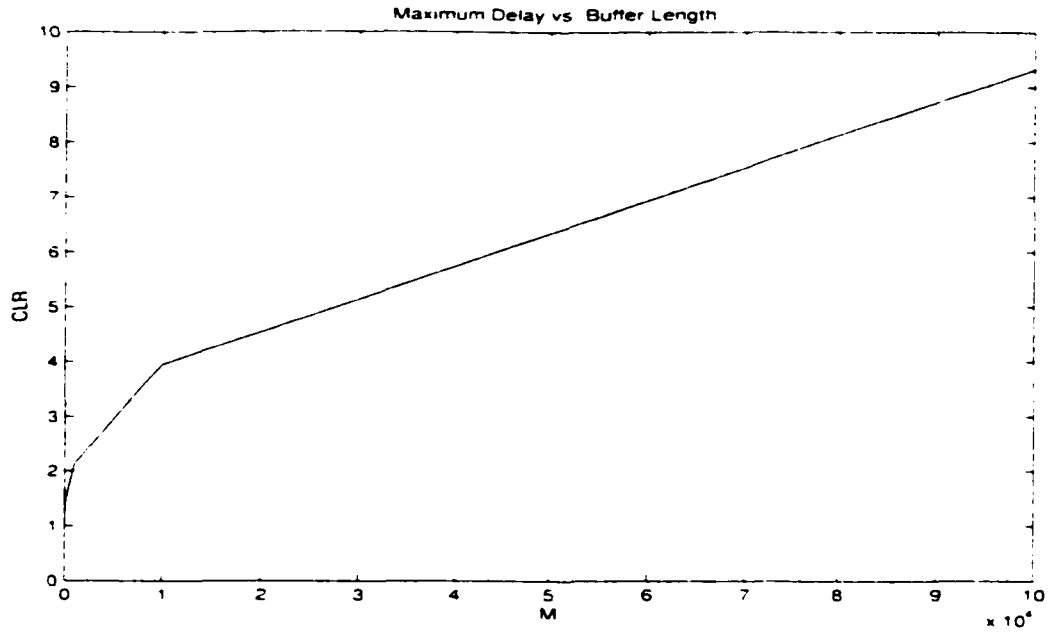


Figure 6.6: MCD vs. M

Here $M = 50$ and $\rho = 512$. The CLR decays exponentially with γ. Although the MCD also decays with γ, it does so at a much slower rate and in an oscillatory fashion.

The utilization is shown in Figure 6.9, and it also decays exponentially with γ. An important design problem is choosing an optimal value for γ. In choosing such a value, one would want CLR and MCD as low as possible with a utilization as close to one as possible. However, since these three parameters decays with γ, there will be a need for some sort of tradeoff. Future work will examine this question.

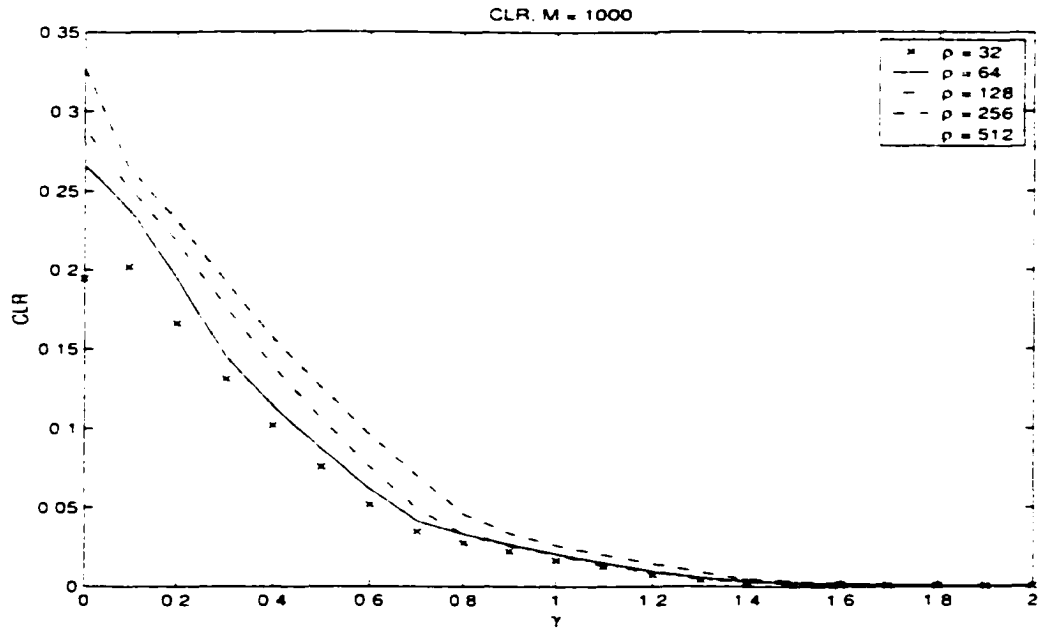


Figure 6.7: CLR vs. γ

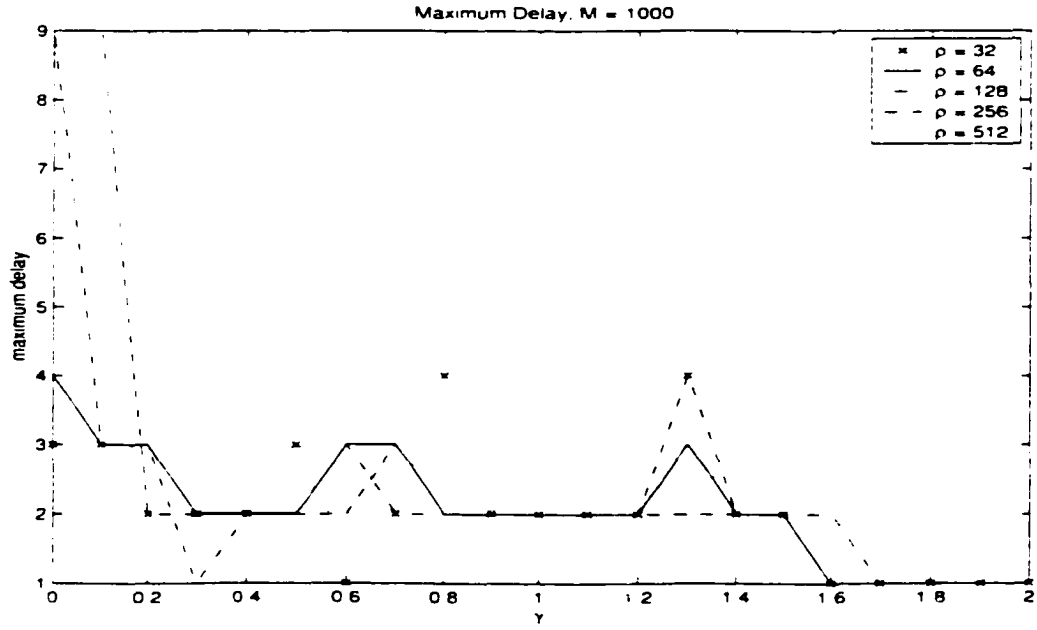


Figure 6.8: MCD vs. γ

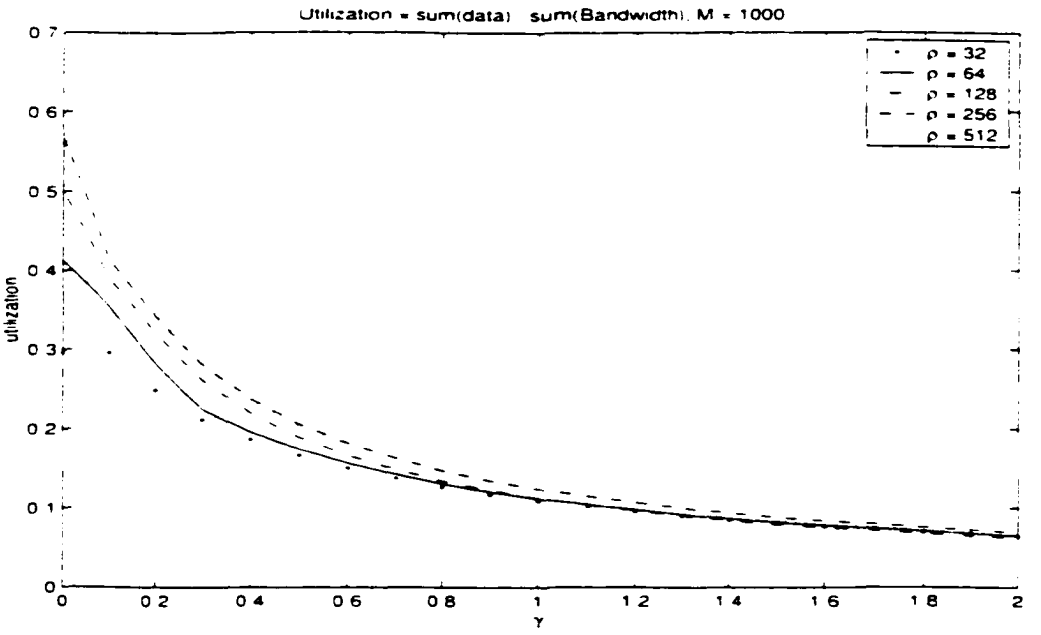


Figure 6.9: Utilization and γ

6.4 The Independent Wavelet Bandwidth Allocation Algorithm

This subsection presents a second bandwidth algorithm—the Independent Wavelet Bandwidth Allocation Algorithm (IWBAA). In many respects the IWBAA is similar to the CMBAA. The major difference lies in the use of the ρSIP algorithm to *generate simulated network traffic ahead of time* and then estimating the bandwidth from this traffic instead of just using past values of the actual network traffic. Let an initial bandwidth of length $\lambda - \rho$ be pre-allocated on the out-going link at a network node, i.e., $(b_k)_{k \in K \cap \Lambda_{\lambda-\rho}}$ are given. Let tolerance $\tau \in \mathbb{R}^+$ be given. Assume that traffic values, $\vec{x} = (x_1, x_2, \cdots, x_\lambda)$, are stored in a buffer with maximum length λ. Without loss of generality, let us also assume that

they are finitely many traffic samples n. The IWBAA is illustrated in Figure 6.10. It is

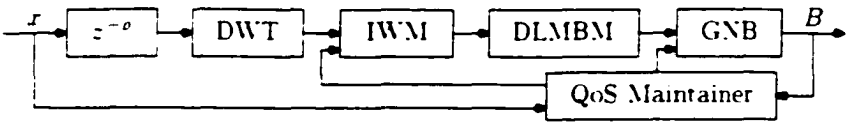


Figure 6.10: Bandwidth Allocation Model

given formally as follows:

Algorithm:

Step 1: Let tolerance τ be given and assume that $B_1, B_2, \cdots, B_{k-\rho}$ have been pre-allocated for the first $\lambda = l\rho$ anticipated traffic values $(x_1, x_2, \cdots, x_\lambda)$ of x.

Step 2: Let $k = \lambda + \rho$ and $w = (x_1, x_2, \cdots, x_{k-2\rho})$;

Step 3: Pass w to ρSIP algorithm to obtain wavelet coefficients $\hat{d} = \hat{d}_1, \hat{d}_2, \cdots, \hat{d}_\rho$ for traffic $x_{k-\rho-1}, x_{k-\rho-2}, \cdots, x_k$;

Step 4: Let $\hat{x}$ be the signal reconstructed via the IDWT of $\hat{d}$;

Step 5: Let $B_{temp} = \mathcal{B}_{\gamma,k}\hat{x}$;

Step 6: Set $B_k = B_{temp}$;

Step 7: Let

$$w^{(1)} = (x_{k-2\rho+1}, x_{k-2\rho+2}, \cdots, x_{(2k-3\rho)/2}) \quad \text{and}$$

$$w^{(2)} = (x_{(2k-3\rho)/2+1}, x_{(2k-3\rho)/2+2}, \cdots, x_{k-\rho})$$

Step 8: Set $b^{(1)} = \mathcal{B}_{\gamma,k}w^{(1)}$ and $b^{(2)} = \mathcal{B}_{\gamma,k}w^{(2)}$

Step 9: If $[(b^{(1)} - B_{temp}) > \tau]$ && $[(b^{(2)} - B_{temp}) > \tau]$

$B_{temp} = max\{b^{(1)}, b^{(2)}\}$

$B_{k-\rho} = B_{temp}$

Else

$B_{k-\rho} = B_{temp}$

Step 10: While $k < n$

$k = k - \rho$

Goto Step 5

6.5 Results

This section presents the results of the IWBAA algorithm for the the wide area TCP trace mentioned in the previous section, LBL-TCP-3[63]. We again use the simplified ATM simulator shown in Figure 6.3 with a finite buffer of length M in order to test the cell loss ratio (CLR), maximum cell delay (MCD) and the utilization of the IWBAA at a network node. The ATM simulator has been discussed in the previous section. The results for the IWBAA are given in Figures 6.11, 6.12, 6.13, 6.14 and 6.15. These results are similar to those for the CMBAA in Section 6.2 (see the discussion there).

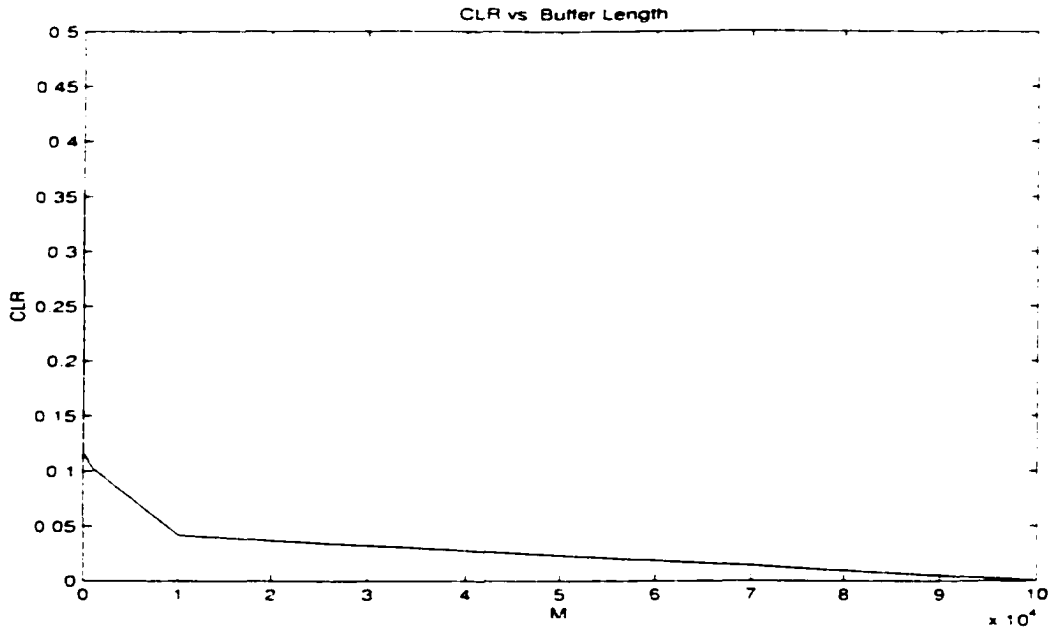


Figure 6.11: CLR vs. M

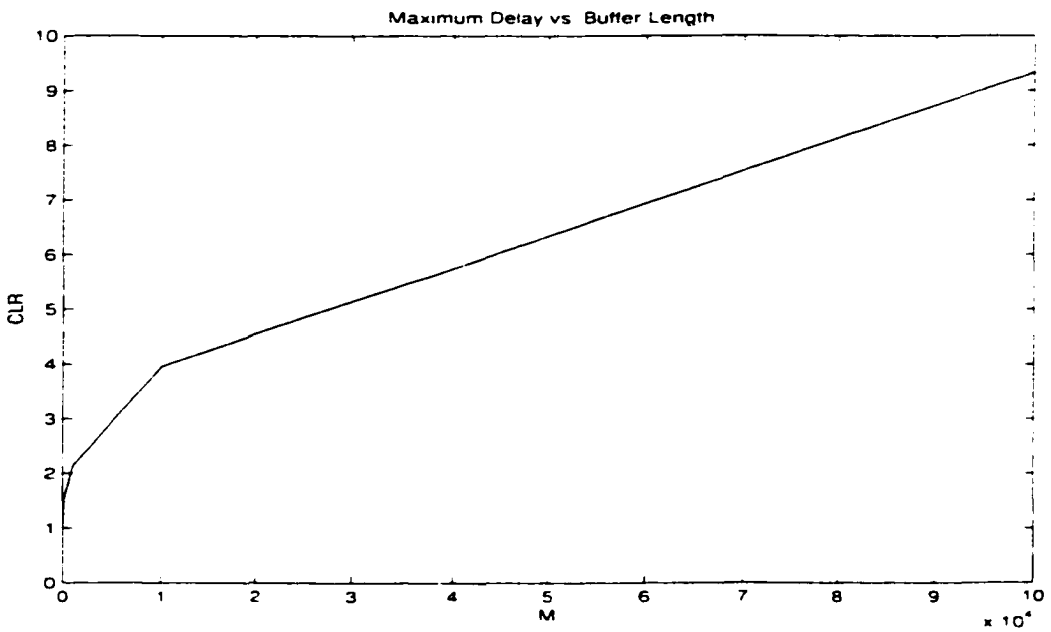


Figure 6.12: MCD vs. M

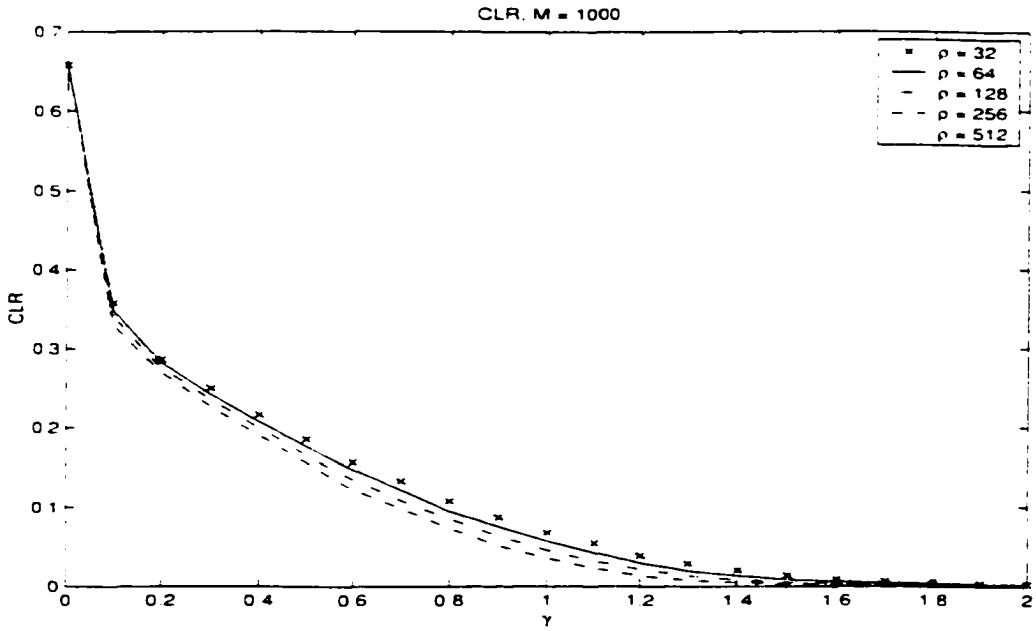


Figure 6.13: CLR vs. γ

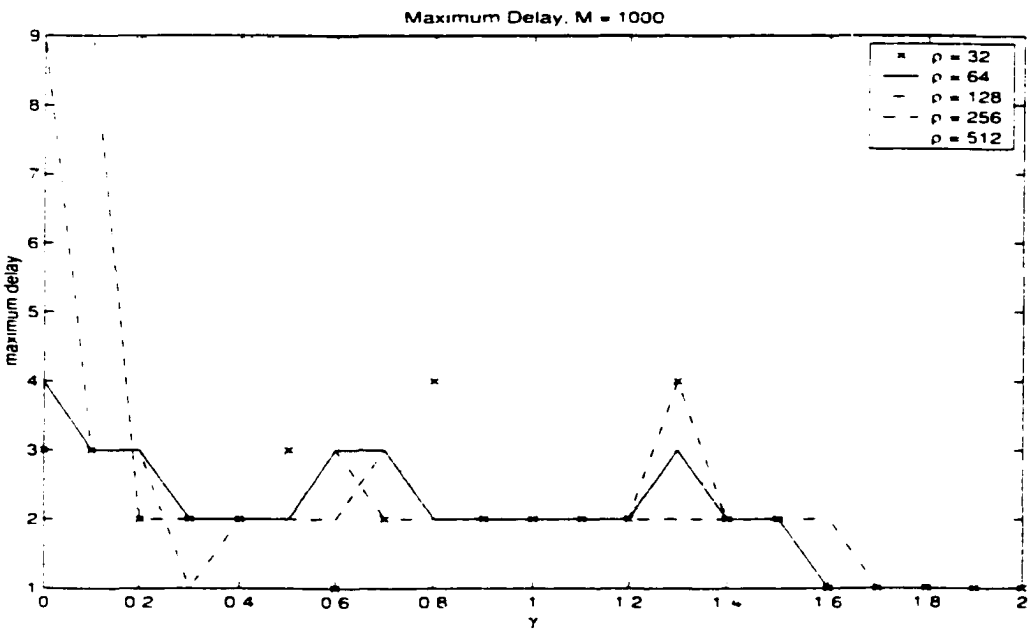


Figure 6.14: MCD vs. γ

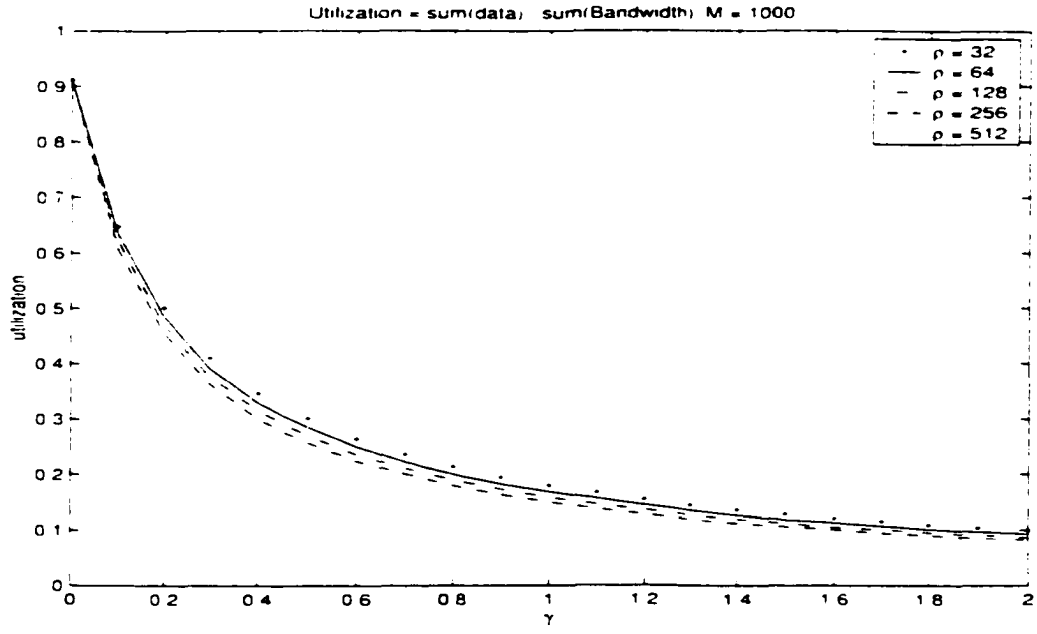


Figure 6.15: Utilization

6.6 Summary

This chapter presents two bandwidth allocation models for to estimate the bandwidth at a network node. The Consecutive Max Bandwidth Allocation Algorithm is based on past values of the traffic as well as previous bandwidth values. The Independent Wavelet Bandwidth Allocation Algorithm is based on past values of the traffic and the bandwidth and also on simulated data generated via the Independent Wavelet Model.

Chapter 7

Conclusion and Future Research

This dissertation investigates the problem of bandwidth reservation and allocation at a communication network node using the concepts of multiresolution analysis, wavelets and time series analysis. Two classes of bandwidth models based on the discrete wavelet transform have been developed to model the traffic bandwidth at a network node. Algorithms for estimating (or predicting) wavelet coefficients needed by the bandwidth models, which themselves are based on the discrete wavelet transform of network traffic have also been provided. The LMBM family of bandwidth models are obtained by taking the maximum of the traffic over the appropriately chosen traffic samples. Another family, the DLMBM family of bandwidth models, are obtained by taking the maximum of the detail signal of the traffic at level 1. We also proposed two prediction algorithms. The ρSIP algorithm, based on the Independent Wavelet Model can be used to estimate wavelet coefficients needed by a chosen bandwidth formula. Similarly, the ρSWD-LP algorithm can also be used to predict the necessary wavelet coefficients for a bandwidth formula. It is based

an on linear prediction or autoregression. An integrated bandwidth allocation algorithm is then constructed by picking an appropriate bandwidth function, and/or a prediction algorithm to predict its wavelet coefficients. We have also developed two special cases of this generalized integrated model. The Consecutive Max Bandwidth Allocation Algorithm (CMBAA) is based on past values of the traffic as well as previous bandwidth values. The Independent Wavelet Bandwidth Allocation Algorithm (IWBAA) is based on past values of the traffic and on simulated data generated via the IWM. Both models have been analyzed with a simulated ATM switch and their performance evaluated in terms of CLR, MCD and utilization. The results for the algorithms are similar. Namely, they show that CLR can be made arbitrarily small by choosing the buffer length at the switch appropriately large. However MCD increases as M gets large while utilization drops below one. Therefore, one important design problem to be investigated is choosing an optimal value for M at a network node. Choosing an optimal values for the parameters γ and α in the DLMBM and LMBM, respectively, are also important problems to be investigated as these parameters control CLR, MCD and utilization. Our results compare favorable with results in [2], and in some cases out-performed those results. The proposed method is also flexible enough to allow further improvements as more sophisticated prediction algorithms become available and as wavelet theory advances.

Future work will continue to investigate more efficient bandwidth models and prediction algorithms that can be used to predict bandwidth components, such as, the large class of nonlinear prediction algorithms that have been developed. We also wish to investigate the problem of using wavelets itself as a prediction tool. Wavelets and multiresolution analysis

can be used to separate traffic into a summation of multiple streams; each stream may then be modelled separately. The overall prediction will then be the sum of the individual predictions. This approach may have several advantages:

- Intuitively, some streams are more important than others and thus need to be modelled more carefully. For example, the traffic streams at either high or low enough resolution may be ignored (i.e. a simple mean model is used).
- Traffic at different resolution may be modelled separately. Intuitively, traffic is composed of multiple variations at multiple time scales (e.g. either net backoff protocol [less than one second], TCP and end-to-end dynamic (1 second), application adaptation (5 seconds), and user behavior (minutes)). This allows us to model different dynamics separately and differently.
- It might be possible to model seasonal traffic variation without much effort.

Bibliography

[1] Gagan L. Choudhury, David M. Lucantoni, and Ward Whitt , "Squeezing the Most Out of ATM," *IEEE Journal on Selected Areas in Communications*, vol. 44, pp. 203 - 217, February 1996.

[2] R. Ting, *A Multiscale Analysis and Analysis Technique for Management of Resources in ATM Networks*, PhD thesis, City University of New York, 1998.

[3] V. Galtier, K.Mills, Y. Carlinet, S. Bush and A. Kulkarni, "Predicting Resource Demand in Heteregeneous Active Networks," in *MILCOM 2001*, Oct.

[4] Claude F. Turner and Rolston M. Jeremiah, "The Independent Wavelet Bandwidth Allocation Algorithm," *IEEE GLOBECOM 2002*.

[5] P. Abry, R. Baraniuk, P. Flandrin, R. Reidi, D. Veitch, "The Multiscale Nature of Network Traffic: Discovery, Analysis, and Modelling," *IEEE Signal Processing zMagazine*, Apr. 2002.

[6] S. Ma, C. Ji, "Modeling Heteregeneous Network Traffic in Wavelet Domain," *IEEE/ACM Trans. on Networking*, Oct. 2001.

[7] Will E. Leland, Murad S. Taqqu, Walter Willinger and Daniel Wilson, " On the Self-Similarity Nature of Ethernet Traffic (Extended Version)," *IEEE/ACM Trans. on Networking*, vol. 2, pp. 1-15, 1994.

[8] Jan Beran, Robert Sherman, Murad S. Taqqu and Walter Willinger, " Long-Range Dependence in Variable-Bit-Rate Video Traffic," *IEEE Trans. on Communications*, vol. 43, no. 2, pp. 165-176, 1995.

[9] Erramilli, Narayan and Willenger, "Experimental Queuing Analysis with Long-range Dependent Packet Traffic," *IEEE/ACM Trans. on Networking*, vol. 5, no. 1, 1996.

[10] S. Ma, *Network Traffic Modeling and Analysis*, PhD thesis, Rensselaer Polytechnic Institue, May 1998.

[11] P. Abry, P. Flandrin, M.S. Taqqu and D. Veitch, "Wavelets for the Analysis, Estimation, and Synthesis of Scaling Data," in *Self-similar Network Traffic and Performance Evaluation* (Kihong Park and Walter Willinger, ed.), Wiley & Sons, Inc, 2000.

[12] A. Grossman and J. Morlet, "Decomposition of Hardy Functions Into Square Integrable Wavelets of Constant Shape," *Siam J. Math. Anal.*, vol. vol. 15, pp. pp. 723–736, July 1984.

[13] R. Jeremiah, *A DCT Multiresolution Analysis for Image Decomposition and Texture Segmentation*. PhD thesis, City University of New York, New York City, NY 10031, Aug. 1998.

[14] R. Jeremiah, R. Ting, and J. Barba, "A bidimensional multiresolution analysis without edge effects.," *Submitted to IEEE Trans. on Image Processing*.

[15] S. G. Mallat, "Multiresolution Approximations and Wavelet Orthonormal Bases of $L^2(\mathbb{R})$," *Trans. of American Mathematical Society*, vol. vol. 315, pp. pp. 69–87, Sept. 1989.

[16] D. Ingrid, *Ten Lectures on Wavelets*. Philadelphia: Siam, 1992.

[17] Eberhard Zeidler, *Applied Functional Analysis - Applications to Mathematical Physics*. New York: Springer Verlag, 1995.

[18] A.N. Kolmogorov and S.V. Fomin, *Introductory Real Analysis*. New York: Dover, 1970.

[19] K. Karhunen, "Ueber Lineare methoden in der Wahrscheinchkeitsrechnung," *Ann. Acad. Sci. Fenn. Ser A.I. Math. phys.*, vol. 37, 1974.

[20] A. K. Jain, *Fundamentals of Digital Image Processing*. Englewood Cliffs, NJ: Prentice Hall, 1989.

[21] A. Aldroubi and M. Unser, "Sampling procedure in function spaces and asymptotic equivalence with shannon's sampling theory," *Numerical Functional Analysis and Optimization*, vol. 15, 1994.

[22] I. D. A. Cohen and J. Feauveau, "Biorthogonal Bases of Compactly Supported Wavelets," *Communication in Pure Applied Mathematics*, vol. XLV, no. 41, pp. 485–560, 1992.

[23] V. P. Walter Willinger and M. S. Taqqu, "Self-similar and heavy tails: Structural modeling of network traffic," in *A Practical Guide to Heavy Tails* (R. Robert J. Adler and M. S. Taqqu, eds.), vol. 1, pp. 201–208, Birkhaüser, 1998.

[24] Kihong Park and Walter Willinger, "Self-Similar Network Traffic: An Overview," in *Self-similar Network Traffic and Performance Evaluation* (Kihong Park and Walter Willinger, ed.), Wiley & Sons, Inc. 2000.

[25] J. Hosking, "Fractional differencing," *Biometrika*, vol. 68, pp. 165–176, 1981.

[26] Benoit B. Mandelbrot , *Fractals: Form, Chance and Dimension* , San Francisco: Freeman, 1977.

[27] Mark W. Garrett and Walter Willinger, " Analysis, Modeling and Generation of Self-Similar VBR Video Traffic ," in *Proceedings ACM Sigcomm*, (London), pp. 269–279, 1994.

[28] G.W. Wornell and A.V. Oppenheim, "Wavelet-based representation for a class of self-similar signals with application to fractal modulation," *IEEE Trans. Inform. Theory*, vol. 40, Aug. 1992.

[29] D.R. Cox, "Long-Range Dependence: A Review," in *Statistics: An Appraisal* (H.D. David and H.T. David, ed.), pp. 55-74, The Iowa State University Press, 1994.

[30] George E.P. Box, Gwilym M. Jenkins and Gregory C. Reinsel, *Time Series Analysis*, Upper Saddle River, NJ: Prentice Hall, 1994.

[31] B. B. Mandelbrot and V. Ness, "Fractional brownian motion, fractional noises and application," *Siam Review*, vol. 10, no. 4, pp. 422-437, 1968.

[32] C. Granger and R. Joyeux, "An introduction to long-range time series models and fractional difference," *Journal of Time Series Analysis*, vol. 1, pp. 15-30, 1980.

[33] Patrice Abry and Darryl Veitch, "Wavelet Analysis of Long Range Dependent Traffic," *IEEE Trans. Inform. Theory*, vol. 44, pp. 2-16, Dec. 1998.

[34] J. Beran, *Statistics for Long Memory Processes*, New York: CRC, 1998.

[35] W. Allan, "Statistics of Atomic Frequency Standards," *In Proc IEEE*, vol. 54, pp. 221-230, 1996.

[36] Darryl Veitch and Patrice Abry, "A Wavelet Based Joint Estimatior of the Parameters of Long-Range Dependence," *IEEE Trans. Inform. Theory*, vol. 45, Apr. 1999.

[37] N. B. Shroff, *Traffic Modeling and Analysis in High Speed ATM Networks*, PhD thesis, Columbia University, 1995.

[38] Mischa Schwartz, *Telecommunication Networks: Protocols, Modeling and Analysis*, Reading, MA: Addison-Wesley Publishing Company, 1987.

[39] Harry Heffes and David Lucantoni, "A Markov Modulated Characterization of Packetized Voice and Data Traffic and Related Statistical Multiplexer Performance," *IEEE Journal on Selected Topics in Communications*, vol. SAC-4, pp. 856 - 868, Sept. 1986.

[40] David N.C. Tse, Robert G. Gallager, and John N. Tsitsiklis, "Statistical Multiplexing of Multiple Time-Scale Markov Streams," *IEEE Journal on Selected Areas in Communications*, vol. 13, pp. 1028 - 1038, August 1995.

[41] Ness B. Shroff, and Mischa Schwartz, "Improved Loss Calculations at an ATM Multiplexer," *IEEE/ACM Transaction on Networking*, vol. 6, pp. 411 - 421, August 1998.

[42] A. Tewfik and M. Kim, "Correlation structure of the discrete wavelet coefficients of fractional brownian motion," *IEEE Trans. Inform. Theory*, vol. 38, no. 2, pp. 904-909, 1992.

[43] P. Flandrin, "Fractional brownian motion and wavelets," in *Wavelets, Fractals and Fourier Transform* (J. H. M. Farge and J. Vassilicos, eds.), Clarendon Press, 1993.

[44] P. Flandrin, "On the spectrum of brownian motion," *IEEE Trans. Inform. Theory*, vol. 35, pp. 197-199, 1989.

[45] P. Flandrin. "Wavelet analysis and synthesis of fractional brownian motion." *IEEE Trans. Inform. Theory*, vol. 38, pp. 910-917, 1992.

[46] Michéle Basseville, Albert Benveniste and Alan Willsky . "Multiscale Autoregressive Processes. Part I: Shur-Levinson Parametrizations." *IEEE Trans. Signal Processing*, vol. 38, pp. 785-800, 1992.

[47] Robert W. Dijkerman and Ravi R. Mazumdar . "Wavelet Representation of Stochastic Processes and Multiresolution Stochastic Models." *IEEE Trans. Signal Processing*, vol. 42, July 1994.

[48] "ITU-T Recommendation I.356, Integrated Service Digital Network (ISDN) - Overall Network Aspects and Function - B-ISDN ATM Layer Cell Transfer Performance," tech. rep., Nov. 1993.

[49] Martin De Prycker, *Asynchronous Transfer Mode: Solutions for Broadband ISDN*, Englewood Cliffs, NJ: Prentice Hall, third ed., 1995.

[50] P. Pancha and M.E. Zarki, "MPEG Coding for Variable Bit Rate Video Transmission," *IEEE Communications Magazine*, pp. 54 - 56, May 1994.

[51] L. Chiariglione, "MPEG and Multimedia Communication," *IEEE Transactions on Circuits and Systems for Vedio Technology*, vol. 7, pp. 5-18, Febuary 1997.

[52] T. Sikora, "The MPEG-4 Video Standard Verificationb Model," *IEEE Transactions on Circuits and Systems for Vedio Technology*, vol. 7, pp. 19-31, Febuary 1997.

[53] M. Nomura, T. Fujii and N. Ohta, "Basic Characteristics of Variable Rate Video Coding ," *IEEE Journal on Selected Topics in Communications*, pp. 305 - 314, Apr. 1991.

[54] R. Grunenberger, J.P. Cosmas, S. Manthorpe and A. Odinma-Okkafor, "Characterization of Video Codecs as Autoregressive Moving Average Processes and Related Queuing System Performance," *IEEE Journal on Selected Topics in Communications*, pp. 284 - 293, Apr. 1991.

[55] R.M. Rodriguez-Dagnino, M.R.K. Kansari and A. Leon-Garcia, "Prediction of Bit Rate Sequences of Encoded Video Signals," *IEEE Journal on Selected Topics in Communications*, pp. 305 - 314, Apr. 1991.

[56] V.G. Kulkani, L. Gun and P.F. Chimento, "Effective Bandwidth Vectors for Multiclass Traffic Multiplexed in a Partitioned Buffer," *IEEE Journal on Selected Topics in Communications*, pp. 1039-1047, Aug. 1995.

[57] R. Guerin, H. Ahmadi and M. Naghshineh, "Equivalent Capacity and its Application to Bandwidth Allocation in High-Speed Networks," *IEEE Journal on Selected Topics in Communications*, pp. 968-981, Sept. 1991.

[58] A Decision-Theoretic Approach to Call Admission Control in ATM Networks, "Effective Bandwidth Vectors for Multiclass Traffic Multiplexed in a Partitioned Buffer," *IEEE Journal on Selected Topics in Communications*, pp. 1101 1114, Aug. 1995.

[59] A. Elwalid, D. Mitra, and R.H. Wentworth, "A New Approach for Allocating Buffers and Bandwidth to Heteregeneous Regulated Traffic in an ATM Node," *IEEE Journal on Selected Topics in Communications*, pp. 1115-1127, Aug. 1995.

[60] E.P. Rathgeb, "Modeling and Performance Comparison of Policing Mechanisms for ATM Networks," *IEEE Journal on Selected Topics in Communications*, pp. 325-334, Apr. 1991.

[61] E.P. Rathgeb, "Effectiveness of the 'Leaky Bucket' Policing Mechanism in ATM Networks," *IEEE Journal on Selected Topics in Communications*, pp. 335-342, Apr. 1991.

[62] Andrew S. Tanenbaum, *Computer Networks*, Upper Saddle River, NJ: Prentice Hall, third ed., 1996.

[63] V. Paxson and S. Floyd , " Wide Area Traffic: The Failure of Poisson modeling ," in *Proceedings ACM Sigcomm*, (London), pp. 269-279, 1994. available through ftp://ftp.ee.lbl.gov/papers/WAN-poisson.ps.Z.