INFORMATION TO USERS

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

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

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

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

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

UMI

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

A Multiscale Analysis and Adaptive Technique for Management of Resources in ATM Networks

by

Rulei Ting

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

1998

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.

9/8/98.
Date

Professor Joseph Barba,
Chair of Examining Committee

9/8/1998
Date

Dean Mumtaz Kassir,
Executive Officer

Professor Mitra Basu

Doctor Joseph Kneuer

Professor Myung Lee

Professor Tarek Saadawi

Supervisory Committee

The City University of New York

Abstract

A Multiscale Analysis and Adaptive Technique for Management of Resources in ATM Networks

by

Rulei Ting

Adviser: Professor Joseph Barba

One of the main advantages of the Asynchronous Transfer Mode (ATM) Broadband-Integrated Services Digital Network (B-ISDN) is the efficient sharing of the network resources through statistical multiplexing of variable-rate traffic streams. Although buffering is provided at the network nodes to relief traffic contention and absorb traffic fluctuations, the burstiness of the ATM traffic presents challenges to guarantee the multiple Quality of Service (QoS) requirements. The effective bandwidth provides an estimation on the bandwidth reservation and admission control, however, its overestimation and underestimation occurs. The ATM traffic bandwidth prediction seems to be an effective way for ensuring the ATM service QoS. During the recent years, neural networks have been utilized as the predictive

mechanisms for the ATM traffic bandwidth predictors. However, limitation exists in the current methods, the neural network parameters need to be specified based on the traffic spectrum, or the traffic details have to be well known before those methods could be employed. During our investigation, a novel mechanism is proposed. The Wavelet decomposition techniques are used for extracting the bandwidth components in ATM traffic. As a result, the Effective Dynamic Bandwidth is introduced. A recurrent neural network with adaptive learning rate is then used for predicting the Effective Dynamic Bandwidth. The algorithms are verified using the Variable Bit Rate (VBR) MPEG traffic volume trace data. The algorithm is also effective on Available Bit Rate (ABR) LAN traffic. Discussion are extended to the future researches in the realtime implementation and the implication with multiple ATM nodes.

Acknowledgments

This dissertation is dedicated to my loving parents. Their wisdom, love and continuous encouragement and support are the major source of inspiration and strength. It is also dedicated to my wife and children for their encouragement, assistance, emotional support and patience in pursuing my Ph.D. studies.

Deepest appreciation to Professor and Dean Joseph Barba, my most respected mentor, who has been providing me with generous guidance and support for many years and introduced me into the magnificent world of Wavelets and their applications. I admire your devotion to the academic program and your kindness to the engineering community at the City University of New York.

My great appreciation also goes to Dr. Joseph Kneuer, a Bell Labs veteran scientist who has been devoting your life to the telecommunications research and produced numerous publications and over fifteen patents in telecommunication arena and provided me with the vision as well as the detailed insights of the rapid growing ATM and related technologies in the fascinating telecommunication industry.

I would like to thank Professor Tarek Saadawi, Professor Myung Lee, Professor Mitra Basu, Professor Ibrahim Habib, Professor Emeritus Donald Schilling and other professors at the School of Engineering for their academic advice, encouragement and continuous support, as well as Dean Mumtaz Kassir, Dean Gerard Lowen, and many faculty members and fellow students at the City University of New York for their guidance and support.

I would also like to thank my professors and mentors at Shanghai Jiao Tong University and McTyeir School, as well as my extended family, for providing me with a solid foundation and guidance over many years.

I am especially grateful to Mr. George MacLachlan and Lucent Technologies Bell Labs for the support during my doctorate research. There, I learned that *INNOVATION* is the engine that keeps the world growing.

Table of Contents

Table of Figures

1. Introduction

1.1 Overview

Asynchronous Transfer Mode (ATM) is a technology recommended by the International Telecommunication Union-Telecommunications Standardization Sector (ITU-T) for the implementation of Broadband-Integrated Services Digital Network (B-ISDN) [1]. It is supported, especially by many in the telecommunications common carrier industry as the backbone for future broadband information services. Using connection-oriented and cell-oriented switching and multiplexing techniques, ATM allows various services such as voice, video and data to be carried in standard 53 byte cells through a single integrated network. ATM provides its customers an all purpose digital network which integrates interactive and distributive services, and guaranteed and best-effort services into one universal broadband network.

One of the main advantages of the B-ISDN is the efficient sharing of the network resources through statistical multiplexing of variable-rate traffic

streams. In order to accomplish that, it requires high-speed transmission and switching facilities, specialized algorithms for congestion control, and new strategies for network management, which ultimately satisfies the requirements for multiple Quality of Services (QoS).

1.2 Research Interests

How to improve QoS in order to meet customer's satisfaction has become an urgent yet difficult research area. Although buffering is provided at the network nodes to relief traffic contention and absorb traffic fluctuations, the burstiness of the ATM traffic presents challenges to guarantee the multiple *Quality of Service* (QoS) requirements to support wide varieties of traffic and diverse services. The cell losses resulted from unanticipated buffer overflow degrades the QoS of ATM services. This dissertation will focus on the consideration dealing with the traffic cell loss and proactive procedures.

Over the past years, a significant amount of research efforts has been in the area of Resources Management in Telecommunication, encompassing areas such as traffic characterization, bandwidth estimation, source policing and admission control, as well as the standards work in explicit feedback & feedforward congestion control protocol, with the goal of improving QoS while increasing the utilization of resources.

For non-stationary signals, especially for the ATM traffic volume, the assumptions for traditionally used statistical models may not be satisfied, the results based on those models may not be valid [73]-[77]. Hence, the traditional traffic studies based on the statistical model have their limitations [78].

It is believed, though, that traffic prediction has become a key element of such improvement due to several reasons:

1. Management of resources in ATM networks require algorithms with "anticipative" or "preventive" actions since the delay-bandwidth product of these networks does not always allow for reactive actions.
2. On certain traffic patterns such as real-time variable bit rate video services, decisions must be made in extremely short time intervals, in order to optimize switching performance.
3. Most of the multimedia traffic sources have poorly understood traffic characteristics. Yet, an effective traffic management strategy requires accurate and simple characterization of the "time-variability" behavior of the traffic.

4. Some traffic management algorithms such as admission control, or policing, should be adaptive in order to respond to the highly dynamic traffic work-load on the links.

However, the existing proposal for the effective bandwidth approximation overestimates when many sources are more bursty than Poisson, and it underestimates the bandwidth for sources less bursty than Poisson. The traffic prediction schemes proposed by Li [54] and Chang [56] are unrealistic because of their limitations on knowing the traffic power spectrum or separating the traffic into I,P,B frames.

1.3 Research Goals

By exploring the research work in this area, we believe that ATM traffic prediction plays an important role in the maintenance of multiple QoS classes, and the development of a coherent traffic management strategy[2]. All of the above require efficient traffic prediction, yet the effective representation of the ATM traffic which may lead to an effective prediction and an integrated ATM traffic management solution, is still under investigation.

In this dissertation, we utilize the Wavelet and Neural Networks techniques for ATM traffic characterization and bandwidth predictions in order to facilitate resource management in the telecommunication systems.

Our research goal is to identify an effective representation of the dynamic bandwidth, and further develop an integrated solution of the dynamic traffic bandwidth prediction contributing to real-time end to end ATM QoS maintenance [79-81], which is robust not only to the existing services but also adaptive to the future new service establishments.

1.4 Organization of the Dissertation

There are five chapters in this dissertation. Chapter One provides an introduction with an overview of ATM, B-ISDN and QoS, our research interests in the network, related research work and our research goals.

The basic and related concepts of ATM Technologies, Quality of Services, Traffic Characterization and Recourse Management in Multimedia Communication are covered in Chapter Two, where the necessity of the traffic prediction in determining the bandwidth and resource allocation is discussed, and the impact of low frequency components on the long term behavior (bandwidth allocation) and the impact of high frequency components on the short term behavior (buffer allocation) are emphasized. The fundamentals on Neural Networks are also presented in this Chapter.

Chapter Three introduces the Wavelets and Multiresolution Decomposition techniques. It further proposes a new model for the Effective Dynamic Bandwidth (EDB), which is based on traffic signal

processing using Wavelets and Multiresolution Decomposition techniques. We will analyze the statistical similarity between the EDB and the original traffic and the effectiveness of the resources allocation when EDB is employed for the bandwidth allocation.

An integrated solution which utilizes the Neural Networks and Adaptive Learning for predicting the EDB are discussed in Chapter Four, where the neural network architectures and adaptive learning schemes are discussed. A recurrent neural network with adaptive learning for predicting the EDB is proposed and numerical results are compared with the Li's and Chang's models.

The ATM traffic prediction on the effective multiscale components demonstrates an important role in the maintenance of multiple QoS classes, and the development of a coherent traffic management strategy. Finally, a summary of the dissertation and suggestions for the future research is in Chapter Five.

2. Background

2.1 ATM Technologies

Asynchronous Transfer Mode (ATM) is a technology recommended by the ITU-T for the implementation of Broadband-Integrated Services Digital Network (B-ISDN). It is supported, especially by many in the telecommunications common carrier industry as the backbone for future broadband information services. Using connection-oriented and cell-oriented switching and multiplexing techniques, ATM allows various services such as voice, video and data to be carried in standard 53 byte cells through a single integrated network, as illustrated in Figure-1.

ATM provides its customers an all-purpose digital network which integrates interactive and distributive services and guaranteed and best effort services into one universal broadband network.

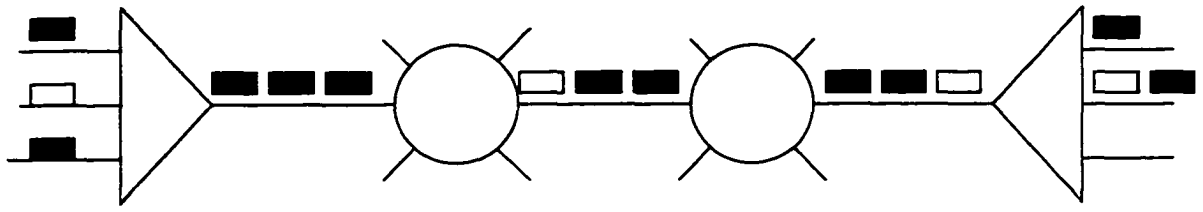

Figure 1 ATM Cells Carried on a Statistically Multiplexed ATM Network

In order to accomplish that, it requires high-speed transmission and switching facilities, specialized algorithms for congestion control, and new strategies for network management, which ultimately satisfies the requirements for multiple Quality of Services (QoS).

2.2 B-ISDN Performance Model & ATM QoS Requirements

The *Network Performance* (NP) [3] provided to B-ISDN users depends on the performance of three layers as shown in Figure-2 :

- *The physical layer* is based on plesiochronous digital hierarchy (PDH), synchronous digital hierarchy (SDH), or cell-based transmission systems. This layer is terminated at points where the connection is switched or cross-connected by equipment using the ATM technique. It does not have end-to-end significance when switching occurs.
- *The ATM layer* is a cell-based switching & multiplexing technique. It is physical media and application independent and

has end-to-end significance. ATM is connection-oriented. The signaling and user information are carried on separate virtual channels. There are two kinds of connections: virtual channel connections (VCC) and virtual path connections (VPC). A VPC can be considered as an aggregate of VCCs. When switching and multiplexing on cells is performed, it must be done first based on the VPC, and then based on the VCC.

- *The ATM adaptation layer (AAL)* enhances the performance provided by the ATM layer to meet the needs of higher layers. The AAL supports multiple protocol types (e.g., SMDS[i], TCP/IP[ii], circuit emulation), each providing different functions and different performance. However, the ATM/AAL boundary is not "pure" in the sense that ATM traffic management does provide some AAL dependent services at the ATM layer, e.g. early and partial packet discard during congestion.

ATM services show very specific characteristics with respect to the traffic parameters, such as peak cell rate, sustainable cell rate[iii], burst tolerance, delay variation tolerance, etc. and a set of QoS parameters.

---

[i] SMDS - Switched Multimegabit Data Service
[ii] TCP/IP - Transmission Control Protocol/Internet Protocol
[iii] See Figure 4, VBR Traffic Fluctuation

The QoS requirements are generally specified in terms of information loss, information delay, and delay variability. The basic QoS parameters for an ATM connection are: cell loss ratio, cell transfer delay, and cell delay variation (jitter). Other QoS parameters include: cell misinsertion rate, cell error ratio, cell transfer capacity (throughput), and skew (difference in representation times of two related objects, e.g. video/audio).

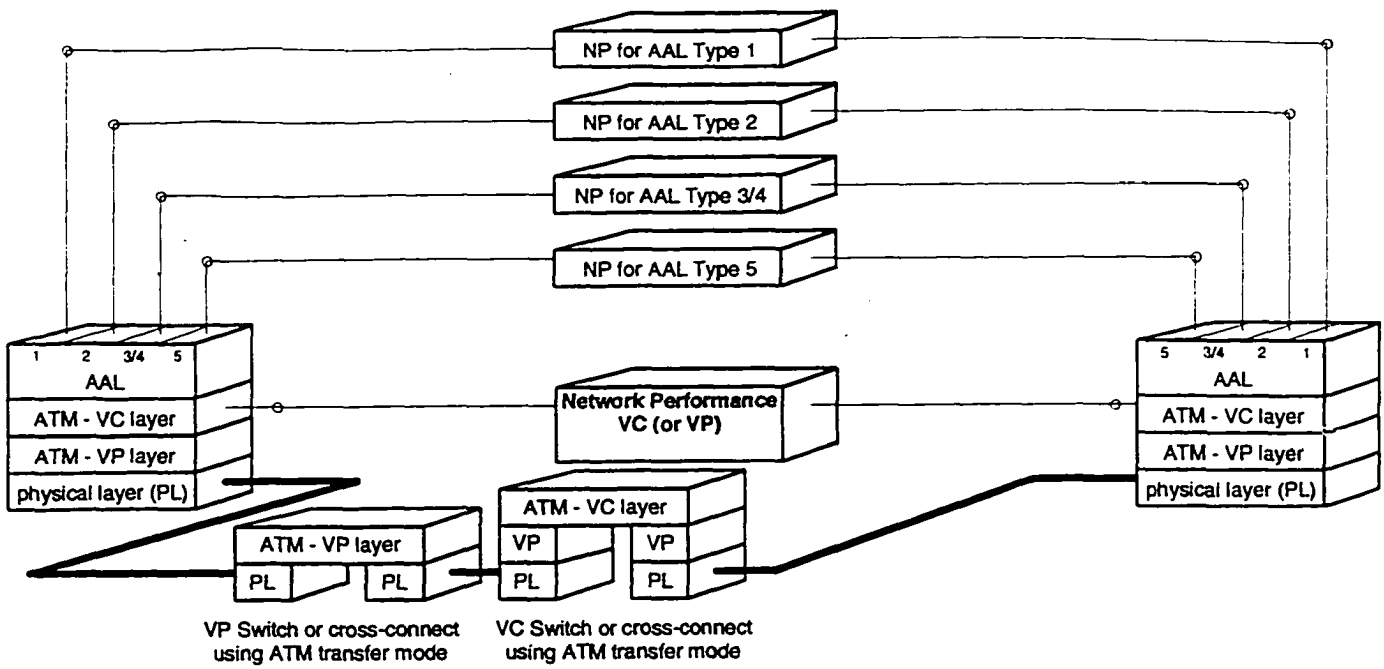

Figure 2 B-ISDN's Layered Performance Model

The major ATM cell transfer performance parameters are defined in [3]:

- *Cell error ratio (CER)* is the ratio of total errored cells to the total of successfully transferred cells, plus tagged cells and errored cells in a population of interest. Successfully transferred cells, tagged cells, and

errored cells contained in severely errored cell blocks are excluded from the calculation of cell error ratio.

- *Cell loss ratio (CLR)* is the ratio of total lost cells to total transmitted cells in a population of interest. Lost cells and transmitted cells in severely errored cell blocks are excluded from the calculation of cell loss ratio. Three special cases are of interest, CLR_0 , CLR_{0+1} , and CLR_1.

  1. *The cell loss ratio for high priority cells (CLR_0)*: $CLR_0 = \frac{N_l(0)}{N_t(0)}$,

     where $N_t(0)$ is the number of CLP=0 cells transmitted[i]; and $N_l(0)$ is the number of corresponding lost cell outcomes plus the number of corresponding tagged cell outcomes. Cells that are tagged by the network (possibly due to overpolicing) are considered lost from the stream of high priority cells.

  2. *The cell loss ratio for the aggregate cell stream (CLR_{0+1}):* $CLR_{0+1} = \frac{N_l(0+1)}{N_t(0+1)}$, where $N_t(0+1)$ is the total number of cells transmitted, and $N_l(0+1)$ is the number of corresponding lost cell outcomes. Tagged cells are not considered lost from the aggregate stream. $CLR_{0+1}=CLR_1$, when all cells are CLP=1.

[i] See Figure 5, Cell Tagging for Traffic Policing

3. *The cell loss ratio for low priority cells (CLR_l):* $CLR_l = \frac{N_l(1)}{N_t(1)}$,

where $N_t(1)$ is the number of CLP=1 cells transmitted, and $N_l(1)$ is the number of corresponding lost cell outcomes. Cells that are tagged by the network (but are still conforming to the aggregate traffic contract) are not considered in either the numerator or the denominator of the expression for CLR_l. As defined, CLR_l quantifies the user's perception of the cell loss ratio for their low priority traffic.

- *Cell misinsertion rate (CMR)* is the number of misinserted cells per connection second. Misinserted cells and time intervals associated with severely errored cell blocks are excluded from the calculation of cell misinsertion rate.[i]
- *Severely errored cell block ratio (SECBR)* is the ratio of total severely errored cell blocks to total cell blocks in a population of interest. The severely errored cell block outcome and parameter provide a means of quantifying bursts of cell transfer failures and preventing those bursts

[i] A misinserted cell is a received cell that has no corresponding transmitted cell on the considered connection. Cell misinsertion on a particular connection is caused by impairments either on physical layer unassigned cells or on cells being transmitted on a different connection. Since the mechanisms that cause misinserted cells have nothing to do with the number of cells transmitted on the observed connection, this performance parameter cannot be expressed as a ratio, only as a rate.

from influencing the observed values for cell error ratio, cell loss ratio, cell misinsertion rate, and the associated availability parameters.

- *Cell transfer delay (CTD)* is the time, $t_2 - t_1$, between the occurrence of two corresponding cell transfer reference events, CRE_1 at time t_1 and CRE_2 at time t_2, where $t_2 > t_1$ and $t_2 - t_1 \leq T_{max}$. The value of T_{max} is for further study, but should be larger than the largest practically conceivable cell transfer delay. This definition can only be applied to successfully transferred, errored, and tagged cell outcomes.
- *Mean cell transfer delay* is the arithmetic average of a specified number of cell transfer delays.
- *Cell delay variation (CDV)* is the absolute maximum variation during the observation interval. There are two cell transfer performance parameters associated with cell delay variation being defined:

    1. *1-point cell delay variation*, is defined based on the observation of a sequence of consecutive cell arrivals at a single Measurement Point (MP). The *1-point CDV* parameter describes variability in the pattern of cell arrival (entry or exit) events at an MP with reference to the negotiated peak cell rate 1/T [23]; it includes variability present at the cell source

(customer equipment) and the cumulative effects of variability introduced (or removed) in all connection portions between the cell source and the specified MP. It can be related to cell conformance at the MP, and to network queues. It can also be related to the buffering procedures that might be used in AAL 1 to compensate for cell delay variation.

2. *2-point cell delay variation*, is defined based on the observations of corresponding cell arrivals at two MPs that delimit a virtual connection portion. The *2-point CDV* parameter describes variability in the pattern of cell arrival events at the output of a connection portion (e.g. measurement point MP_2) with reference to the pattern of corresponding events at the input to the portion (e.g. measurement point MP_1); it includes only the delay variability introduced within the connection portion. It provides a direct measure of portion performance and an indication of the maximum (aggregate) length of cell queues that may exist within the portion.

Additional information on the QoS requirement can be found in [3].

2.3 Studies on Traffic Characterization & Resources Management

The ATM technology is based on the statistical multiplexing of variable bit rate sources. Due to the burstiness of the traffic sources, the QoS in an ATM network might be degraded if there is no intelligent control mechanisms. Studies made in this area exist from source characterization to congestion control, with a goal of improving QoS and increasing the network utilization. This section summarizes these recent studies.

2.3.1 Types of Sources

The communication network provides services to many types of sources, such as remote local area networks, high-resolution still image transfer, mixed-mode document exchange and retrieval, high quality interactive videotext, videophone, video conference, distributive TV, voice and emerging telemetric services. All of these sources have different characteristics[4] in terms of traffic burstiness, bit rate, and call duration, as illustrated in Table-1[i]. Each of them also imposes different service requirements as shown in Table-2[ii]. Both Table-1 and Table-2 are at their best representatives.

---

[i] Average Bit Rate: E[s(t)], Burstiness: B = max[s(t)]/E[s(t)]. The duration over which the peak and average is calculated may be an important parameter to characterize a service.

[ii] Bit Error Rate (BER), Packet Loss Rate (PLR), Packet Insertion Rate (PIR)

Table 1 Broadband Services and Their Characteristics

Service	E[s(t)]	B
Voice	32 Kb/s	2
Interactive Data	1 - 100 Kb/s	10
Bulk Data	1 - 10 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

Table 2 Service Attributes for an ATM Network

Service	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^{-3}	10^{-3}	1000 ms

The ATM services are generally classified as *Constant Bit Rate (CBR), Variable Bit Rate (VBR)* and *Available Bit Rate (ABR).* CBR is used for circuit emulation where a constant bandwidth is required. ABR is subscribed for bulk data transmission, LAN or TCP/IP traffic. VBR is the most challenging service type in terms of the resource management due to the fact that heterogeneous traffic sources are multiplexed together and traffic volume fluctuates with time.

ABR is defined, and supported in varying degrees in the network, but it has not "taken off". This is partly due to the lack of endpoint support, but also there is considerable concern about interaction of ABR and TCP control mechanisms.

So in current systems the broad VBR category is pressed into service for distinct service classes: for contracted real-time VBR with committed delay QoS; for contracted non-real-time VBR with weighted fairness and loss probability QoS; and for Unspecified Bit Rate (UBR) for best-effort and "take your chances" service. This latter, with AAL5 packet discard mechanisms, is argued as an alternative to ABR for TCP traffic.

With the potentially vast amount of real-time services such as the image and video application, the multimedia VBR traffic such as JPEG & MPEG encoded data [5],[6] has drawn more interest in telecommunication traffic research and applications.

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

The standard supported by the Motion Picture Experts Group (MPEG) is for video compression. It comes with three versions, MPEG-I, MPEG-II, and lately MPEG-IV. MPEG-I is used for non-interlaced television at 30 frames/second and with a compressed bit rate at the order of 1 Mbits/sec; MPEG-II is used for interlaced television at 60 fields/sec and with a compressed bit rate from 5 to 10 Mbits/sec.

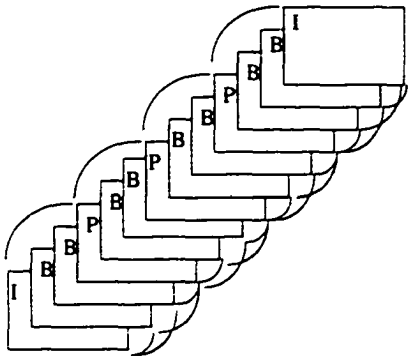

Figure 3 The Inter Frame Relations of MPEG GOP

MPEG uses both hybrid motion compensated predictive DCT coding for a subset of frames and bi-directional motion interpolation. The MPEG data is structured into *Group of Pictures* (GOP's). Each GOP consists of numerous I, P, B frames. The I frame is an intra-picture coded without reference to other frames. It serves as entry points for the random access. The P frame is a predictive picture coded with reference to previous I or P frames. The B frame is a bi-directionally predictive picture coded with reference to both a previous I or P frame and a future I or P frame. Example of such sequences is: "I B B P B B P B B P B B I .....", where each GOP contains 12 frames as in Figure-3.

The international standard MPEG-IV is currently being developed and expected to be released by the end of 1998 [40]. It is intended for a flexible content-based audio and visual environment ranging from conventional storage and transmission of audio and video to truly

interactive audio-visual services requiring content based database access, remote editing of audio visual, games, tele-shopping, remote monitoring and control, low bit rate public switched network, etc.

2.3.2 Model Based Traffic Source Characteristics

Most of the model based studies rely on the statistical characteristics (imposed with special assumptions), using a suitable approximation method such as fluid model, phase type process, Markov Modulated Poisson-Process (MMPP), Intercepted Poisson Process (IPP), Fractal Brownian Motion (FBM) process among many others. Nomura[7] used three measures: distribution, autocorrelation, and coefficient of variation to evaluate the burstiness. Video sources are modeled and characterized by the autoregressive (AR) process and coefficient of variation. Autoregressive Moving Average Process(ARMA) was used by Grunenfelder[8]. A parametric model of the encoding algorithms was introduced by Rodriguez, which generalized video traffic characterization regardless of their actual encoding scheme being employed [9].

Within the models being considered, the Markov models have been the most widely used ones due to their simplicity. Some recent developments by Hui[10], proposed to use thermodynamic theory for

characterizing sources for purposes of bandwidth assignment and buffer assignment.

2.3.3 Non-Model Based Traffic Characterization

Non-model based traffic characterization studies bypass the modeling procedure ranging from the entropy & energy functions to traffic components decomposition. By exploring the analogy between the rate-function and thermodynamic theory, the rate-function[11] can be estimated directly and is used to estimate the bandwidth by computing the tail distribution of the buffer occupancy stochastic process.

Fowley[12] examined an existing LAN traffic at high time-resolution. The author reported that LAN data has extreme traffic variability on time scales ranging from milliseconds to months, and conventional traffic models do not capture this behavior, which has a profound impact on the nature of traffic congestion. The findings indicate that during a congestion period, congestion persists and data losses may be significant, and congestion losses cannot be avoided by modest increase in buffer capacity.

Tse and Gallager[13] studied the problem of statistical multiplexing of cell streams that have correlation at multiple time-scales, with an approach of multiple time-scale decomposition of the traffic process. Their results emphasized the dominant role of the slow time scale correlations.

2.3.4 Bandwidth Analysis and Approximation

ATM is capable of supporting a wide range of connections with different bandwidth requirements and traffic characteristics. While this environment provides an increased flexibility in supporting various services, its dynamic nature poses difficult traffic control problems when trying to achieve efficient use of network resources, especially for the bandwidth management and allocation for real-time traffic. Providing the deterministic guarantees requires the network to allocate resources according to the worst-case scenario, which in turn requires the peak-rate resource allocation scheme that would significantly under-utilize the network for VBR traffic where the peak to average ratios is high[i].

Stochastic fluid models and the theory of large deviations have been used to define effective bandwidth or equivalent capacity for each source. The effective bandwidth depends on the source characteristics, the QoS requirement of the source and the buffer size[14]. An approximation expression for the "equivalent capacity" of both individual and multiplexed connections was proposed by Guerin[15]. Resource management using Effective Bandwidths has been demonstrated by De Veciana,

[i] In the presence of non-real-time traffic, the cell buffer resource is a significant problem.

which is based on the general classification of the ATM connections into: deterministic, statistical, and best-effort[16].

However, a study by Choudhury [17], 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 connection admission control algorithm using the effective bandwidths based solely on tail-probability asymptotic decay rates may not be as effective as hoped. 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 and the effective bandwidth approximation can greatly under-estimate the target blocking probabilities.

Hence, new approximation mechanisms are required.

Call Admission Control, which determines whether the network has sufficient resources to support a new connection without degrading the service of existing connections[33], has drawn significant attention as well. Gibbens[18] presented a decision-theoretic approach which employs Bayesian decision theory with time-scale decomposition. Elwalid[19] and others developed QoS assurance, especially protection of one connection QoS from behavior of another connection, 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 respect of allocating buffers and bandwidth to heterogeneous regulated traffic in ATM node.

2.3.5 Source/Traffic Policing Function

In order to ensure an acceptable quality of service for all coexisting calls sharing the same network resources, control of the individual cell streams within the entire duration is imposed. This procedure is called policing, or usage parameter control function. The traffic in excess of its specified characterization may not be permitted to enter the network. These mechanisms include: *Cell Tagging, Leaky Bucket Mechanism (LB), Jumping Window Mechanism (JW), Triggered Jumping Window Mechanism (TJW), Moving Window Mechanism (MW),* and *Exponentially Weighted Moving Average Mechanism (EWMA)*[20]. The Cell Tagging and Leaky Bucket as respectively shown in Figure-5 and Figure-6, are orthogonal to the violation measurement technique in the UPC. However, UPC recommendations and most real implementations are Leaky Bucket, where it isn't really possible to distinguish a bucket overflow due to operation above PCR from one due to an excessively long burst at or below PCR.

The most popular traffic policing function, Leaky Bucket, was analyzed as a G/D/1/N queue with finite waiting room N and a suitable arrival process by Butto[21]. There are other analyses such as the one by Kushner [22], where several standard control schemes including leaky bucket/token-type control were presented. In his work, a very important point was amply demonstrated that if the rule for the marking and deletion can depend on the system state, then the operation of the system can be greatly improved, since the deletion occurs only at those (rarer) moments when a problem needs to be managed.

In order to incorporate VBR video traffic into the network with service guarantee, a key challenge still remains in the difficulty of finding an appropriate traffic characterization that captures the dynamic of the sources. We try to illustrate an instantaneous view of the cell arrival rates in a figure, however, the short term jitter may make that appear very large. A windowed low-pass filtered view of the traffic arrivals is shown in Figure-4.

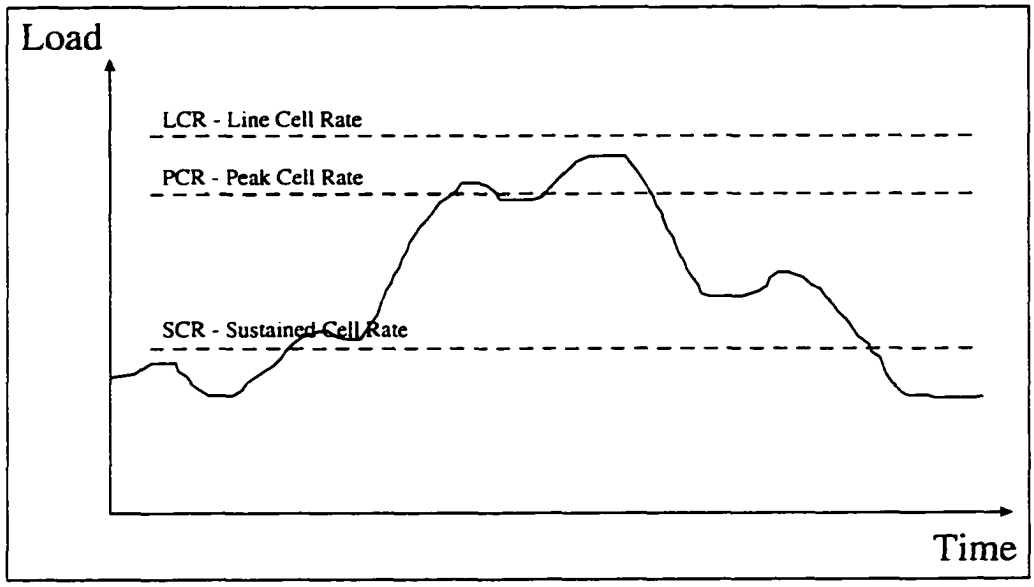

Figure 4 VBR Traffic Fluctuation

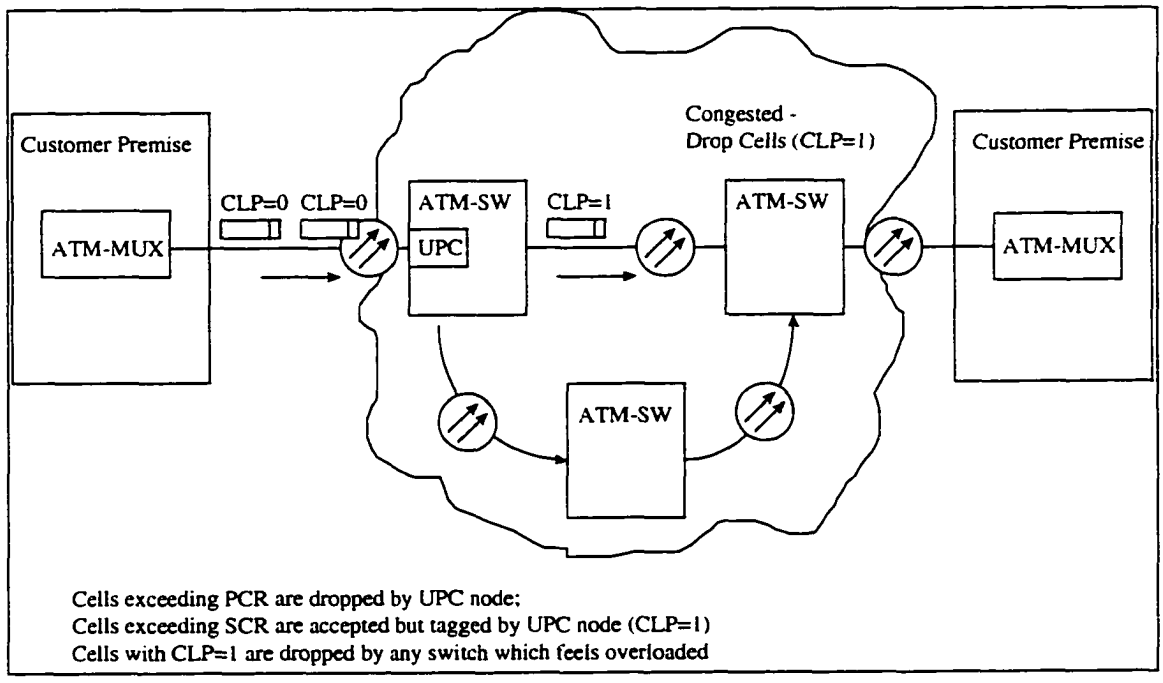

Figure 5 Cell Tagging for Traffic Policing

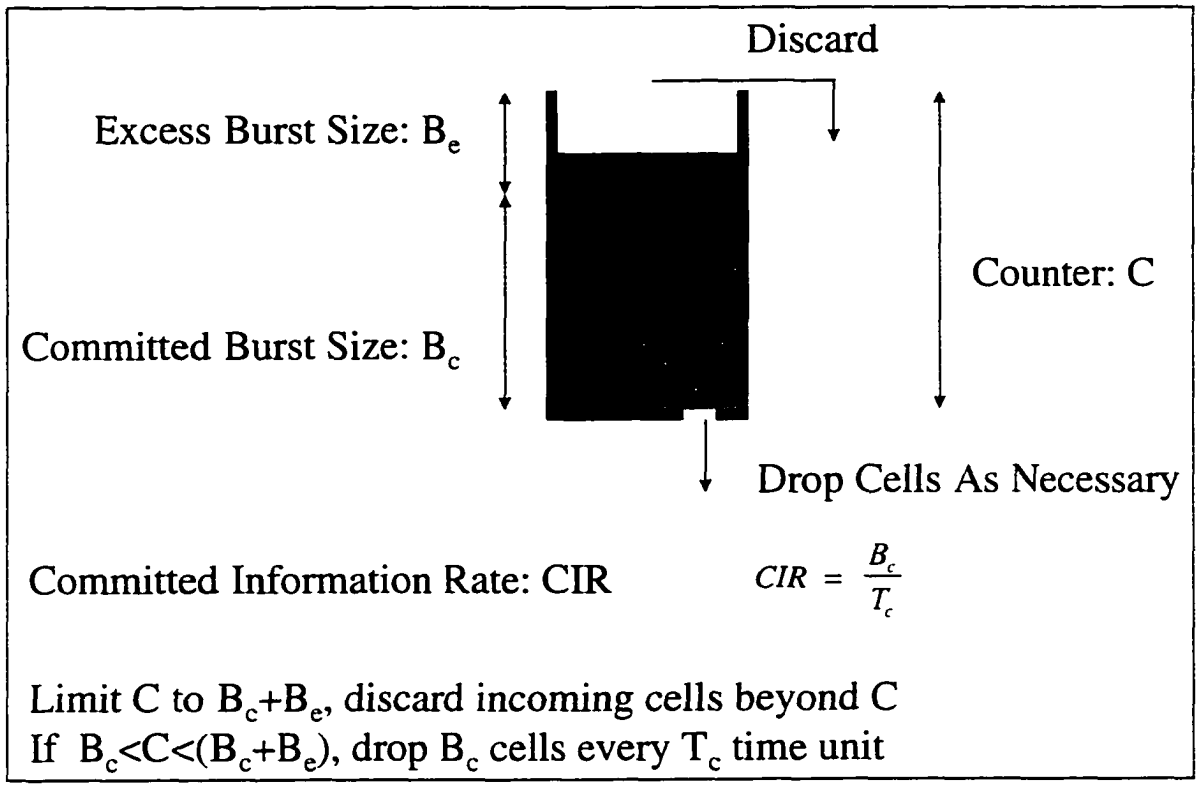

Figure 6 Leaky Bucket Algorithm

2.3.6 Feedback & Feedforward Congestion Control Protocol

The studies on ATM congestion control have yielded the techniques in [23], such as *selective discard of violating cells*, *discard block of cells*, and *end node notification* techniques. The end node notification techniques include: *estimation by the end nodes*, *explicit backward congestion notification* (EBCN), *and explicit forward congestion notification* (EFCN) [23].

When estimation by the end nodes is in use, a source sends time-stamped probe cells along the connection route periodically to measure the

response time between the source and the destination. They are used to estimate one-way delays. Destination node notifies the source to adjust its rate in case of abnormalities. With thousands of *Virtual Channel Connections* (VCCs) per access node, the processing burden is significant.

The EBCN is used in current low-speed networks as a back-pressure mechanism, however, it is not adopted by ITU-T due to the requirement of considerable processing at intermediate nodes and resulting excessive delays.

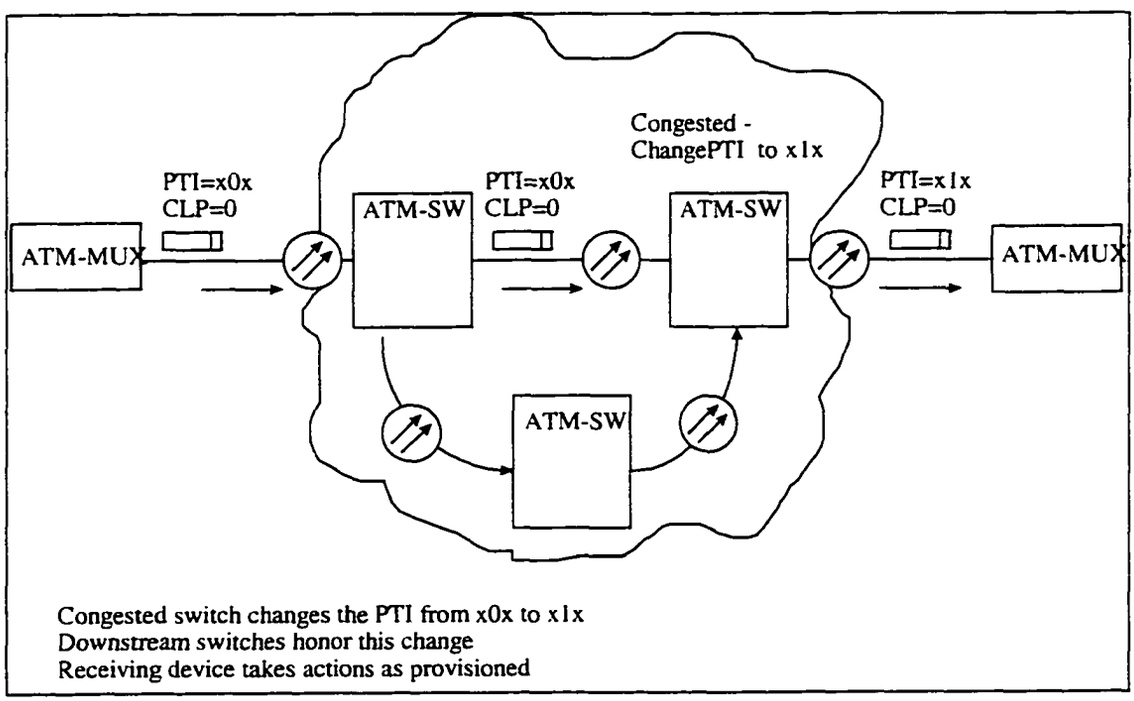

Figure 7 Explicit Forward Congestion Notification (EFCN) for Congestion Control

In EFCN as shown in Figure-7, a special indication, PTI, is carried in user data cell. It is changed by congested switches to indicate that congestion is experienced. All downstream switches honor this change and

let it go through. The receiving device may take appropriate actions as provisioned, while feedback controls are generally defined as a set of actions taken by the network and by the users to regulate the traffic submitted on ATM connection according to the state of the network elements.

EFCN/EBCN are normally associated with ABR (although they have other proprietary uses). More recently *Explicit Rate* (ER) feedback has gained favor for its faster, more (or less) stable, and fair (min-max) operating characteristics. The EFCN is no longer confined to narrow bandwidth applications - the largest commercial switches (25 Gbps aggregate capacity, 622 Mbps link rates) currently support EFCN and will add ER calculation next year.

The overhead for ABR feedback is comparable to that of other necessary switching functions such as address translation, policing, buffer management, routing, scheduling, data collection, etc. The current state of VLSI capacity makes the calculation relatively easy, the costs for all of these functions are heavily weighted by the need for read-update access to large off-chip RAM tables on a per-connection basis. In aggregate the control bandwidth requirements significantly exceed the cell buffer bandwidth requirements.

In these schemes, the traffic congestion controls are *reactive*.

2.3.7 Layered Source Encoding

Layered video coding is a source encoding scheme coded with priority levels which compensates packet loss in a packet-based network. When the coding information is separated into the most significant parts (MSP's) and the least significant parts (LSP's). The MSP packets take priority over the LSP packets. Hence, controlled and limited packet losses can have minimal effect on picture quality if the switching systems can support preferential cell and packet discard policies [24].

2.3.8 Neural Network Applications

Artificial Neural Networks are information processing systems implemented on general or special purpose computer system and electronics circuits. The concept was originated from natural nervous systems, however, it may or may not have the exact mapping to a biological function and behaviors.

The element in a neural network is called *neuron.* It has an output, which is a function of the weighted sum of a set of inputs [26], as shown in Fifure-8.

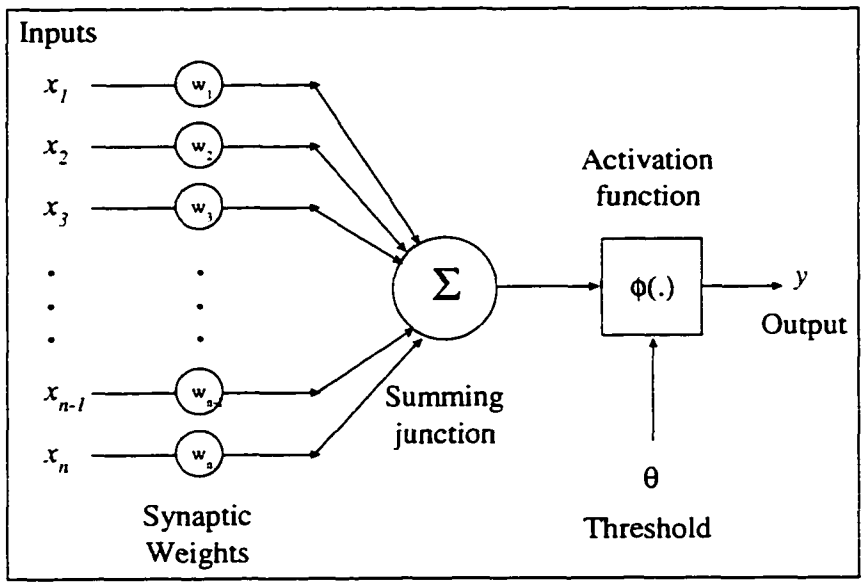

Figure 8 Neuron - An Element In Neural Network

The power of the neural network mainly comes from:

1. The network interconnected by groups of neurons;
2. The learning capabilities of the network through the weight variations;
3. *Kolmogorov's Three-Layer Theorem,* which states that any input-output mapping may be approximated to any degree of closeness by a structure equivalent to a three-layer neural network.

The main reasons for using an artificial neural network are: *versatility, tractability, learning capability, robustness,* and *speed.* The convergence theorems of most of the networks are readily proved. The learning could take place automatically. Neural networks can learn from

noisy data and generate optimal solutions to probabilistic problems. Additionally, they could be implemented in hardware for speed due to their inherently parallelism.

A fully connected neural network is shown in Figure-9.

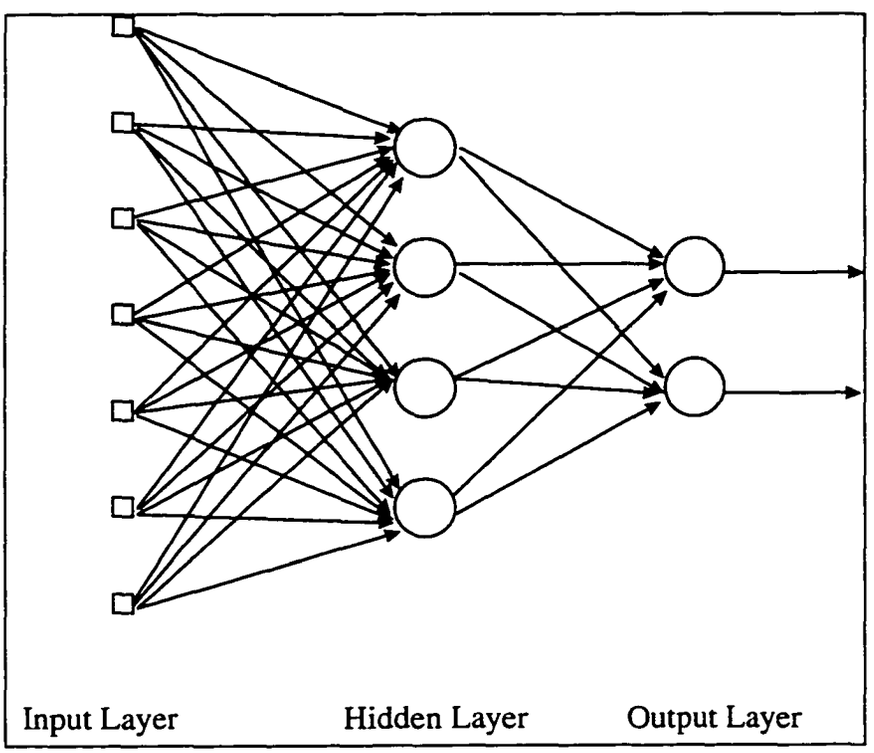

Figure 9 A Fully Connected Multilayer Feedforward Network

When a fully connected multi-layers neural network architecture is employed, there are two major operational stages: *Learning* and *Generalization*, where the learning rules enforce the weights to converge to the optimal range during the training processes, and ultimately making the network capable of generalizing its knowledge to the generalization data

set. Some repetitive cycles may make the system learning as a real-time or semi-realtime process.

Because of that a three-layered neural network can approximate an arbitrary nonlinear function by precisely adjusting weights between neurons, Hiramatsu introduced a scheme for call admission control and link capacity control using neural networks in [25]. The neural network learns those relations as a nonlinear function from the noisy data. Traffic control using neural network adapts to the changes in traffic characteristics and to the addition of new communication services. This scheme is suggested for multimedia services with unknown network characteristics. Detail works on the ATM traffic prediction using Neural networks have been studied in [36] and [41]. The potentials of traffic prediction without statistical model have been demonstrated.

2.4 Effective Traffic Prediction For Resource Management

How to improve QoS in order to meet customer's satisfaction is an urgent yet difficult research area.

ATM advantages rely on the multiplexing of various statistically independent traffic sources. The unpredictable burstiness in traffic streams may degrade the network's quality of service.

From the above discussion, we learn that:

- Traffic models need to be developed to exhibit the variations on many time scales and represent the characteristics in new traffic sources and services;
- By evaluating the tradeoffs between the controlled-loss and buffer-loss, ATM traffic subscriber favors the controlled-loss versus the buffer-loss, as the control-loss generally causes less damage to the traffic.

In the area of resource management, we also learn that:

- Increasing the capacity of the link buffer is an approach for reducing the sensitivity of the system to small changes in offered traffic. While this helps loss, it may even increase delay problems;
- The magnitude of low-frequency traffic variability implies that precise engineering of components such as inter-office links will be difficult. Hence, dynamic bandwidth allocation might be required.

It is concluded that mechanisms anticipating traffic fluctuations will contribute to the improvement of the ATM service quality.

These lead to the need in the study of the high-speed network traffic prediction. Source Regulation will directly benefit from the traffic prediction in regulating the single source with more accurate state information of other sources traffic rate and the total network traffic

utilization status. Dynamic Network Resource Allocation will also benefit from the traffic prediction with precise information on the long-term demand for the bandwidth allocation and short-term demand for controlling congestion in the small buffer size.

We believe that the effective modeling of ATM traffic and the prediction of the effective bandwidth components may provide more promising results for improving the quality of ATM services.

3. Modeling of the Effective Dynamic Bandwidth

3.1 Multiscale Phenomenon In Traffic

By reviewing some of the research highlights in ATM traffic engineering area, we understand that:

- The studies on the problems of statistical multiplexing of cell streams that have correlation at multiple time-scales, revealed *the dominant effect of the slow time scale correlations*.
- The congestion losses generally cannot be avoided by modest increase in buffer capacity during a congestion period where congestion persists and data losses may be significant.

During our studies, the self-similarity phenomenon at different time scales [66], i.e. fractal effect, of the multimedia traffic, became obvious with the aid of Wavelet multiscale signal decomposition, as illustrated in Figures10 and 11.

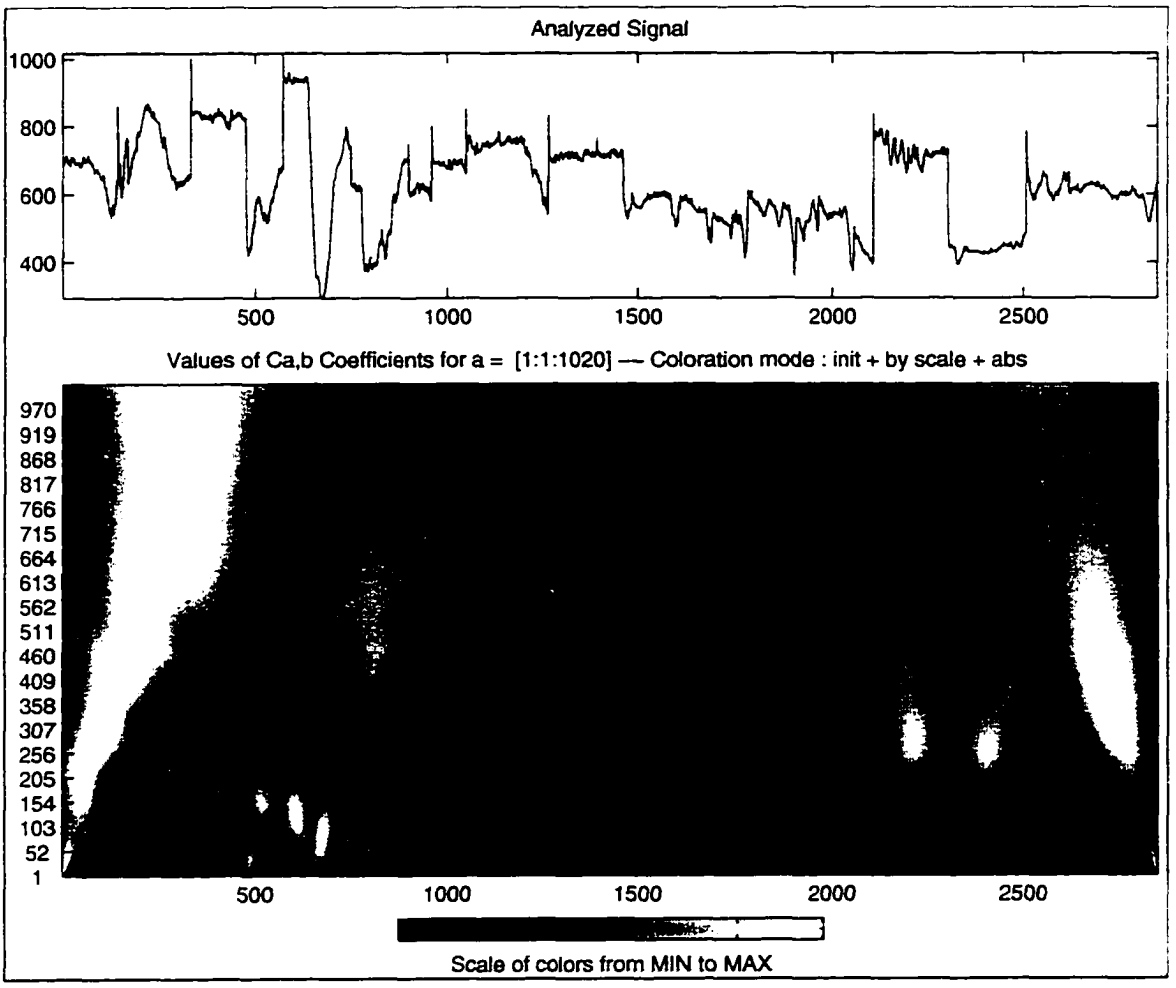

Figure 10 Continuous Wavelet Analysis of ATM traffic Volume-1

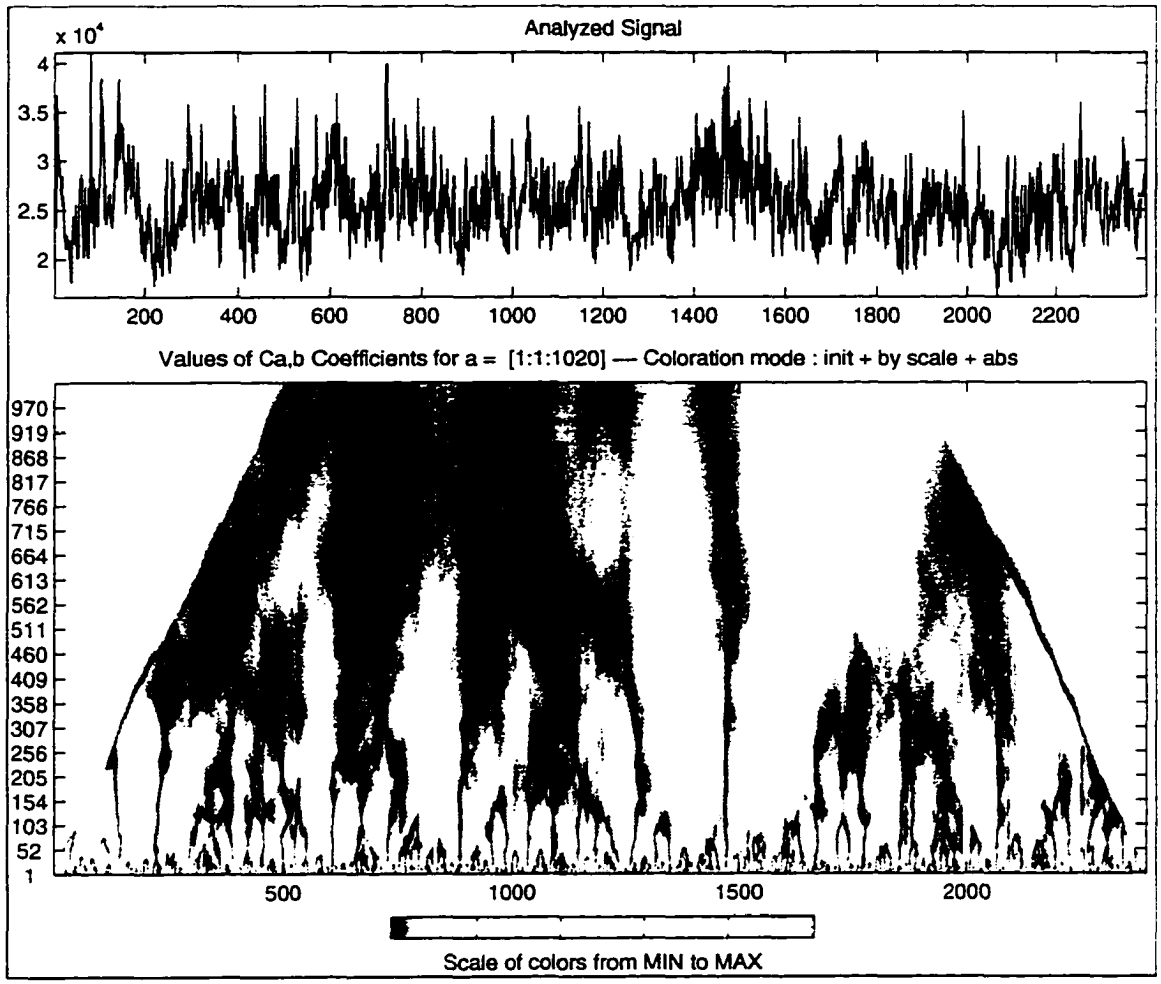

Figure 11 Continuous Wavelet Analysis of ATM traffic Volume-2

The Wavelet coefficients of the ATM traffic volume traces are shown in the illustrations. The vertical axis represents the scale factor in the Wavelet operation, which will be discussed in the next section - an intuitive understanding of the scale factor is that it represents the frequency components in traffic volume or the degree of the traffic variations. The horizontal axis represents the corresponding time stamps when these traffic volumes are recorded and Wavelet coefficients are generated. The absolute

values of the wavelet coefficients are represented by the darkness in the figure, i.e. the magnitude of the traffic volume changes relative to the specified scale factors during in a particular time interval.

It is illustrated in these figures that the traffic fluctuations at smaller time scale are repeated while the traffic fluctuations at larger time scale build on top of them and repeat the similar convergence patterns being presented in smaller time scales. This is a clear evidence of the self-similarity phenomenon in ATM traffic which was discovered and reported in [66]. This phenomenon becomes obvious by Wavelet Analysis. This example typifies the power of multiscale analysis with Wavelet techniques.

3.2 Wavelets and Multiresolution Decomposition

This section discusses the wavelets and multiresolution signal decomposition which lays the foundation for the discussions in ATM traffic signal processing, understanding, prediction and resource management.

3.2.1 Prior Art

Understanding real life signals has been one of the long term goals in signal processing. Traditionally, the linear signal expansion has always played a significant role in such analyses and understandings.

Given any signal x from a space S, it can be expanded into a linear combination

$$x = \sum_i \alpha_i \cdot \varphi_i \qquad (1)$$

where $\{\varphi_i\}_{i \in Z}$ is a set of elementary signals for space S.

If the set $\{\varphi_i\}$ is complete for the space S, there will also be a dual set $\{\varphi^*_i\}_{i \in Z}$ such that the expansion coefficients α_i can be computed as

$$\alpha_i = \int \varphi_i^*(t) \cdot x(t) dt \qquad (2)$$

assuming x and φ^*_i are integrable functions.

Karhunen-Loeve (K-L) Expansion is a special case of time series representation in particular one in which the basis functions, $\{\varphi_i\}$, yield statistically uncorrelated coefficients, $\{a_i\}$. There are at least two major problems with the K-L expansion: K-L expansion is signal dependent; K-L expansion is computationally complex since there is no fast algorithm available when $\{a_i, \varphi_i\}$ is arbitrary.

Fourier expansion[58],[59] breaks down a signal into sinusoids at various frequencies, i.e. it transforms the signal from the time domain to the frequency domain. However, the time-based information is lost during such transformation. When the signal is non-stationary, its transitory characteristics, such as drifts, trends, abrupt changes, the beginning and ends of events, are not detected.

The *Short-Time Fourier Analysis* provides a compromise. It analyzes the signal at small sections by applying the Fourier Analysis on windowed signal pieces. However, the window size is the same regardless of the frequencies being analyzed.

A figure comparing Fourier Analysis and Short-Time Fourier Transfer with Wavelet Series is given in the next section.

The Discrete Cosine Transform (DCT) is a commonly used approximations to the K-L expansion [90]. It is used in several standards for speech, image and video compression. However, the input stream to the DCT must be blocked in order to perform the transformation. Hence, the correlation across the boundaries is not removed and causing blocking effects, and is one reason for using lapped transforms and subband or wavelet schemes.

3.2.2 Characteristics of Wavelet

The *Wavelet Transform* was popularized by Grossman and Morlet [87], [88]. It is capable of localizing its analyses with various length at various time via its two basic operations: *scaling* and *shifting* [60]-[64].

The continuous wavelet transform is defined as:

$$CWT_f(a,b) = \frac{1}{\sqrt{a}} \int_R \psi^* \left(\frac{t-b}{a} \right) f(t) dt \qquad (3)$$

where $a \in$ R+ and $b \in$ R, and $\psi(t)$ is the impulse response of a bandpass filter with zero mean $\int_{-\infty}^{\infty} \psi(t)dt = \Psi(0) = 0$.

By defining

$$\psi_{a,b}(t) = \frac{1}{\sqrt{a}}\psi\left(\frac{t-b}{a}\right) \tag{4}$$

we get

$$CWT_f(a,b) = \left\langle \psi_{a,b}(t), f(t) \right\rangle \tag{5}$$

The *Wavelet Analysis* employs a variable windowing technique. It allows long time interval for analyzing the low frequency information whereas short time interval are used for high frequency information [63-64]. For example, the Haar wavelet is defined as

$$\psi_{m,n}(t) = 2^{-m/2}\psi(2^{-m}t - n), \qquad m,n \in Z \tag{6}$$

where

$$\psi(t) = \begin{cases} 1 & 0 \le t < 1/2, \\ -1 & 1/2 \le t < 1, \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

Here, m represents the scale factor, and n represents the shift factor. Notice that $\psi_{m,n}(t)$ is of length 2^m, and the shift is $2^m n$.

A comparison of the time-frequency analyses among Fourier Analysis, Short-Time Fourier Transfer and Wavelet Series is illustrated in Figure-12.

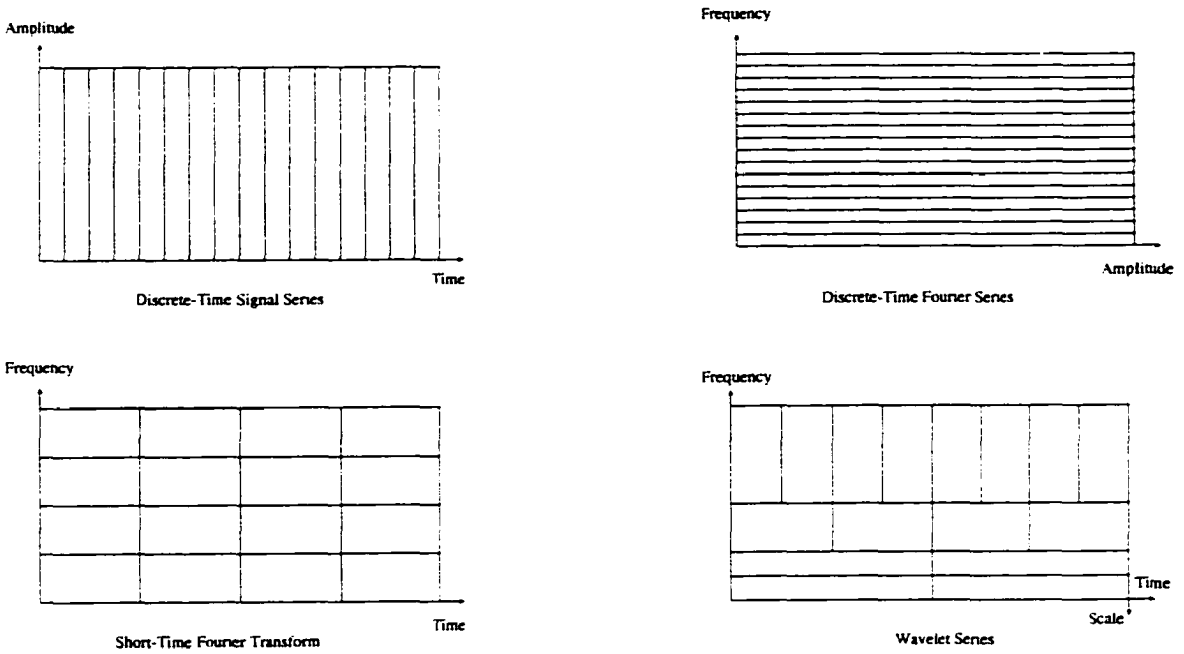

Figure 12 Comparison of Time Series, DFT, STFT and Wavelet Series

The wavelet approach is also linked to subband signal decomposition via the concept of multiresolution analysis.

A multiresolution analysis consists of a sequence of embedded closed subspaces

$$...V_2 \subset V \subset V_0 \subset V_{-1} \subset V_{-2}... \tag{8}$$

such that

Upward Completeness:

$$\overline{\bigcup_{m \in Z} V_m} = L_2(R) \tag{9}$$

Downward Completeness:

$$\bigcap_{m \in Z} V_m = \{0\} \tag{10}$$

Scale Invariance:

$$f(t) \in V_m \Leftrightarrow f(2^m t) \in V_0 \tag{11}$$

Shift Invariance:

$$f(t) \in V_0 \Rightarrow f(t-n) \in V_0, \quad \text{for all } n \in Z \tag{12}$$

Existence of a Basis:

There exists $\varphi \in V_0$, such that $\{\varphi(t-n) | n \in Z\}$ is an orthonormal basis for V_0. (13)

A given signal can be represented by a coarse approximation and added details, i.e. the details of a signal is the difference of the fine version and the coarse version of the signal. After applying such successive approximation recursively, the space of signals $L_2(R)$ is spanned by the spaces of these successive details at all resolutions.

Hence, a signal can be decomposed at multiple levels through Wavelet operations as illustrated in Figure-13, at each level, the signals decomposed into a coarse and a fine component.

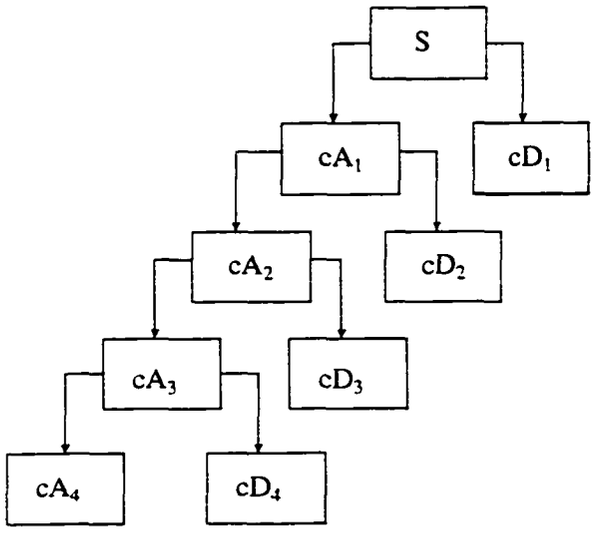

Figure 13 Wavelet Signal Multi-level Decomposition

The coarse approximation of the signal is concise and requires less number of coefficients terms, the details of the signal are more precise and requires greater number of coefficients terms. Figure-14 shows the distribution of Wavelet coefficients in the frequency (scale) domain.

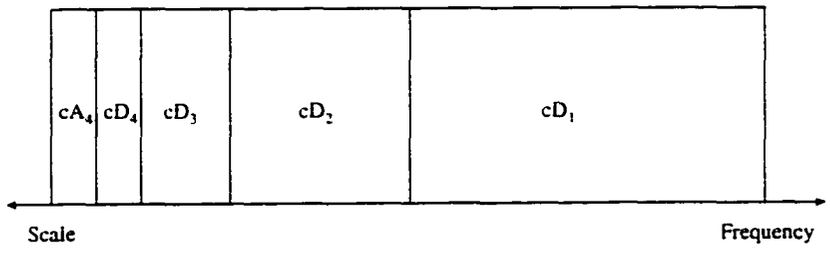

Figure 14 Wavelet Coefficients Distribution In Frequency (Scale) Domain

A signal could be studied very well through a set of its corresponding Wavelet coefficients. Typically, the *zoom-in* process, which starts from the

coarse down to the fine, provides the capability to analyze the signal in all levels of details based on different needs via appropriate scaling. All portions of the signal are analyzed through the appropriate shifting. Figure-15 illustrates the effective Wavelet zoom-in capability.

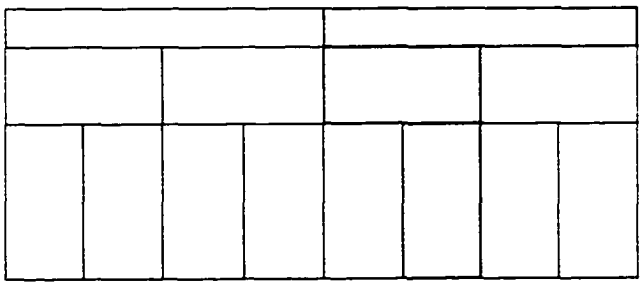

Figure 15 Wavelet Zoom-In Capability

3.2.3 Multiresolution Decomposition Operation

An efficient implementation of the Discrete Wavelet Transform is the recursive use of a set of Wavelet filtering processes, which also called the *Fast Wavelet Transform.* In this two-channel subband coder as shown in Figure-16, the L is a low pass filter which generates the *Approximations*; the H is a high pass filter which generates the *Details.* A downsampling operation immediately follows each of these steps.

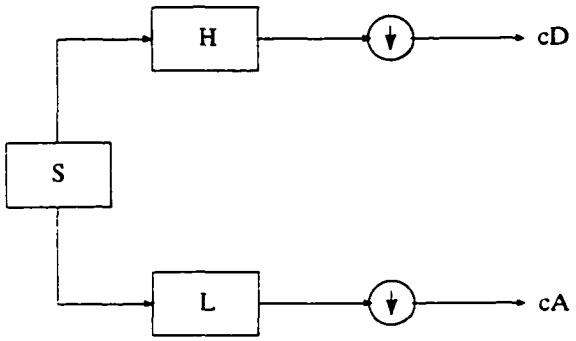

Figure 16 Wavelet Filtering Process

A necessary condition for the H and L filters is that

$$H^*H+L^*L=1 \tag{14}$$

where * is the complex conjugate.

The selection of the filters provides challenges to the studies in Wavelet theorem and their applications. Among the widely used Wavelets [61],[65], there are *Haar, Daubechies, Biorthogonal, Coiflets, Symlets, Morlet, Mexican Hat, Meyer*, etc. The *Haar* and *Daubechies wavelets* [62],[65] are illustrated in Figures 17 through 19.

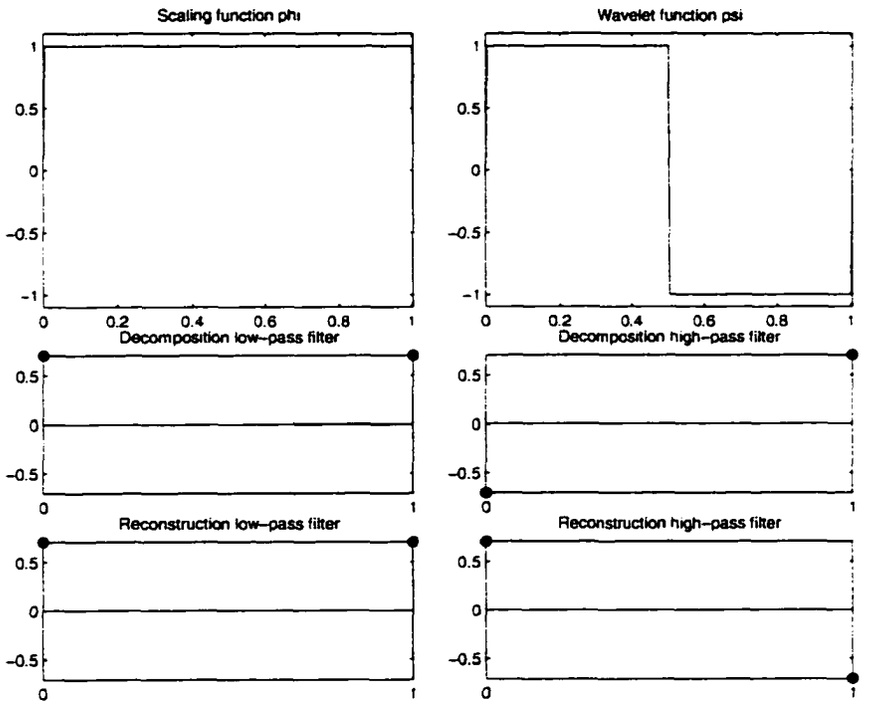

Figure 17 Scaling & Wavelet function, Decomposition & Reconstruction filters of Haar

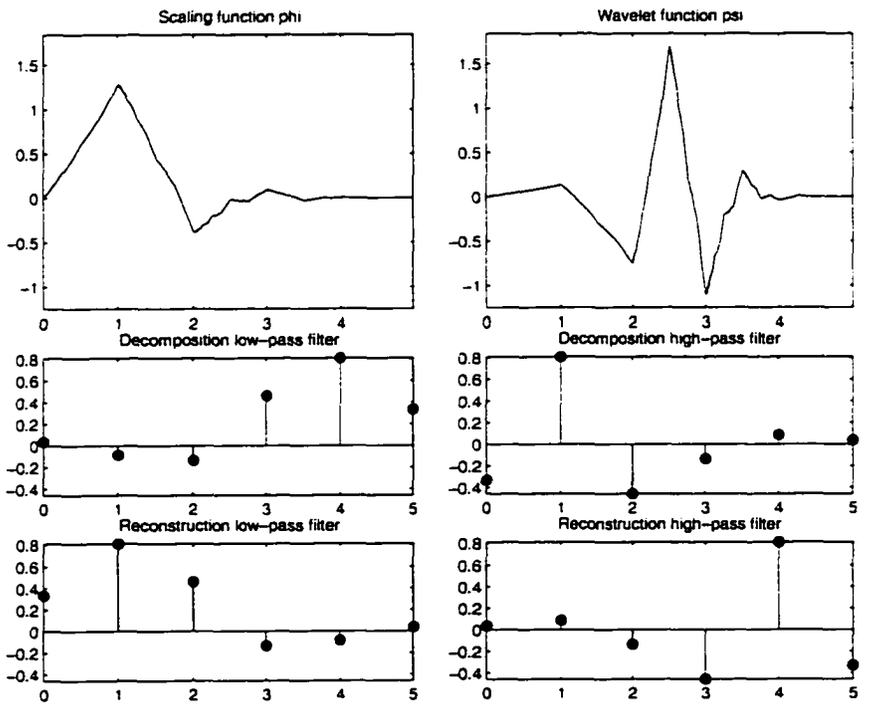

Figure 18 Scaling & Wavelet function, Decomposition & Reconstruction filters of DB-3

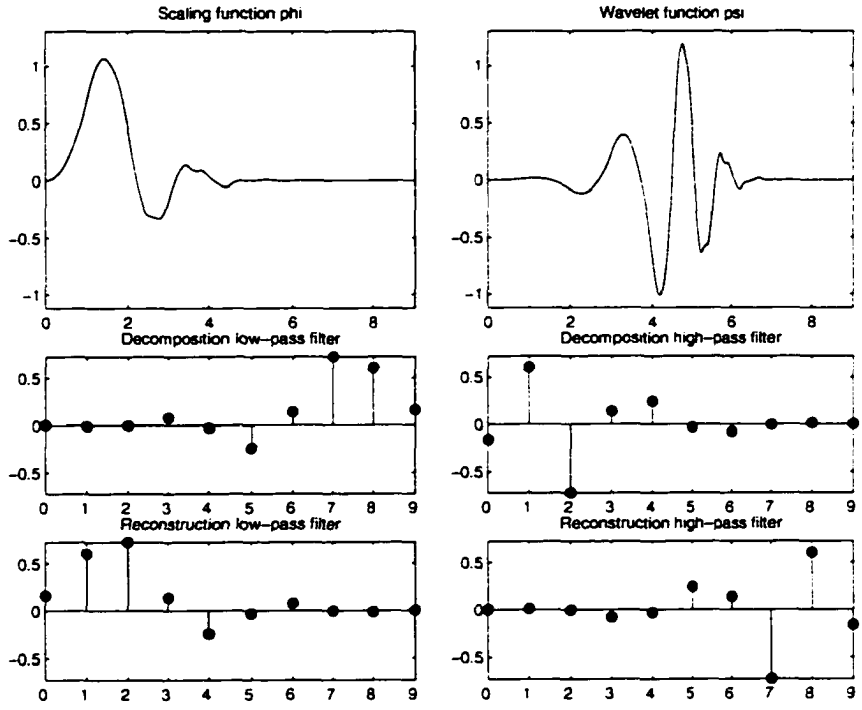

Figure 19 Scaling & Wavelet function, Decomposition & Reconstruction filters of DB-5

Generally, the subspace should be orthogonal, i.e.

$$HL^* = 0 \tag{15}$$

in a decomposition to eliminate the correlation in the filtered signal sequences.

The cascaded sets of the filtering and downsampling operations constitute the *Wavelet Signal Decomposition process*.

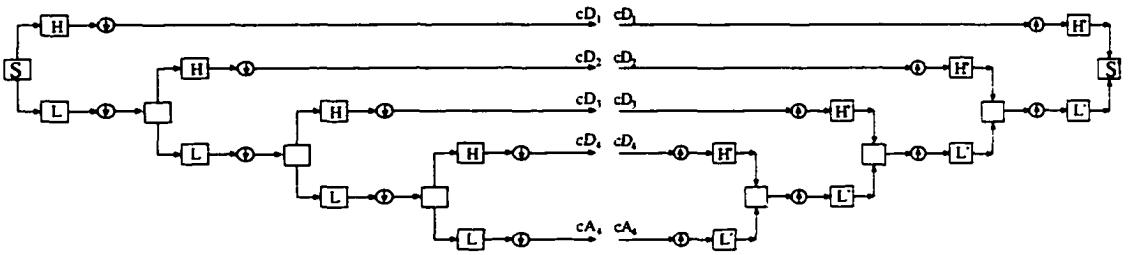

Figure 20 Wavelet Signal Decomposition and Reconstruction Process

The *Reconstruction process* is a reversed version of the Decomposition operation, and is called the Inverse Discrete Wavelet Transform (IDWT), where H and L are replaced by their quadrature mirror filters H' and L' and the downsampling operation is replaced by the upsampling operation. Figure-20 illustrates both the Wavelet signal decomposition process and the reconstruction process.

By selecting the interested components during the Wavelet signal decomposition and feeding these specific components to the reconstruction process, the *signal subband filtering* is accomplished, i.e. the signal is decomposed at specific resolutions as shown in Figures 21 and 22.

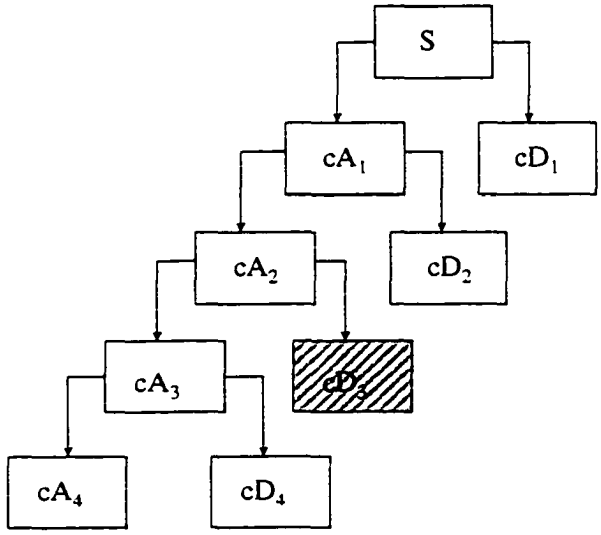

Figure 21 Signal Extraction During Multiresolution Decomposition

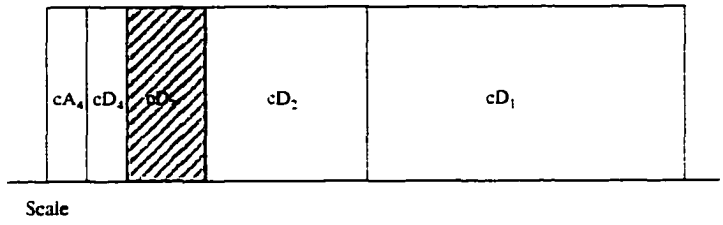

Figure 22 Effective Subband Filtering Through Multiresolution Decomposition

For example, the signal S(t)'s components at D_3 level could be obtained by the forwarding the cD_3 component in the decomposition operations to the reconstruction operations in Figure-23.

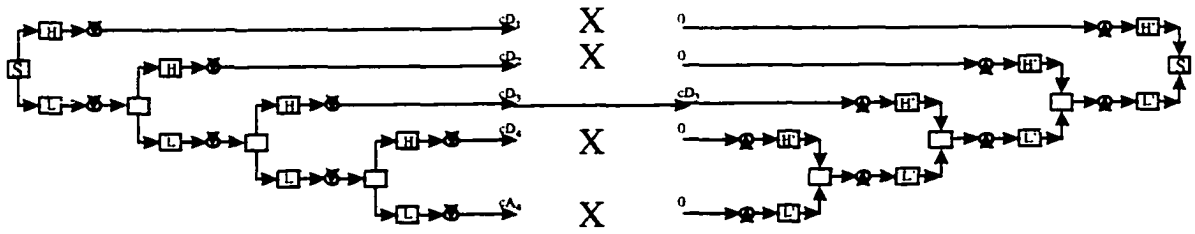

Figure 23 Process of Subband Filtering

Figure-24 shows the direct operations of the signal feature extraction for the details at D_3 level.

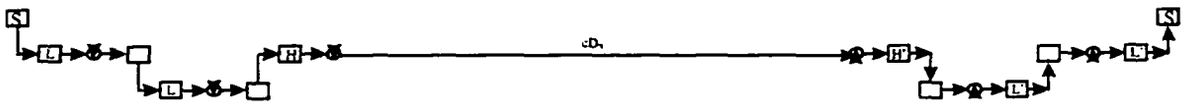

Figure 24 Example of Subband Filtering - Extraction of $D_3[S(t)]$

3.3 Variable Bit Rate Traffic and Their Multiscale Decomposition

The wavelet multiscale decomposition technique provides an efficient yet flexible mechanism for studying the ATM traffic volume at multiple time scales. The effective traffic bandwidth modeling based on multiscale analysis and neural network learning and prediction in order to provide an integral ATM traffic management solution is our goal. To continue our investigation, we start with a discussion on *Variable Bit Rate* (VBR) traffic.

Due to the fact that telecommunication technologies have been experiencing significant changes during recent years [67]-[70], evolving from the circuit switched network toward studies of the packet/cell oriented network implementation, the traffic being carried is becoming more diversified and includes multimedia communication such as voice, data, image and video, etc. Among the potentially vast amount of image and video application, the multimedia traffic such as JPEG & MPEG encoded

data has drawn more interest in telecommunication traffic research and applications.

Because of the nature of these encoding schemes, such traffic has *Variable Bit Rate* (VBR) as shown in Figure-4. VBR traffic is also sensitive to delay because of its application nature as illustrated in Table-2. When resources are reserved based on the average rates of the VBR sources, unacceptable delays and packet losses may happen when the sources are transmitting near their peak rate; when a static or constant service rate based on peak bit rate is used in order to maintain the QoS, the bandwidth utilization tends to be very low in these networks. These bring up the challenges in bandwidth utilization for VBR traffic. Hence, MPEG VBR traffic have been chosen as the primary test data for verifications of our challenging research work in dynamic bandwidth allocation. A varieties of the test datasets are available as described in Appendix A, ranging from single source MPEG, multiple sources MPEG, to ABR/LAN traffic.

With the Wavelet's zoom-in capabilities, the ATM traffic volume can be decomposed into multiple components at different time scales. In order to demonstrate the concept of a multiscale decomposition of the VBR traffic, we use a Haar Wavelet and apply the multiscale decomposition

operation repeatedly on a trace of MPEG GOP[i] traffic volume into D1 and A1, and subsequently A1 into D2 and A2, ..., A5 into D6 and A6.

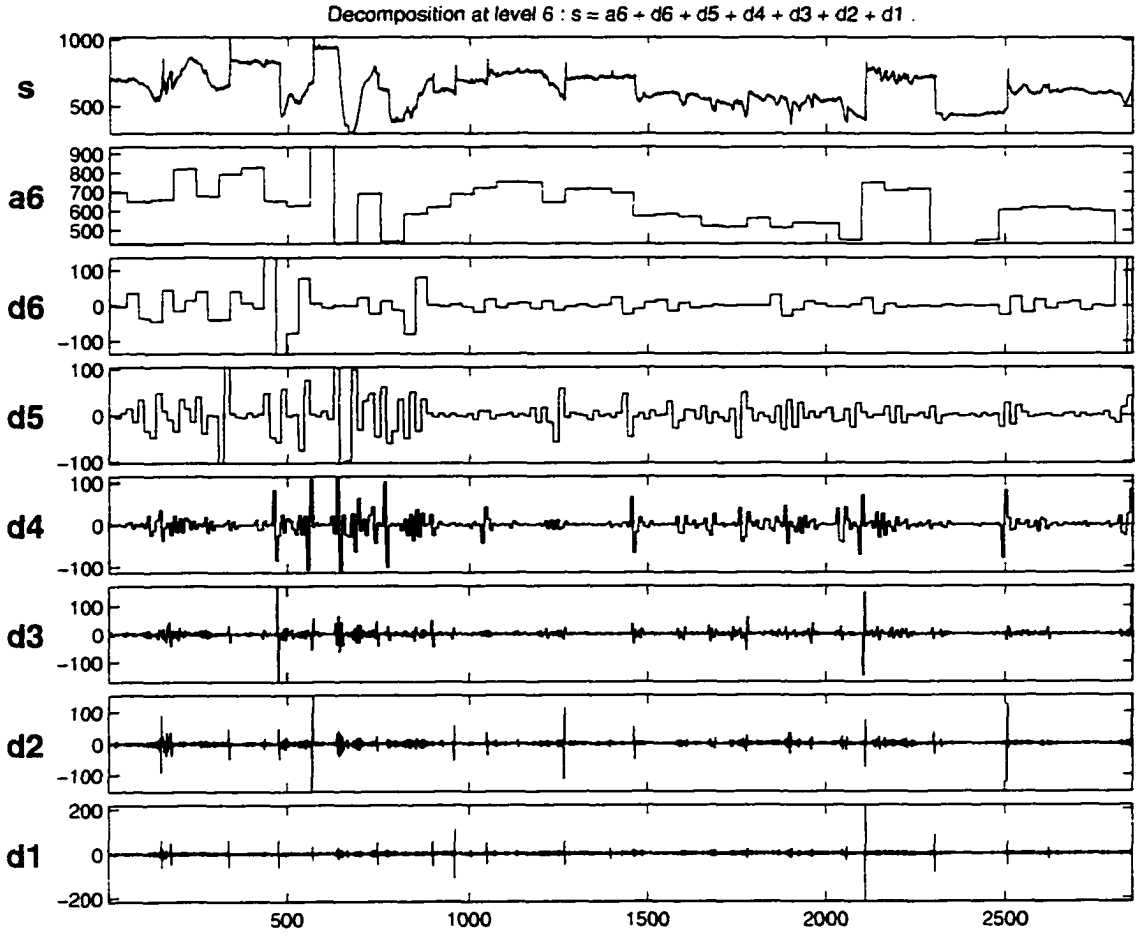

Figure 25 Multiscale Decomposition Of Traffic Volume

The original traffic volume and the decomposed components a6, d6, d5, d4, d3, d2, and d1 are illustrated in Figure-25. The approximation component a6 is overlapped with the traffic volume trace in Figure-26. The decomposed signal detail components, d1 through d6, are overlapped in

[i] A duration of 0.5 second per Group of Pictures (GOP) in this example

Figure-26 (b), and the Figure-26(c) shows the these Wavelet coefficients in gray levels.

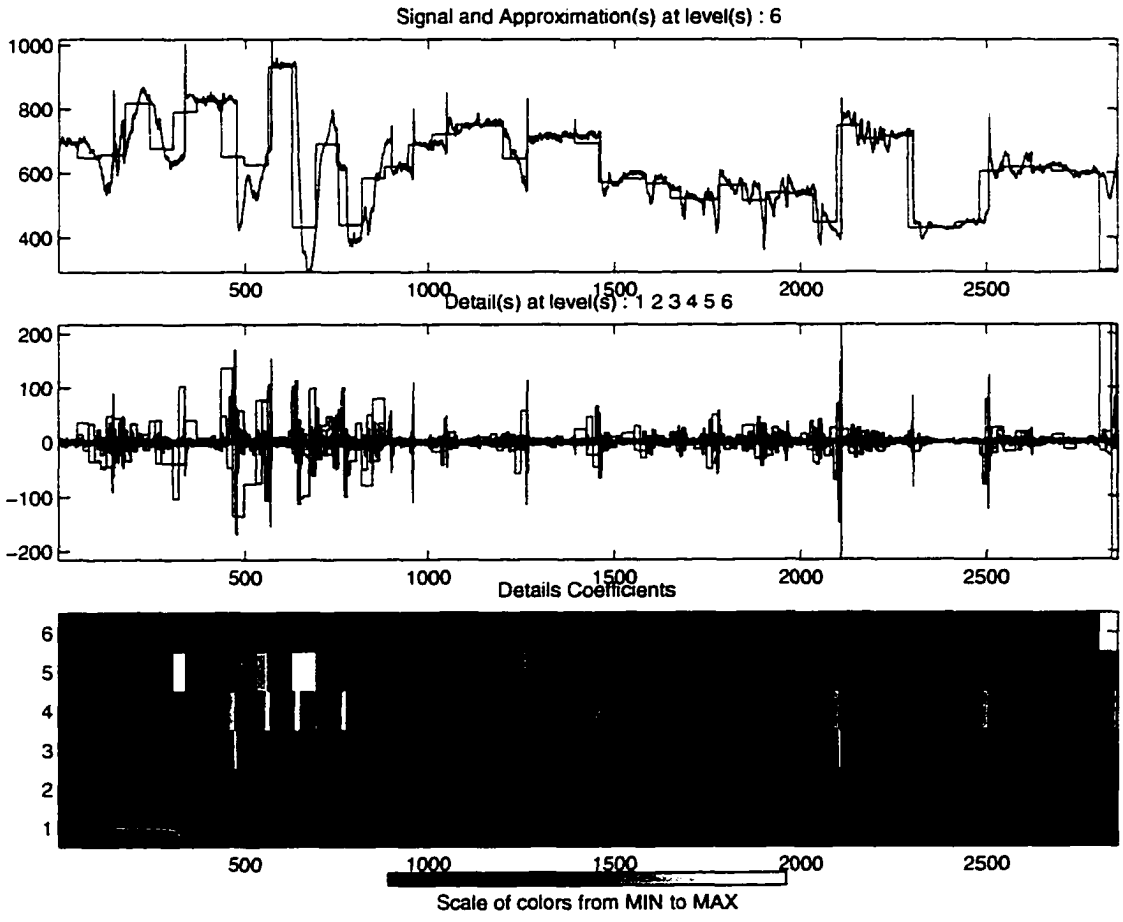

Figure 26 (a) Traffic Volume Overlapped with Low Frequency Component; (b) High Frequency Components; (c) Wavelet Coefficients

In Figures 25 and 26, the decomposed approximation at level 6 (a6) is chosen for visual inspection and it is updated every 16 seconds. It does piece-wisely represent the major bandwidth components, and the details d1, d2, ..., d5 and d6 represent the traffic fluctuations components at the

intervals of 0.5 second, 1 second, ..., 8 seconds and 16 seconds, respectively.

Here, $\delta = \frac{|S - A6|}{A6}$ may approach 10%, which translates to a possible large buffer requirement and hence a large delay may be introduced if only A6 is utilized as the bandwidth allocation requirement. By comparing the |S-A6| and the accumulations of |D1| through |D6|, it is also noticed that the bandwidth shortage based solely on A6 component could be compensated by the contributions from |d1| through |d6| in order to achieve less delays.

Our studies, which is expanded in Section 3.5, revealed that these multiple time-scale components indeed correspond to their frequency components. A general observation is that the decomposition at the low frequency component does piece-wisely reflect the major bandwidth requirements; the high frequency components contain the traffic fluctuation, which is generally eliminated by buffering; some of the intermediate components may be smoothed by buffering as well, yet some may require the increases in the path bandwidth allocation.

This approach is a promising candidate for representing the dynamic bandwidth of ATM traffic volume.

3.4 Representation of Effective Dynamic Bandwidth

From the above description, the low frequency components have higher correlation with the MPEG traffic volume - the bandwidth, while the higher frequency components can be eliminated by the buffering schemes. The intermediate frequency components could be partially smoothed by buffering. However, due to the QoS constraints, appropriate bandwidth calculation with these components deems to be necessary in order to satisfy the maximum delay requirement.

With the discussions in the previous section and the statistical verification and software simulation in the next section, we propose that the effective dynamic bandwidth be:

$$B(t) = B_{n,m}(t) = A_n[S(t)] + A_n\left[\sum_{i=m}^{n} \left| D_i[S(t)] \right|\right] \qquad (16)$$

where $A_n[.]$ denotes the n^{th} Approximation of a signal; $D_i[.]$ denotes the i^{th} Detail of a signal; m is a parameter related to the bandwidth utilization rate, for example, m=3; n is a parameter defining the updating interval for the effective dynamic bandwidth. For example, the bandwidth updating interval is 2^n=16, when n=4. Note that B(t) remains the same during each update interval, 2^n, i.e.

$$B(k \cdot T_s + a_t) = B(k \cdot T_s), \qquad \text{for any integer } k \text{ and } a_t \in [0, 2^n),\ T_s = 2^n \quad (17)$$

The first term, A_n, in (16) represents the n^{th} Approximation of S(t) through the multiscale decomposition. The magnitudes of the details D_i's of S(t), are also utilized by taking their absolute values and subsequently applying the n^{th} approximation. The summations of these terms is used for the bandwidth approximation.

A more general form of the effective dynamic bandwidth, which satisfies more stringent QoS requirements[i], is provided as:

$$B(t) = \lambda_0 \cdot A_n[S(t)] + A_n\left[\sum_{i=1}^{n} \lambda_i \cdot |D_i[S(t)]|\right] \quad (18)$$

Definition in (18) is equivalent to (16) when

$$\lambda_0 = 1,\ \lambda_1 = \lambda_2 = \ldots = \lambda_{m-1} = 0 \quad (19)$$

$$\text{and} \quad \lambda_m = \lambda_{m+1} = \ldots = \lambda_n = 1 \quad (20)$$

3.5 Effective Multiscale Decomposition In Multimedia Traffic

The Effective Dynamic Bandwidth representation is evaluated by examining its statistical characteristics with the original ATM traffic and the resource utilization when EDB is used for bandwidth allocation.

Appendix-A provides a list of the test data used for our evaluation, including both the MPEG coded VBR traffic traces and LAN based ABR

traffic traces. There are over 1,036,800 traffic trace volumes used in our evaluation, which represents a total of over twelve (12) hours of MPEG traffic. There are another 1,000,000 traces of LAN traffic used in our evaluation which represents about an hour of ABR traffic.

3.5.1 Statistical Characteristics of Effective Dynamic Bandwidth

We have studied the statistical similarities between the effective components via the Wavelet multiscale decomposition and the original traffic volumes. The traffic volume of each GOP is used for these evaluations. A bandwidth update interval of 16 GOP's, or 8 seconds is used for practical reason. As shown in Figure-27, the corresponding decomposed approximation at level 2^4, $A_4(t)$, of a single VBR/MPEG source, provides the major traffic volume piecewisely.

For example, cell loss rate$<10^{-9}$

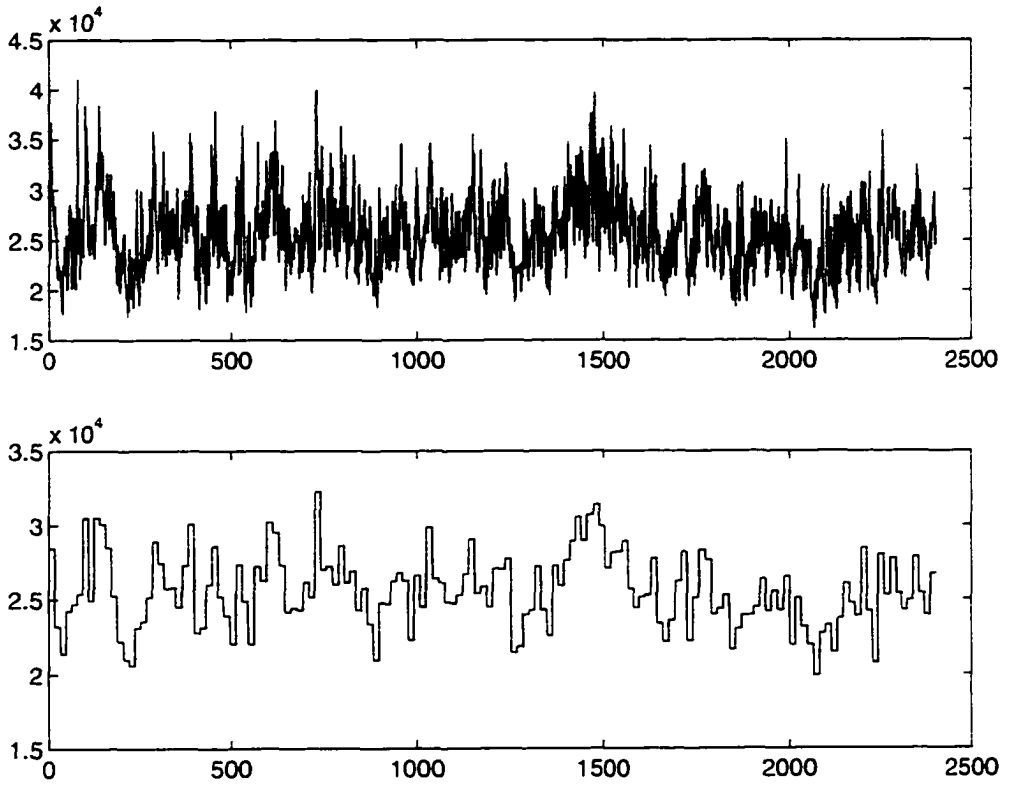

Figure 27 Single VBR - Traffic Volume And Its Low Frequency Components $A_4(t)$

The effective dynamic bandwidth B(t) = $B_{4,3}(t)$ is calculated following (16). It consists the approximation A_4 and the 4^{th} approximation of the magnitudes of the details from D_1, D_2 and D_3. As illustrated in Figure 28, B(t) provides an envelope-like curve above the actual traffic volume. It is updated at the same interval as $A_4(t)$, which is every 8 seconds as previously mentioned.

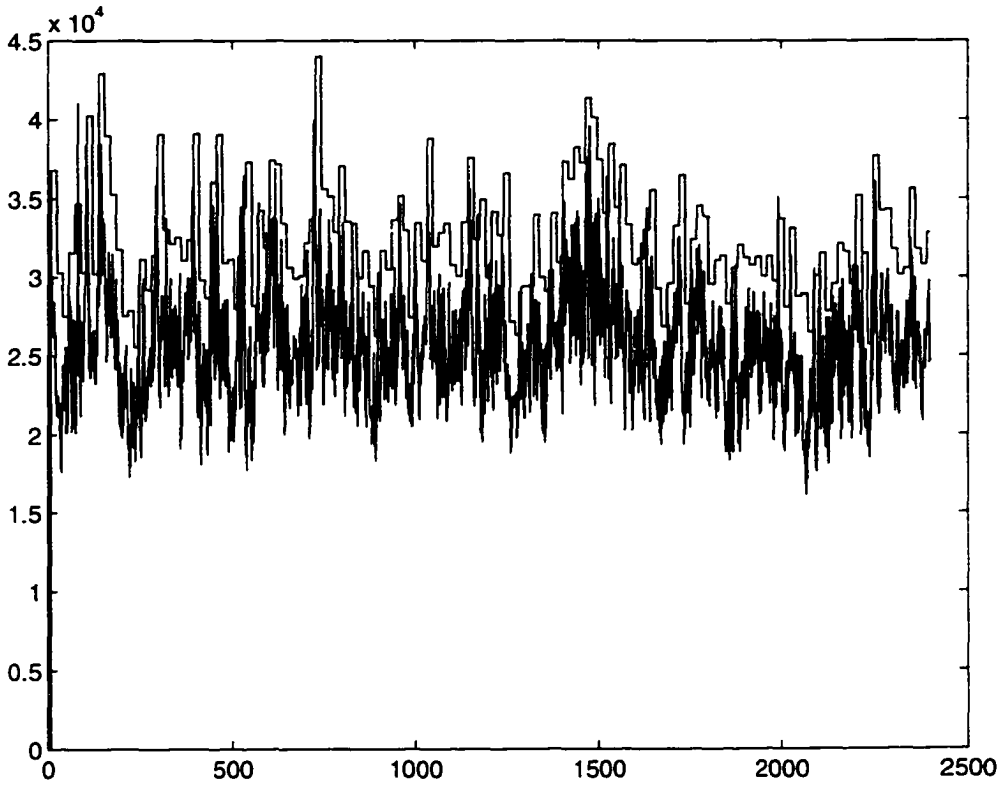

Figure 28 Single VBR - Traffic Volume and its Effective Dynamic Bandwidth

For the effective dynamic bandwidth B(t) to serve as a bandwidth allocation requirement, it is necessary to demonstrate its similarities to the corresponding traffic volume in respect to the statistical characteristics [7]. Hence, the first order and the second order statistics of effective dynamic bandwidth and VBR traffic volume will be illustrated in this section.

Over one million VBR traffic traces have been utilized in our evaluations and the evaluation results based on a piece of 2400 traffic traces are shown in the following figures.

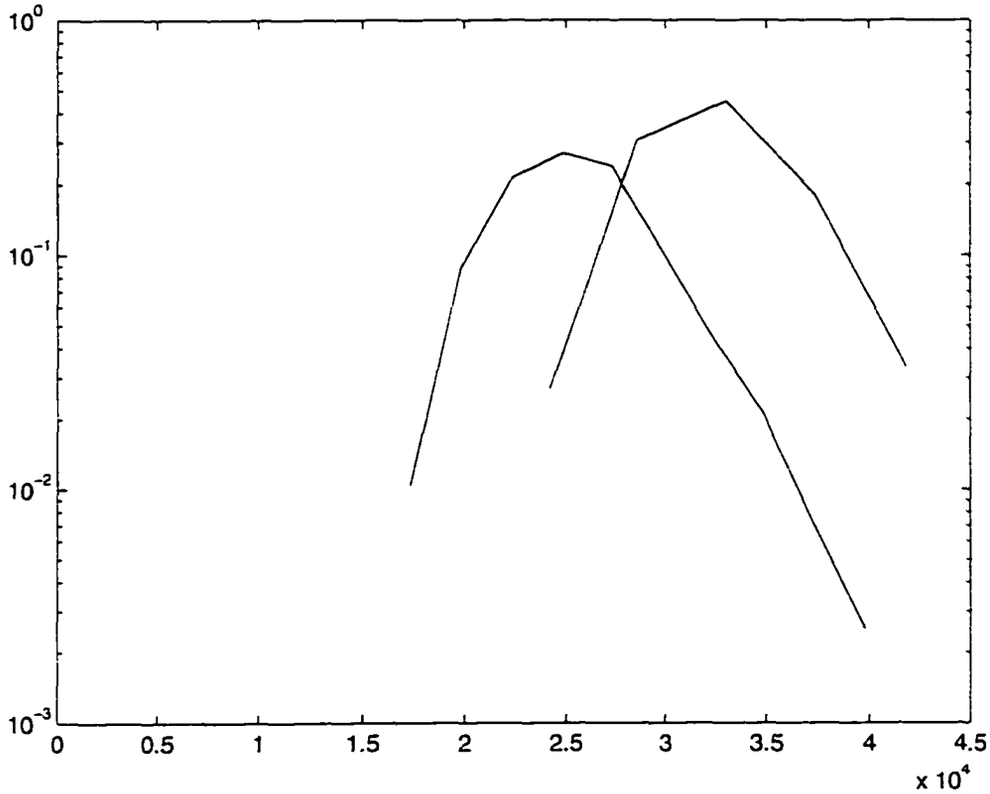

Figure 29 Single VBR Source - PDF's Of S(T) And B(T), S(T) On Left, B(T) On Right

Figure-29 shows the similarity of the *Probability Density Function* (PDF), a first order statistics, between B(t) and S(t). The distributions are similar. However, the B(t)'s mean is shifted to the right which indicates its characteristics as an envelope function of S(t).

The *Auto-Correlation Function* (ACF), a second order statistics as in Figure-30, also demonstrates significant similarity of B(T) and S(t). This shows the substantial similarities in their second order dynamics - the variations. It also indicates that the effective dynamic bandwidth, B(t),

varies with changes in the traffic volume, however, B(t) is only updated every 8 seconds in this case, at a much slower pace than S(t).

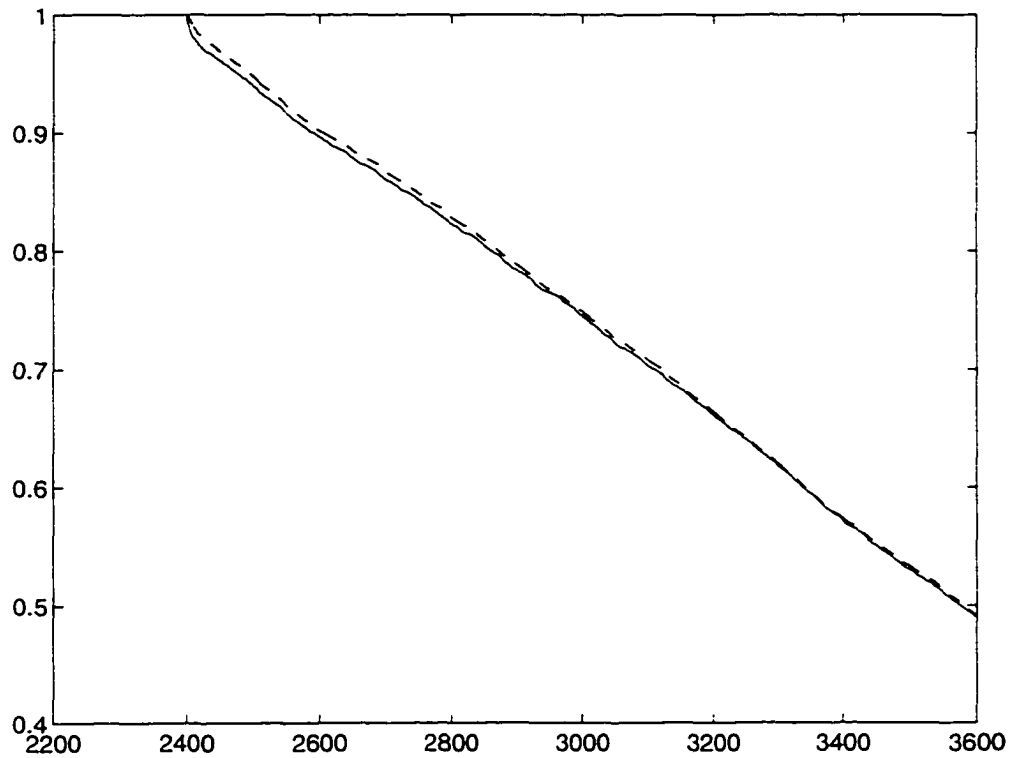

Figure 30 Single VBR Source - ACF's Of S(T) And B(T), S(T) On Solid, B(T) On Dash

The statistical characteristics for other ATM traffic are illustrated in the next figures.

To simulate the effect of multiple traffic sources arriving at an ATM switch node, we multiplexed five MPEG traffic traces from various sources in Appendix A, which consist of movies with fast and slow motions as well as soccer play, to achieve a representative multiplexing of multiple traffic sources. The multiplexed traffic volume is in Figure-31(a).

Assuming the bandwidth allocation of the ATM switch is updated every 8 seconds, the approximation A_4 of the multiplexed VBR/MPEG sources is calculated following the flow in Figure-23. As in the previous simulation, the approximation A_4 of the multiplexed VBR/MPEG sources, also provides a piecewised bandwidth estimate as illustrated in Figure-31(b). The effective dynamic bandwidth, $B_{4,3}$, is obtained following (16), which is calculated based on the Approximation, A_4, with compensations from the Details, D_1 through D_3. It provides a bounding function for the bandwidth requirement, as illustrated in Figure-32. Both the approximation A_4 and the effective dynamic bandwidth $B_{4,3}$ are updated every 16 GOPs.

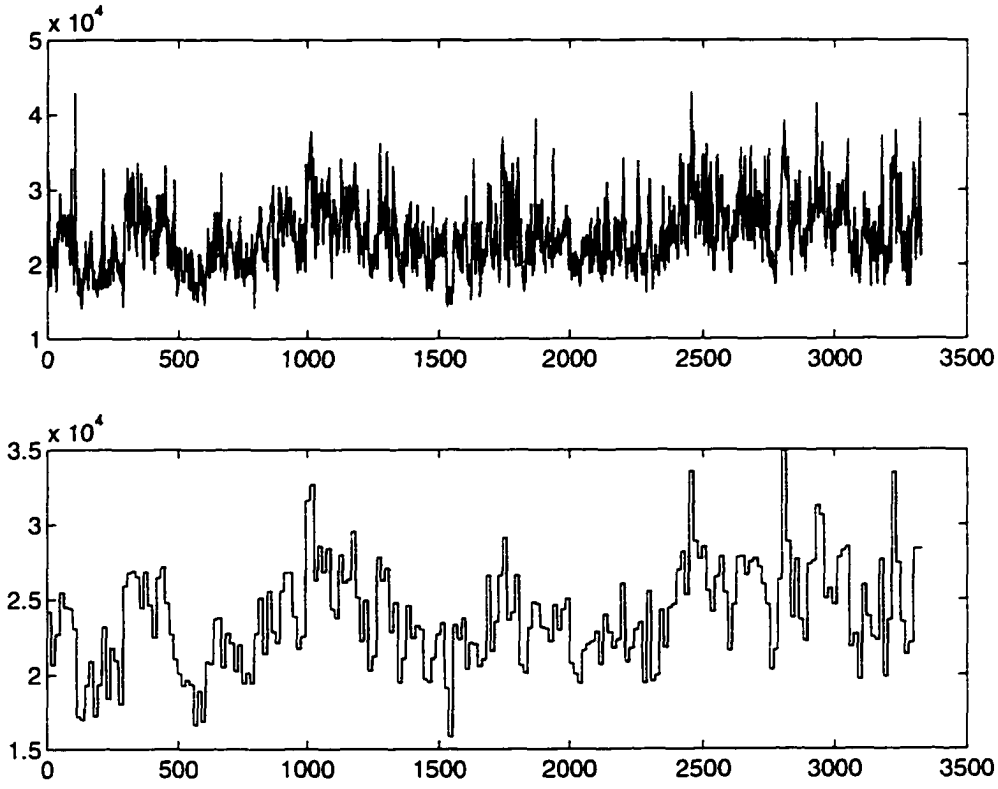

Figure 31 (a) Multiplexed VBR Traffic Volume; (b) Low Frequency Components, $A_4(t)$

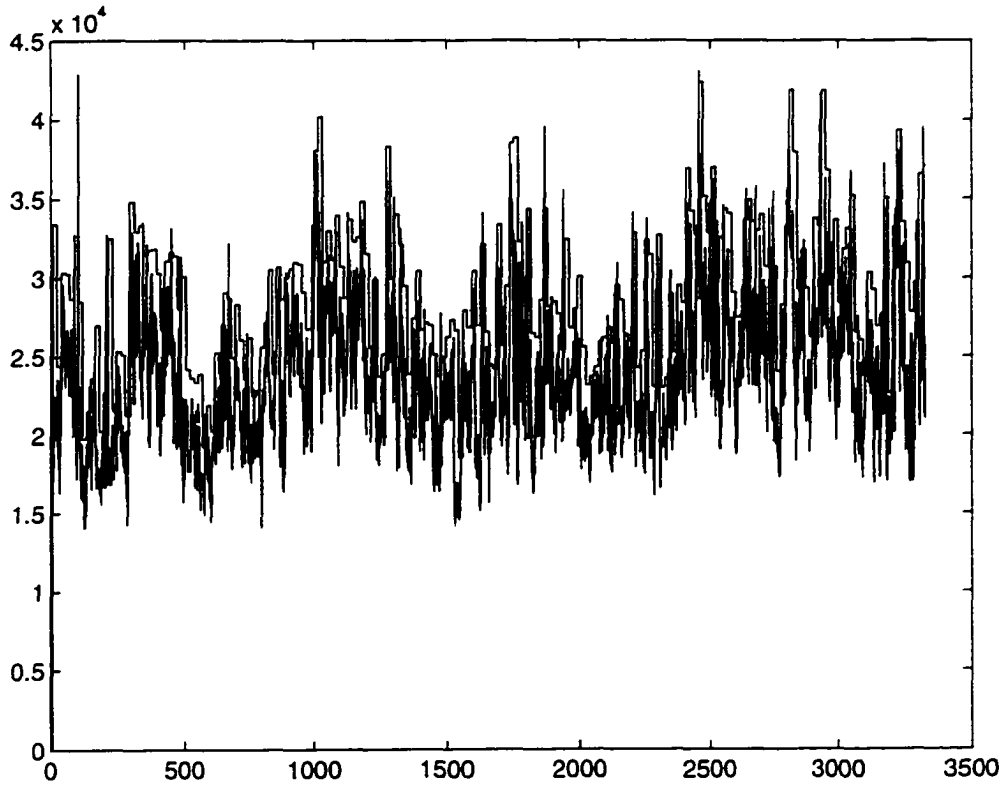

Figure 32 Multiplexed VBR Traffic Volume And Its Effective Dynamic Bandwidth

The Probability Density Function and the Auto Correlation Function also demonstrate the significant similarities between B(t) and S(t), for the multiplexed traffic as in Figure-33 and Figure-34.

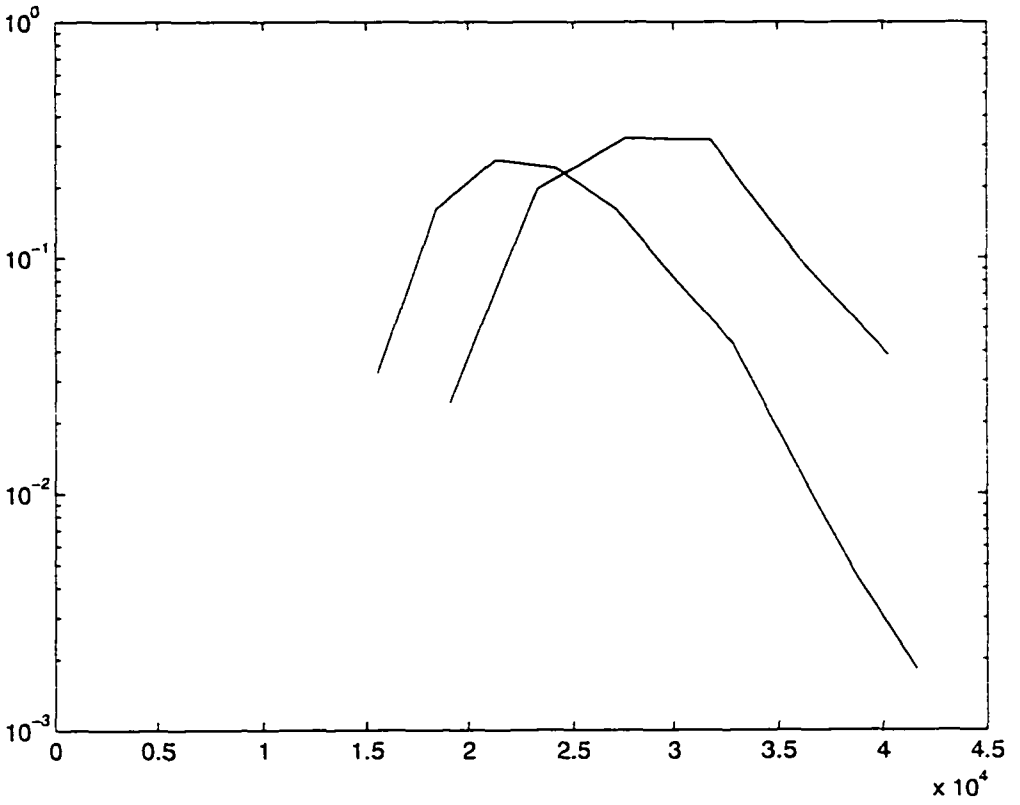

Figure 33 Multiplexed VBR Source - PDF's Of S(T) And B(T), S(T) On Left, B(T) On Right

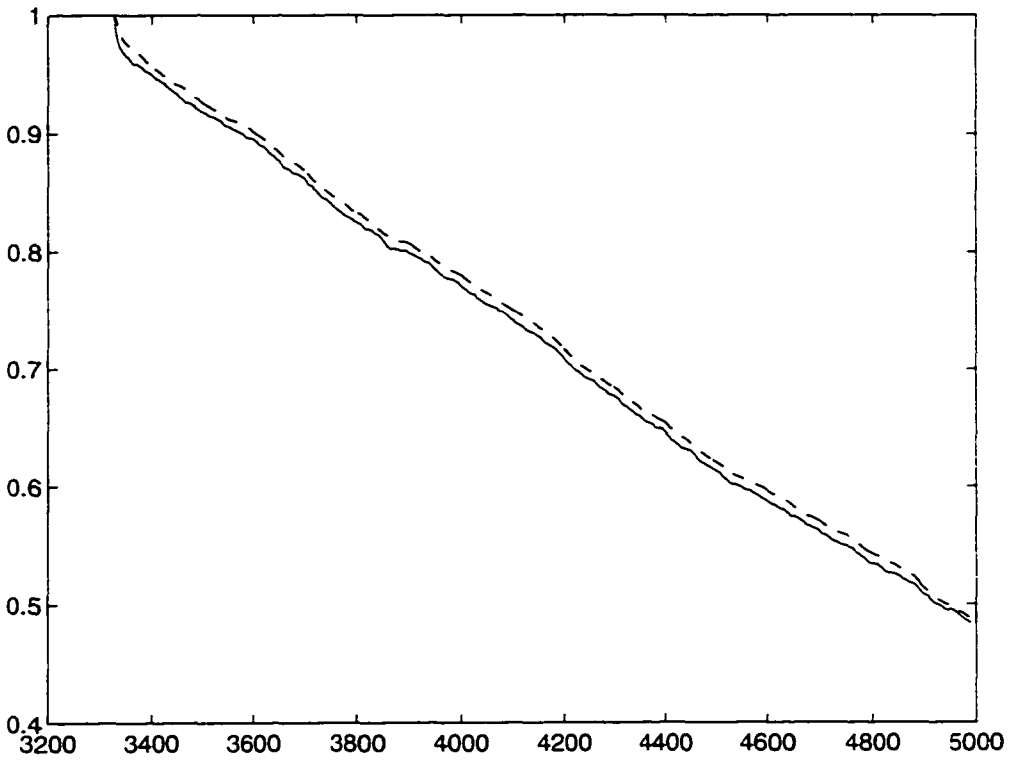

Figure 34 Multiplexed VBR Source - ACF's Of S(t) And B(t), S(t) On Solid, B(t) On Dash

We have further applied the Effective Dynamic Bandwidth model to Available Bit Rate (ABR) traffic. The traffic volume traces from Local Area Network (LAN) typifies the ABR traffic. Such traffic as detailed in Appendix A, is used for our evaluation. Due to the nature of the ABR traffic, it is delivered on the best effort basis to achieve its economic result and the allowed delay is generally in 1000 ms or even longer. The traffic data in every 500 ms interval is grouped together.

Figure-35(a) illustrates the traffic volatility on a LAN during an one hour period in a corporate network environment. The multiscale analysis is applied to derive the approximation (A_4), as in Figure-35(b) and the details are compensated to achieve the effective dynamic bandwidth ($B_{4,3}$), as in Figure-36, which is updated every 8 seconds.

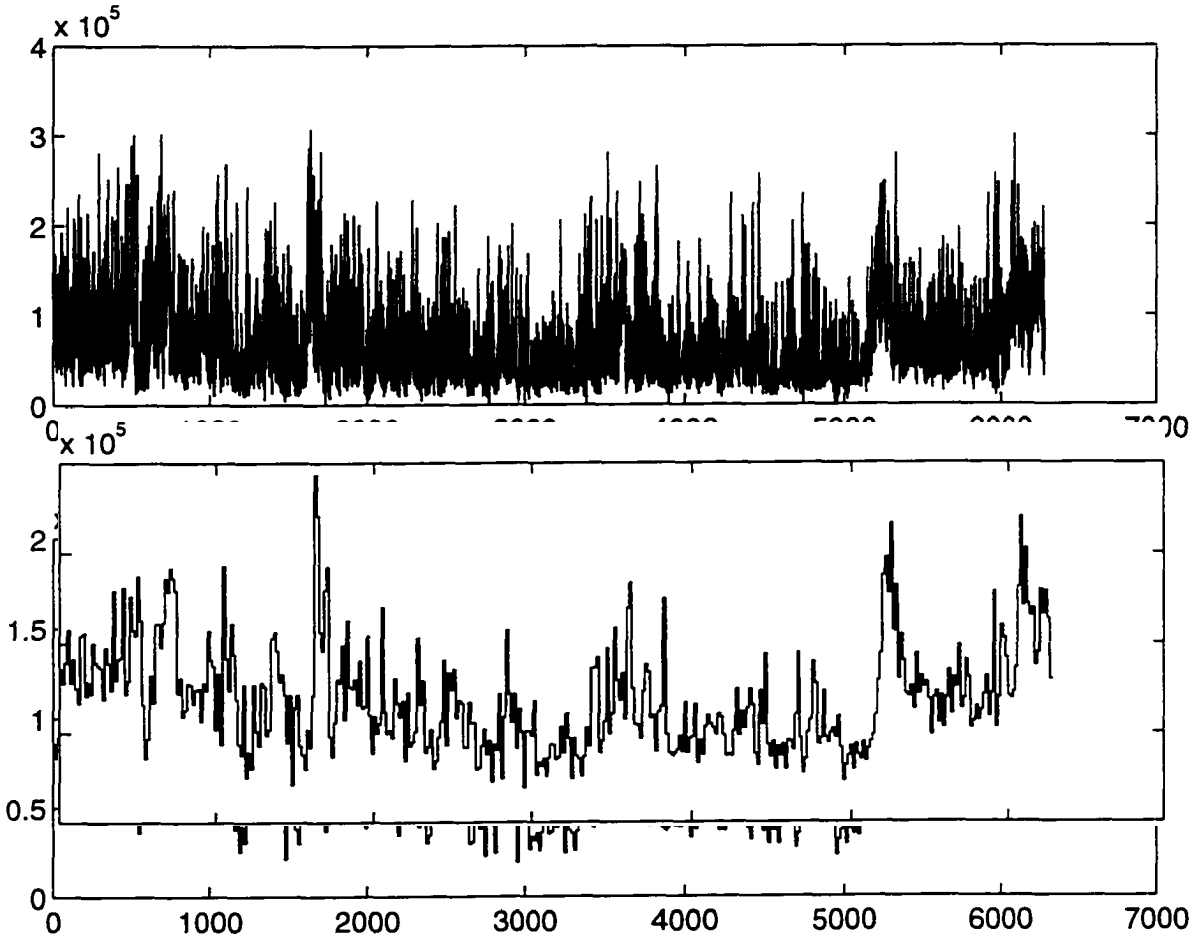

Figure 35 (a) Multiplexed ABR Traffic Volume; (b) Its Low Frequency Components

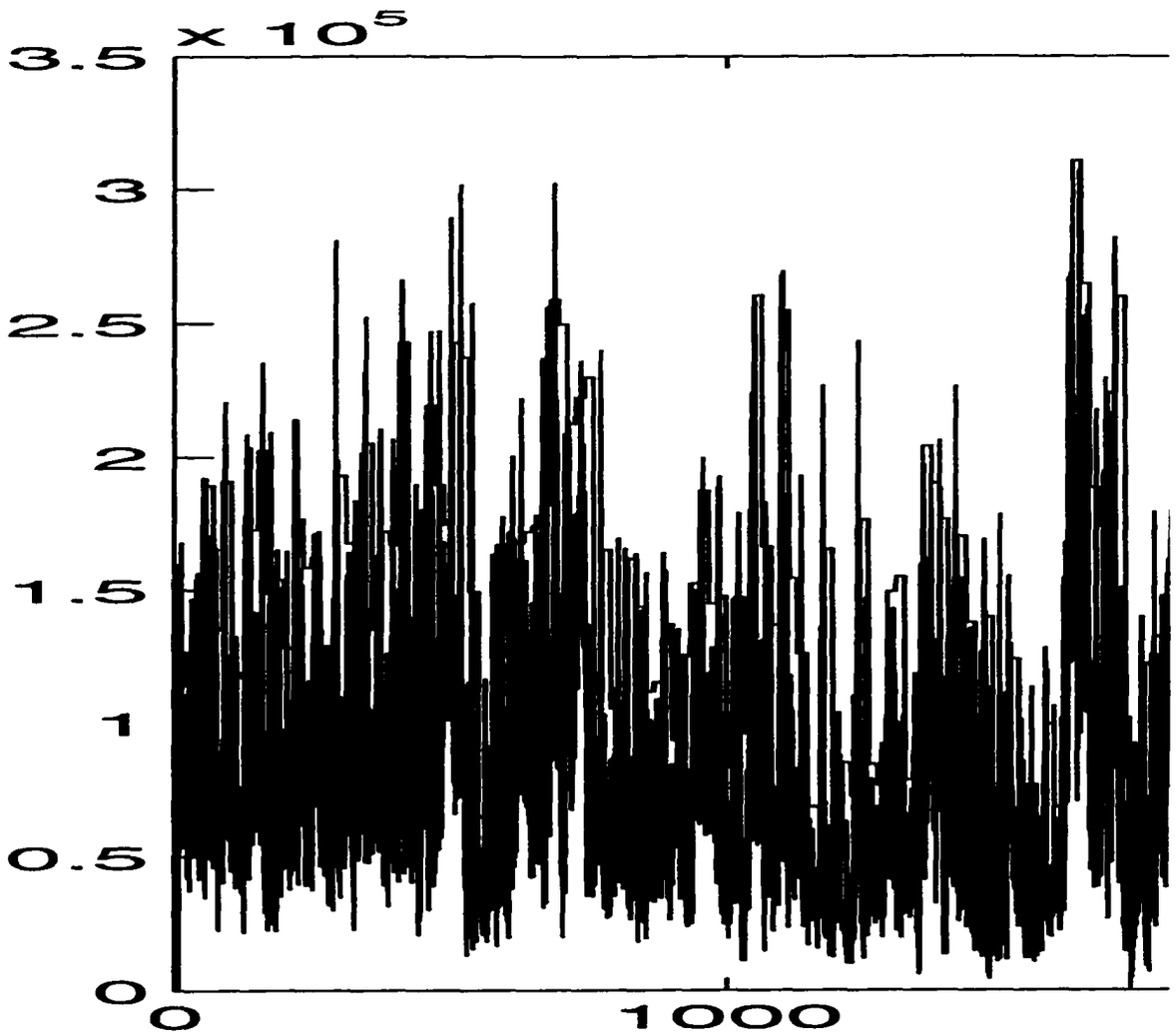

Figure 36 Multiplexed ABR Traffic Volume And Its Effective Dynamic Bandwidth

The Probability Density Function of the S(t) as presented in Figure-37 shows more concentration in the low traffic volume area indicating the silent durations in the monitored period, which is a typical ABR traffic characteristics. During those silent periods, the ATM switch node utilizing the effective dynamic bandwidth will allow the transmission of accumulated traffic in its buffers, thus, the Probability Density Function of B(t) has its peak around the $1x10^5$ area as shown in the same figure. The Auto

Correlation Functions in Figure-38 indicates more variations for S(t), which confirms the volatility of the ABR/LAN traffic. It also demonstrates that ACF of B(t) is smoother and shares the same trend as for S(t). Both ACF's will be close enough when normalization is applied.

These results demonstrate the robustness of the effective dynamic bandwidth on multiplexed ABR/LAN traffic.

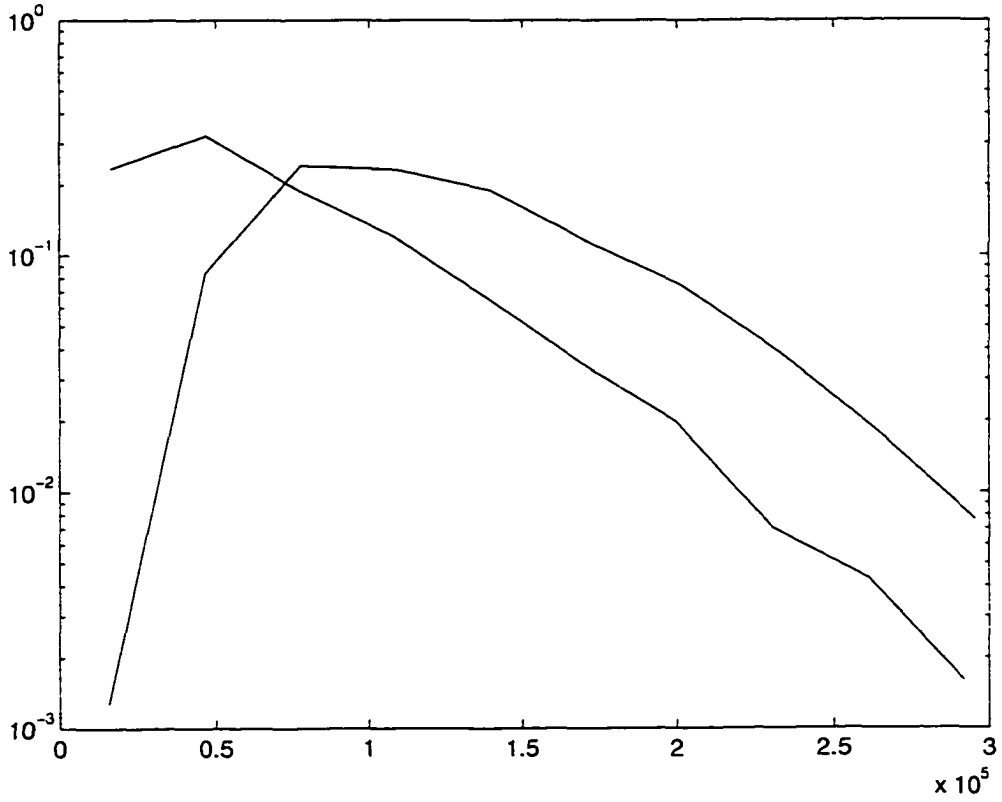

Figure 37 Multiplexed ABR Source - PDF's Of S(t) And B(t), S(t) On Left, B(t) On Right

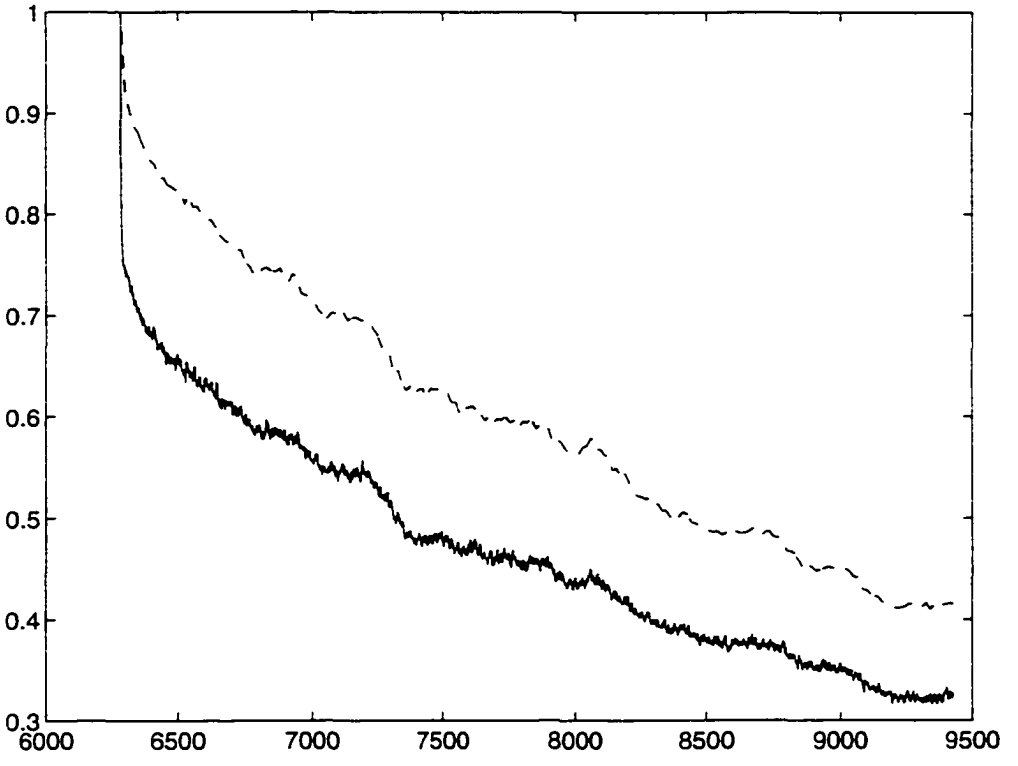

Figure 38 Multiplexed ABR Source - ACF's Of S(t) And B(t), S(t) On Solid, B(t) On Dash

3.5.2 Resource Allocation Using Effective Dynamic Bandwidth

The Effective Dynamic Bandwidth $B_{4,3}$ has been utilized in a simulated ATM switch to model the queuing effect at the switch. The Tail Distribution Function (TDF) has been obtained in simulations with a single VBR/MPEG source, multiplexed VBR/MPEG sources and ABR/LAN sources.

Figure-39 is a simplified version of the simulator we built to obtain the TDF's. When the VBR or ABR traffic stream arrives at our ATM simulator, S(t) is generated as a traffic cell number count in a unit time

interval and the Effective Dynamic Bandwidth $B(t)=B_{4,3}$ is calculated using (16) and the traffic volume S(t) is compared against the EDB. When S(t) is less than B(t), all of the traffic cells S(t) are passed to the downstream switching node. When S(t) is greater than B(t), the traffic cells are bounded by the bandwidth B(t), thus only B(t) traffic cells are passed to the downstream switching node, and S(t)-B(t) traffic cells remain in the local switch buffer when the switch buffer is unbounded. In the subsequent operations, the residual traffic cells are transmitted whenever $S(t) < B(t)$, and only a maximum of B(t)-S(t) is released from the local switch buffer at time t. The Tail Distribution Function is obtained as a probability density function of the traffic cell occupancy in the local switch buffer.

The simulations with bounded local buffers are conducted in our integrated system, which will be addressed in Section 4.5.

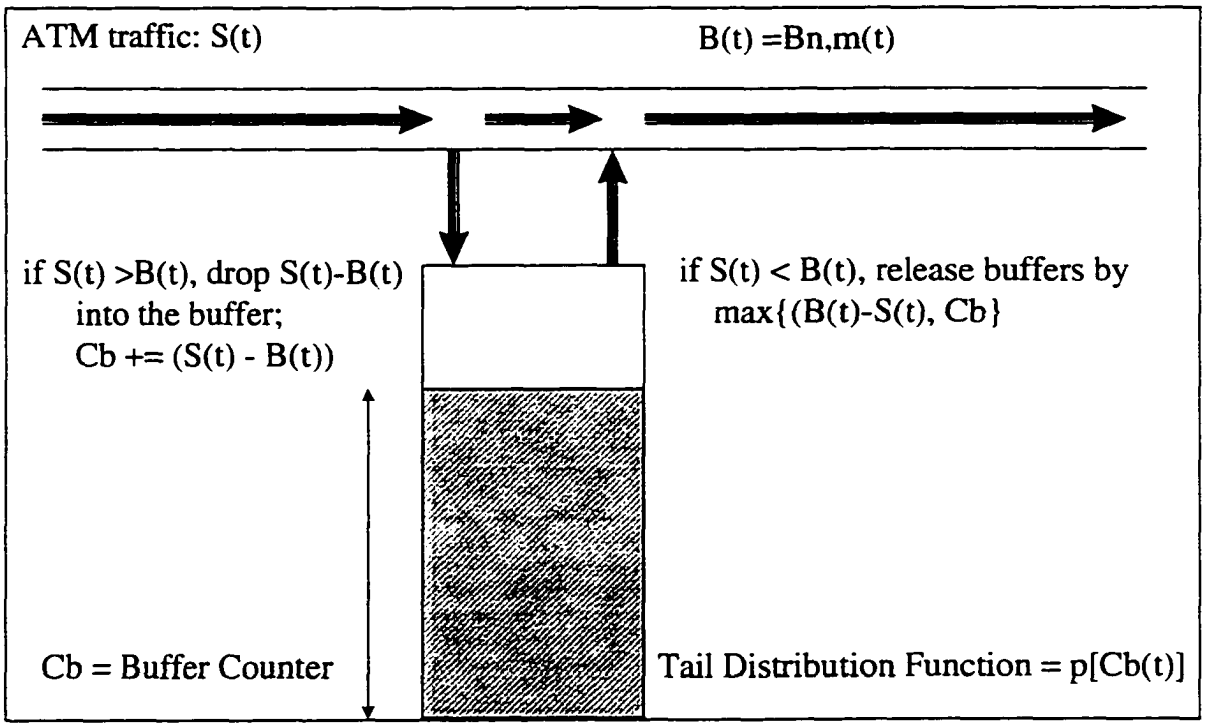

Figure 39 Bandwidth Enforcement with Effective Dynamic Bandwidth

The same VBR/MPEG single traffic trace in the Appendix A, which is also used for the statistical verification in the previous section, has been utilized in our simulations. And the following satisfactory results was generated as in Figure-40.

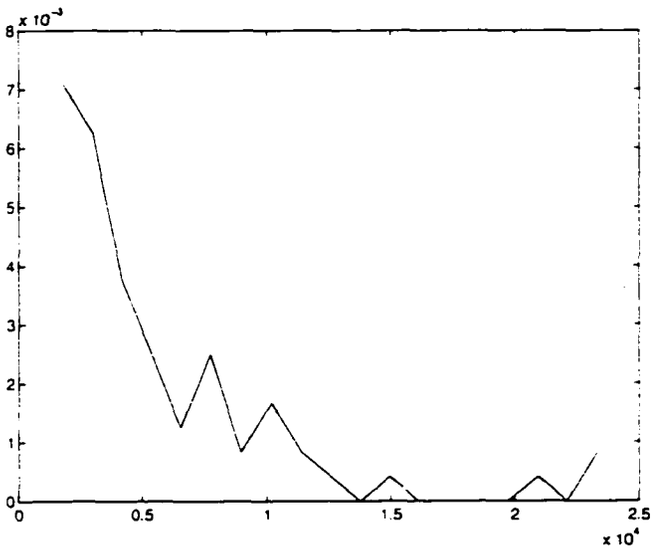

Figure 40 Tail Distribution Function During A Simulation With Single VBR Source

The ATM switch simulation also yields to the following Tail Distribution Function with multiplexed VBR/MPEG sources in Figure-41.

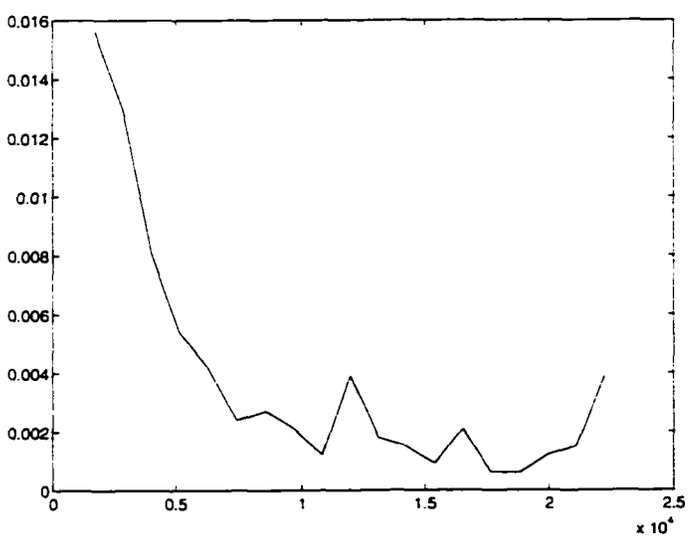

Figure 41 Tail Distribution Function During A Simulation With Multiplexed VBR Source

The simulation with ABR/LAN traffic generates the following Tail Distribution Function as in Figure-42.

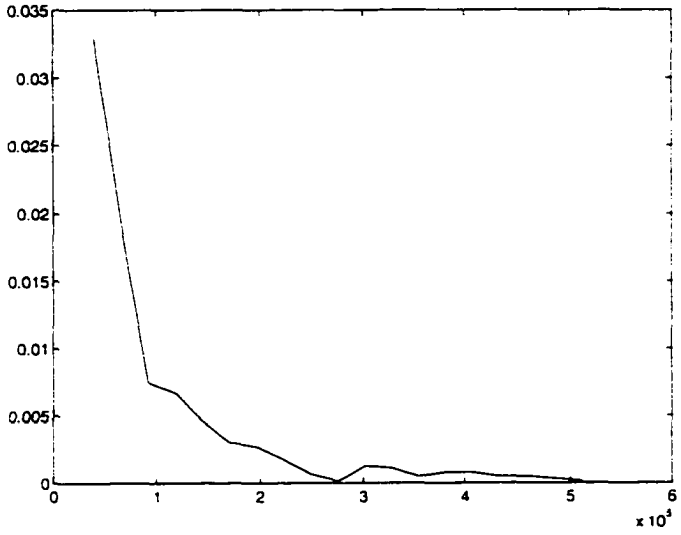

Figure 42 Tail Distribution Function During A Simulation With Multiplexed ABR Source

The discussions up to this point assume that the Effective Dynamic Bandwidth is provided to the ATM switch node, the next section will address the practical scenarios when the Effective Dynamic Bandwidth is not pre-generated.

4. Integrated Solution in ATM Switches

4.1 Predicting the Effective Dynamic Bandwidth

While the Effective Dynamic Bandwidth for ATM traffic obtained from the Wavelet decomposition reflecting the bandwidth consumption requirement within each bandwidth update interval, B(t) is only calculated after the traffic is arrived at each switching node. Naturally, the next question is "How can we obtain such information in advance, so that the bandwidth could be allocated accordingly?". This brings up the challenges in predicting the Effective Dynamic Bandwidth.

Through discussions in the previous chapters, Neural Network has the strength in learning and generalizing its knowledge base from the past patterns through its training processes.

4.2 Network Architecture and Training Techniques

Our research starts from identifying an effective network architecture and an efficient training process.

4.2.1 Network Architectures

Although a fully-connected neural network architecture could be implemented to support arbitrary systems in theory, the advantages of designing a specific neural network architecture include: the reduced training effort of the system, better convergence and stability of the system [28]-[29].

Examples of networks based on interpolation and extrapolation algorithms are *Radial Basis Networks* and *Local Learning* [30]-[32]. The *Associative Learning* and *Ising Models* are networks tightly coupled with the physical phenomenon such as that the state change of any particular element is determined by the state of its neighboring elements as illustrated in Figure-43(a).

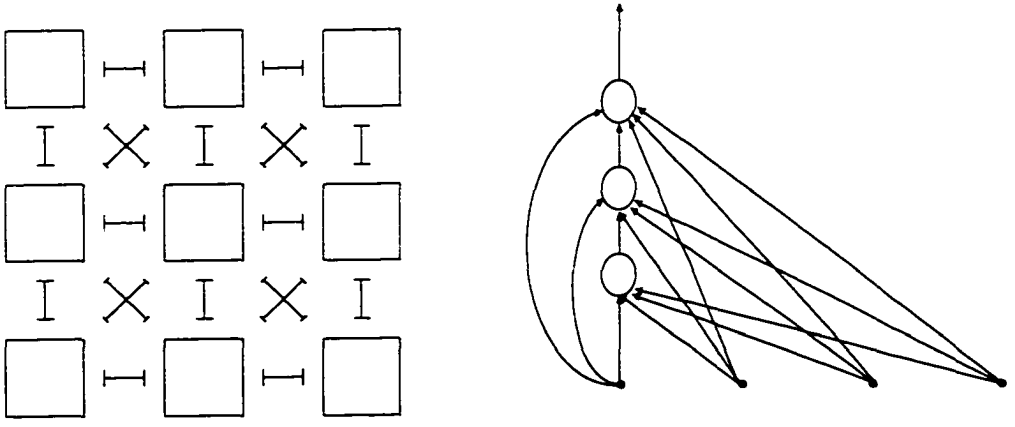

Figure 43 Associative Memory - (a) Ising Model; (b) Cascaded Network Architecture

Other variations in network architecture can be obtained by changing the associative feedback routes to *cascaded network* as shown in Figure-43(b). By adding time delay units to selected outputs and feeding them back as inputs, thus *recurrent network* is created, which is well suited for temporal processing. A recurrent network is illustrated in Figure-44. Example of recurrent networks includes *Hopfield* and *Elman networks.*

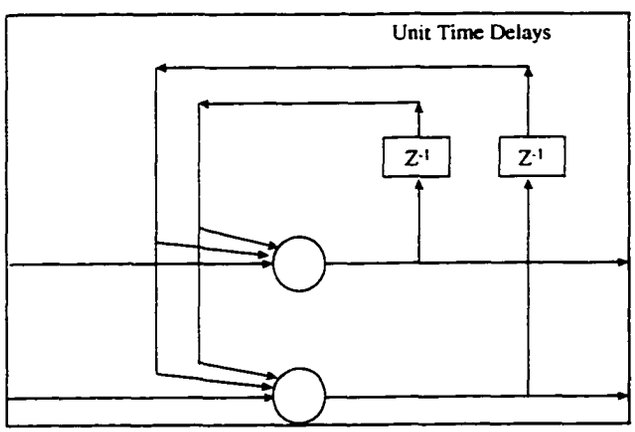

Figure 44 A Recurrent Network

4.2.2 Training Techniques

When a fully connected multi-layers neural network architecture is employed, there are two major operational stages: *Learning* and *Generalization.* The learning rules enforce the weights to converge to the optimal range during the training processes, and ultimately making the network capable of generalizing its knowledge to the generalization data

set. Some repetitive cycles may make the system learning as a real-time or semi-realtime process.

Backpropagation, a popular learning scheme, was introduced based on the gradient descent method, where the weight change is defined as:

$$\Delta w = lr \cdot Dx^T \tag{21}$$

and the threshold change is:

$$\Delta \theta = lr \cdot D \tag{22}$$

where x is the input vector, w is the weight matrix, θ is the threshold vector, and lr is the learning rate, D is the output error vector (or derivative) during the training.

However, the learning process through backpropagation is relatively slow and it may get stuck in a local minima.

In thermodynamic process, the cycle of successive heating (tempering) and slow cooling (annealing) could substantially improve the strength of metal objects. The researches in this area showed that such cycles increase the size of the crystalline alignment domains within the metal. And metal objects tend to break at boundaries of the alignment domains; they are made stronger by increasing the size of these domains. Further studies revealed that if the cooling is slow enough, the atoms in small domains will tend to get aligned with their neighbors in larger

domains, leading the effective growth of those larger domains. A substantial fraction of atomic bonds remain aligned with the direction of existing domains because the heating is controlled. When the metal is cooled, the largest of the old domains are restored first, and grow at the expense of their neighbors [26], [91], [92].

These results are also used in our research for the neural network learning rules. And the simulated annealing could be achieved by dynamically varying the learning parameters and this has been one of our proven methods of finding global minima.

Other improvements in learning rules are made such as the Levenberg-Marquardt Approximation [27]:

$$\Delta w = (J^T J + \mu \mathrm{I})^{-1} J^T e \tag{23}$$

where J is the Jacobian matrix of derivatives of error to weight, μ is a scalar, e is the error.

4.3 Recurrent Network with Adaptive Learning For Predicting EDB

A survey of some recent results on ATM traffic prediction is found in [34]. The potentials of traffic prediction without statistical model, have been demonstrated and reported in [36]-[52],[84]. Various neural network architectures and learning algorithms have been explored.

In [54], Li used a *Pi-Sigma Network (PSN)* which was initially proposed in [55], to predict the JPEG traffic. Instead of summing the inputs to the transfer functions in the output layer, this network multiplies them in its third layer as shown in Figure-45. The synchronous PSN dynamic bandwidth allocation is adapted by $C\hat{x}_{\max}(t)$, where $\hat{x}_{\max}(t)$ represents the maximum prediction of the $x(t)$ in the update duration L. The asynchronous PSN dynamic allocation is determined by $C\max\left\{\varphi, \hat{x}_{\max}(t)\right\}$, where $\varphi = E[x(t)] + \{Var[x(t)]\}^{1/2}$ is a pre-assigned nominal video bandwidth.

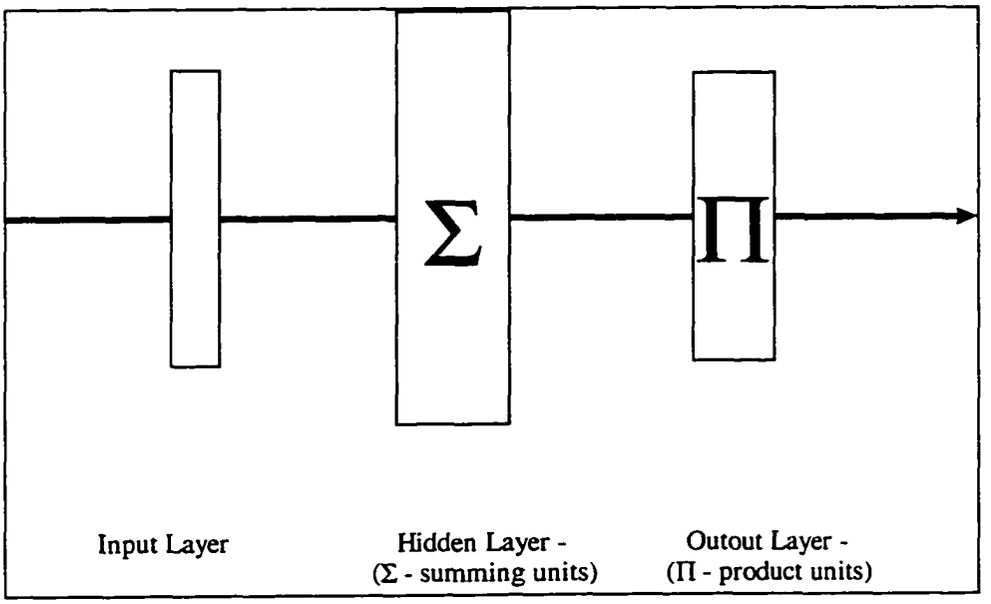

Figure 45 A Pi-Sigma Network

The reported transmission efficiencies have been achieved at 0.74 and 0.62 for synchronous PSN and asynchronous PSN, respectively. However, a

critical process in this proposal is the low-pass filter design for filtering out the high bandwidth components in the traffic volume trace. This in turn requires the power spectrum of the traffic volume be known for determining the cut-off frequency in the low-pass filter and impacting the parameters in the neural network such as the number of neurons and delay units. This proposal is practically not acceptable when dealing with unknown or new ATM traffic sources where the power spectrum of the traffic volume is not available.

In [56], Chang used a pipe-lined recurrent neural network with modularized network entities originally proposed by Haykin [57], to predict the MPEG traffic volume where the annealing learning algorithms are utilized. The architecture of this neural network is shown in Figure-46.

The underlining structure of this neural network is indeed a cascaded network. Rotating Figure-46 by 90 degrees counter-clock wise, it shares the same structure as the cascaded network as shown in Figure-43(b). The modifications in this pipe-lined recurrent network are through replacing the neurons by the network modules and adding time-delays in the feedback associations. The relative RMS prediction errors for I, P, and B frames at 1.0%, 2.3%, and 6.7% are achieved, respectively.

However, these results are only achievable after separating the traffic into three independent groups, i.e. group of I frames, group of P frames, and group of B frames. And such separation is practically not feasible in an ATM switch element where the traffic sources may not identified and the arriving traffic is not limited to the MPEG's.

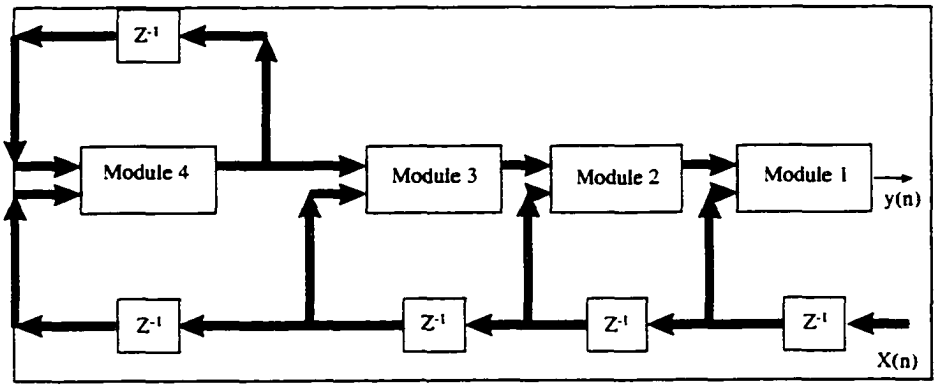

Figure 46 A Pipe-Lined Recurrent Network

Hence, limitations exist in the above methods. In general, an end-to-end realtime mechanism, especially with the knowledge in multiscale components, requires further investigation.

We have researched the neural network architectures for predicting the temporal behaviors and focused on the utilization of the recurrent architectures. A Recurrent Neural Network is a neural network with feedback to its input layer. By doing so, it extends its memory and its impact of the temporal patterns to the future recognition and prediction processes.

We have designed and simulated the illustrated recurrent architecture as in Figure-47, which consists of the followings:

- Time-delay units at the input layer to retain EDB truth inputs from the near past;
- Time-delay units in the recurrent associative learning to retain the EDB dynamics from the past;
- Transfer Functions at the first output layer are sigmoid functions,

$$\varphi(s)=\frac{1}{1+e^{-as}} \quad \text{where } a \text{ is a slope parameter.} \tag{24}$$

to achieve maximum variations;

- Transfer Functions at the second output layer are piecewise linear,

$$\varphi(s)=\begin{cases}\varphi_1 & s\le s_1,\\ a\cdot s+b & s_1\le s<s_2,\\ \varphi_2 & s\ge s_2\end{cases} \tag{25}$$

where a is an amplification parameter and b is a shift parameter to achieve maximum scalability;

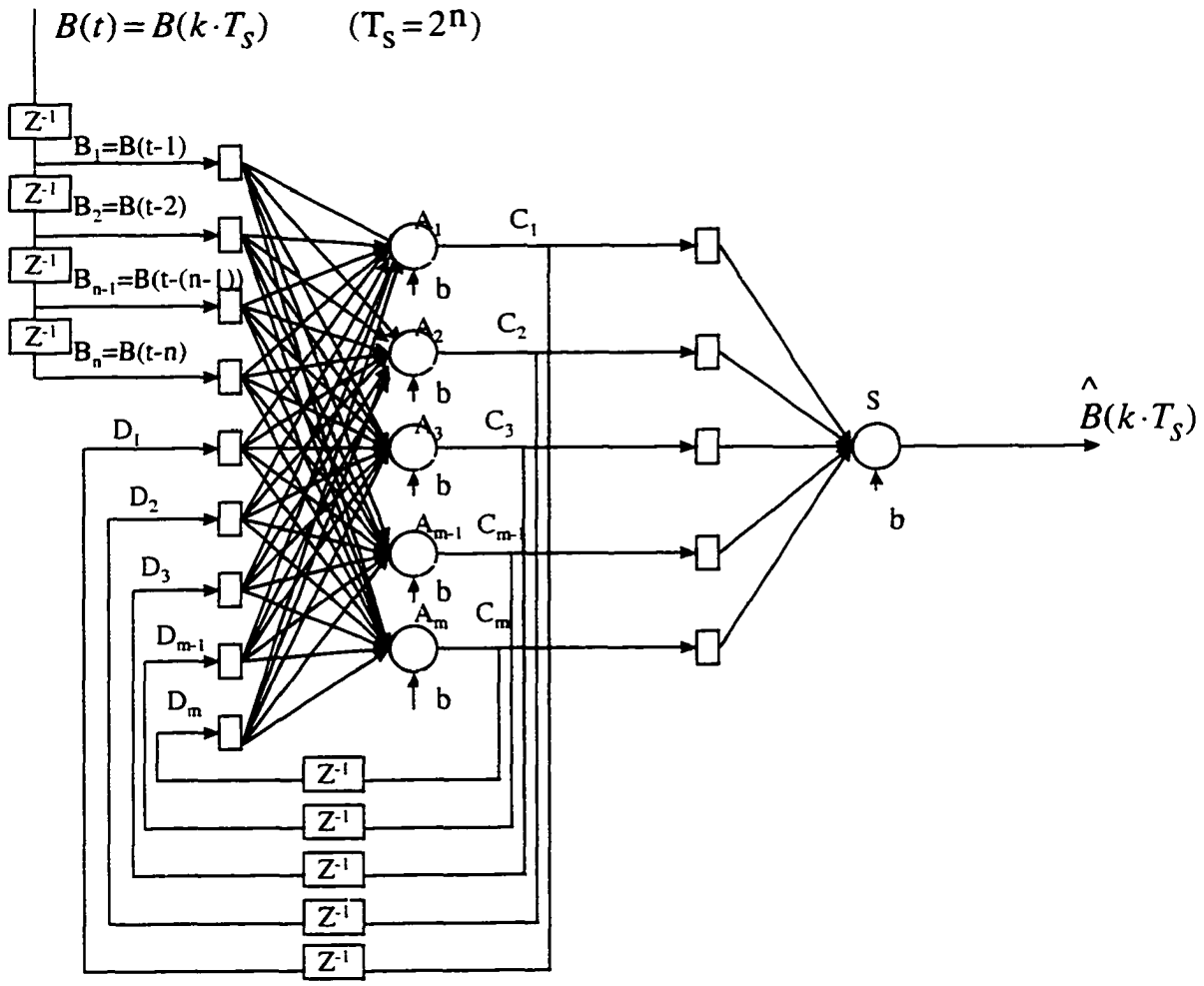

Figure 47 A Recurrent Neural Network Architecture

The inputs to the first layer of this network are the delayed traffic volumes:

$$\begin{bmatrix} B_1 \\ B2 \\ . \\ B_{n-1} \\ B_n \end{bmatrix} = \begin{bmatrix} B(t-1) \\ B(t-2) \\ . \\ B(t-(n-1)) \\ B(t-n) \end{bmatrix} \tag{26}$$

The sigmoid transfer functions operate on the combination of the input layer and delayed feedback from their own outputs.

$$\begin{bmatrix} C_1 \\ C_2 \\ . \\ C_{m-1} \\ C_m \end{bmatrix} = \begin{bmatrix} \frac{1}{1+e^{-a_1 A_1}} \\ \frac{1}{1+e^{-a_2 A_2}} \\ . \\ \frac{1}{1+e^{-a_{m-1} A_{m-1}}} \\ \frac{1}{1+e^{-a_m A_m}} \end{bmatrix} \tag{27}$$

where

$$\begin{bmatrix} A_1 \\ A_2 \\ \dots \\ A_{m-1} \\ A_m \end{bmatrix} = \begin{bmatrix} \lambda_{11} & \cdots & \lambda_{1n} & \lambda_{1(n+1)} & \cdots & \lambda_{1(n+m)} \\ \lambda_{21} & \cdots & \lambda_{2n} & \lambda_{2(n+1)} & \cdots & \lambda_{2(n+m)} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \lambda_{(m-1)1} & \cdots & \lambda_{(m-1)n} & \lambda_{(m-1)(n+1)} & \cdots & \lambda_{(m-1)(n+m)} \\ \lambda_{m1} & \cdots & \lambda_{mn} & \lambda_{m(n+1)} & \cdots & \lambda_{m(n+m)} \end{bmatrix} \begin{bmatrix} B_1 \\ \cdots \\ B_n \\ D_1 \\ \cdots \\ D_m \end{bmatrix} + \begin{bmatrix} b_1 \\ b_2 \\ \cdots \\ \cdots \\ b_m \end{bmatrix} \quad (28)$$

and

$$\begin{bmatrix} D_1 \\ D_2 \\ \cdots \\ D_{m-1} \\ D_m \end{bmatrix} = \begin{bmatrix} C_1(t-1) \\ C_2(t-1) \\ \cdots \\ C_{m-1}(t-1) \\ C_m(t-1) \end{bmatrix} \quad (29)$$

The third layer of the neural network is capable of scaling the traffic bandwidth estimations to the range of the line speed.

$$\hat{B}(t) = \begin{cases} \varphi_1 & s \le s_1, \\ a \cdot s + b & s_1 \le s < s_2, \\ \varphi_2 & s \ge s_2 \end{cases} \tag{30}$$

where

$$S = \begin{bmatrix} S_1 & S_2 & \cdot & S_{m-1} & S_m \end{bmatrix} \begin{bmatrix} C_1 \\ C_2 \\ \cdot \\ C_{m-1} \\ C_m \end{bmatrix} \tag{31}$$

Using the results from dynamics, a modified version of learning algorithm with adaptive learning rate and with momentum, is used in our research. The weight modification during a learning process is defined as:

$$\Delta w = mc \cdot \Delta w + (1 - mc) \cdot lr \cdot Dx^T \tag{32}$$

where Δw in the first term represents the momentum and mc is a momentum constant; Dx^T in the second term represents the gradient and the learning rate, lr, was originally proposed to be a constant, an adaptive learning rate deemed necessary in our research to simulate the annealing and cooling effect as mentioned earlier.

The learning rule enforced with the momentum reduced the chance to get stuck in the local minimum as illustrated in Figure-48 and Figure-49. With the accelerated hill-climbing method with fastest descent and conjugate gradient, as well as the concept of using momentum, and the artificial annealing and cooling, the learning speed has been accelerated in our simulation.

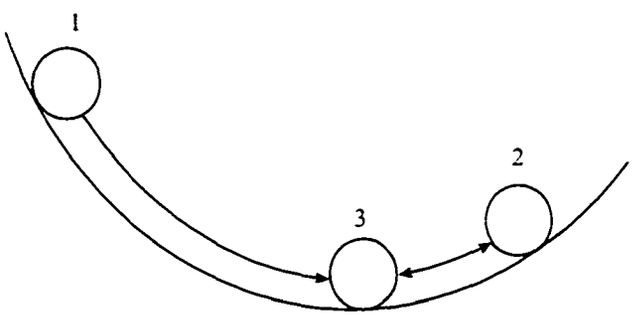

Figure 48 Learning Rule with Momentum Enforcement

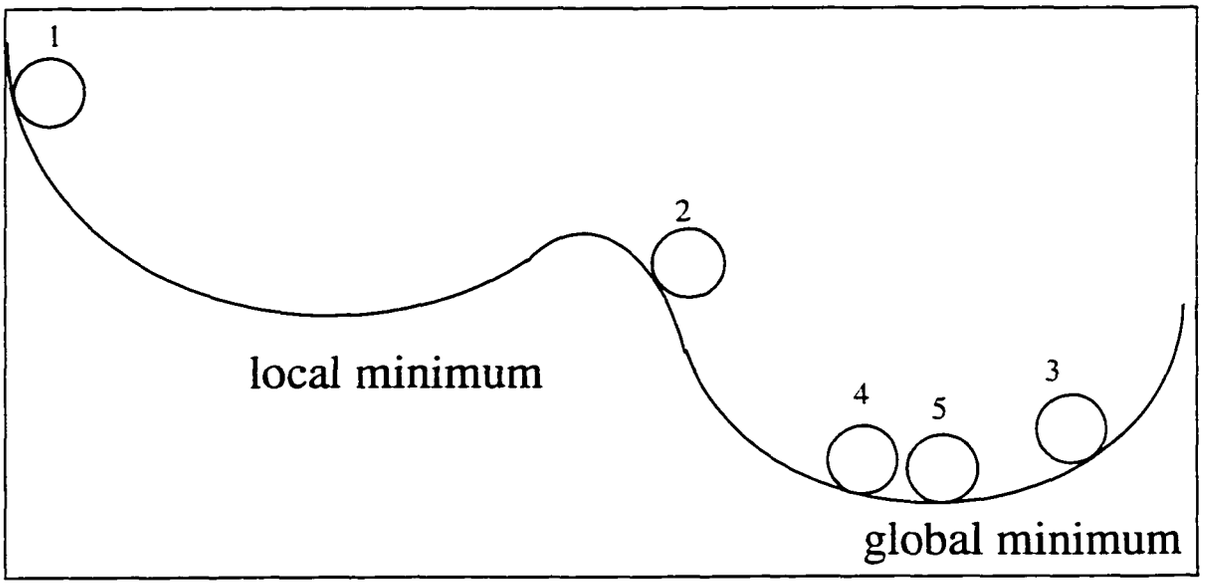

Figure 49 Finding Global Minimum - Momentum and Annealing Enforced Learning

During our simulation, the above neural network architecture and learning algorithms have been employed for one-step prediction of the effective dynamic bandwidth based on the Wavelet signal decomposition. Because the updating interval of the effective bandwidth is 2^n, it is in effect a predictor of the next 2^n GOP's for the bandwidth requirement.

4.4 EDB Prediction of Multimedia Traffic

The prediction of the effective bandwidth components has been applied to the Bellcore's Starwar movie data, as illustrated in Figure-50.

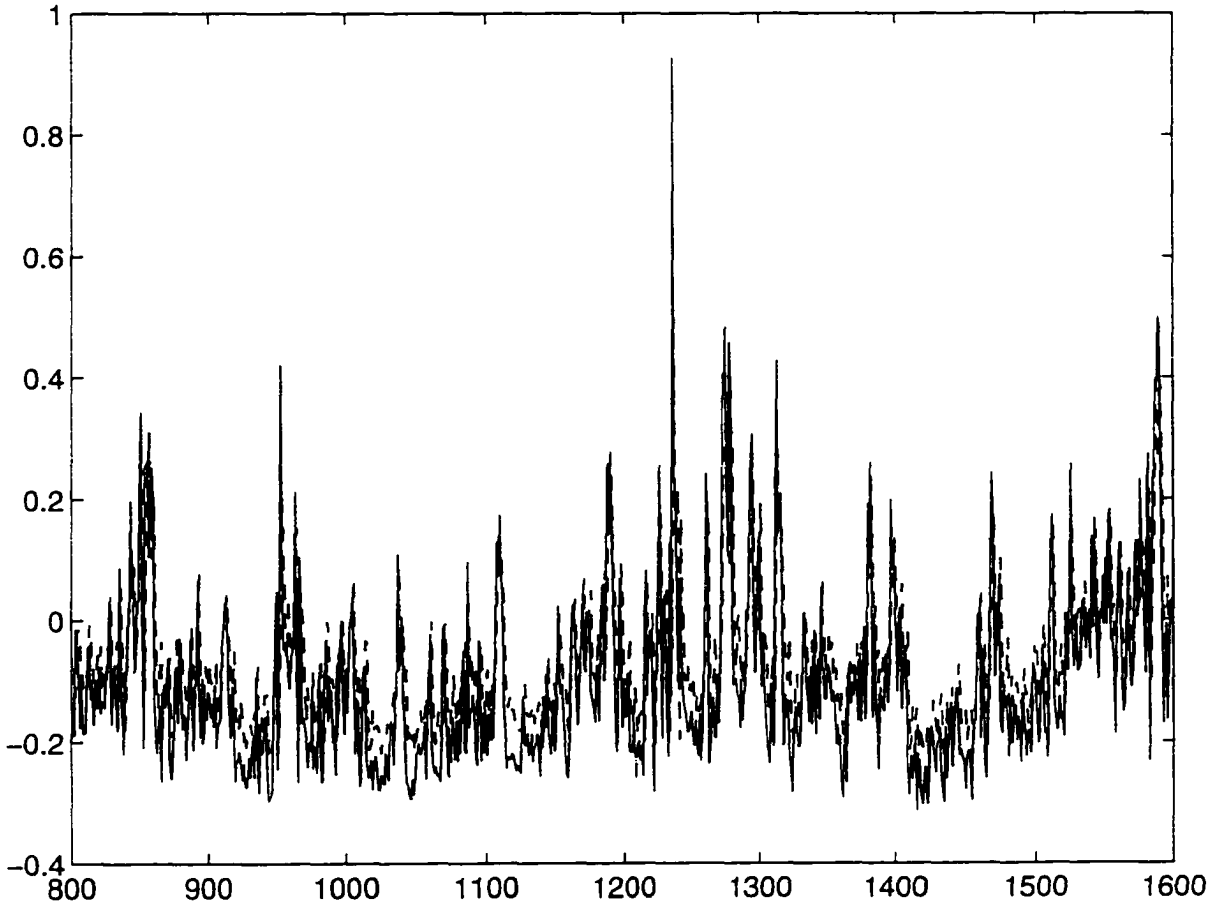

Figure 50 Prediction Results of Bellcore Starwar Movie

While the bandwidth is updated every 8 seconds, the neural network is re-trained progressively. The traffic prediction resulted from the above architecture demonstrated less than 1% error rate, which is significantly better than the results reported in [56], i.e. 6.7% for groups of B frames, 2.3% for groups of P frames, and 1.0% for groups of I frames.

The Prediction of the effective dynamic bandwidth has been applied to the multiplexed five movies using UPenn's data:

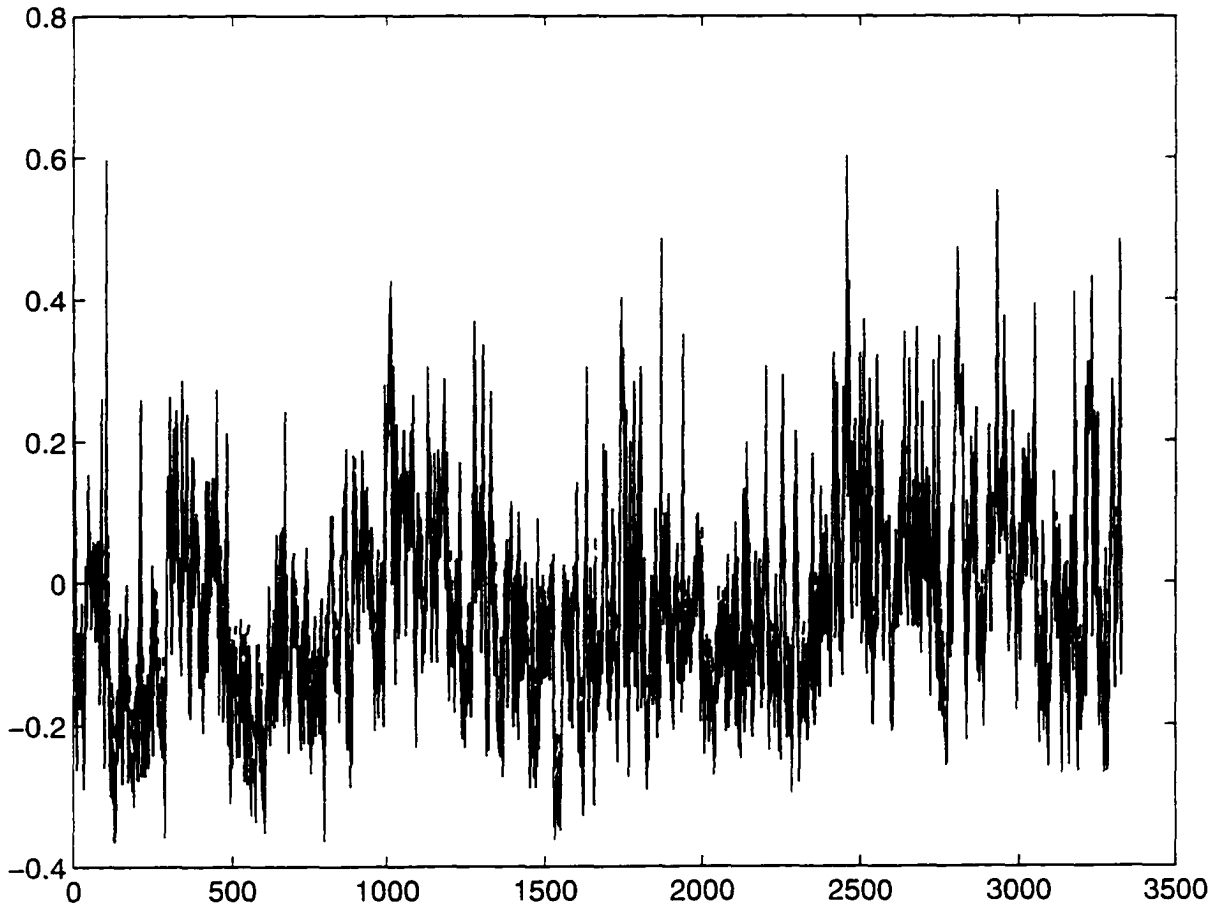

Figure 51 Prediction Results of UPenn's five movies-1

The scale of a segment of this prediction result is enlarged in Figure-52 where the predictions on the dashed line is shifted for display purpose.

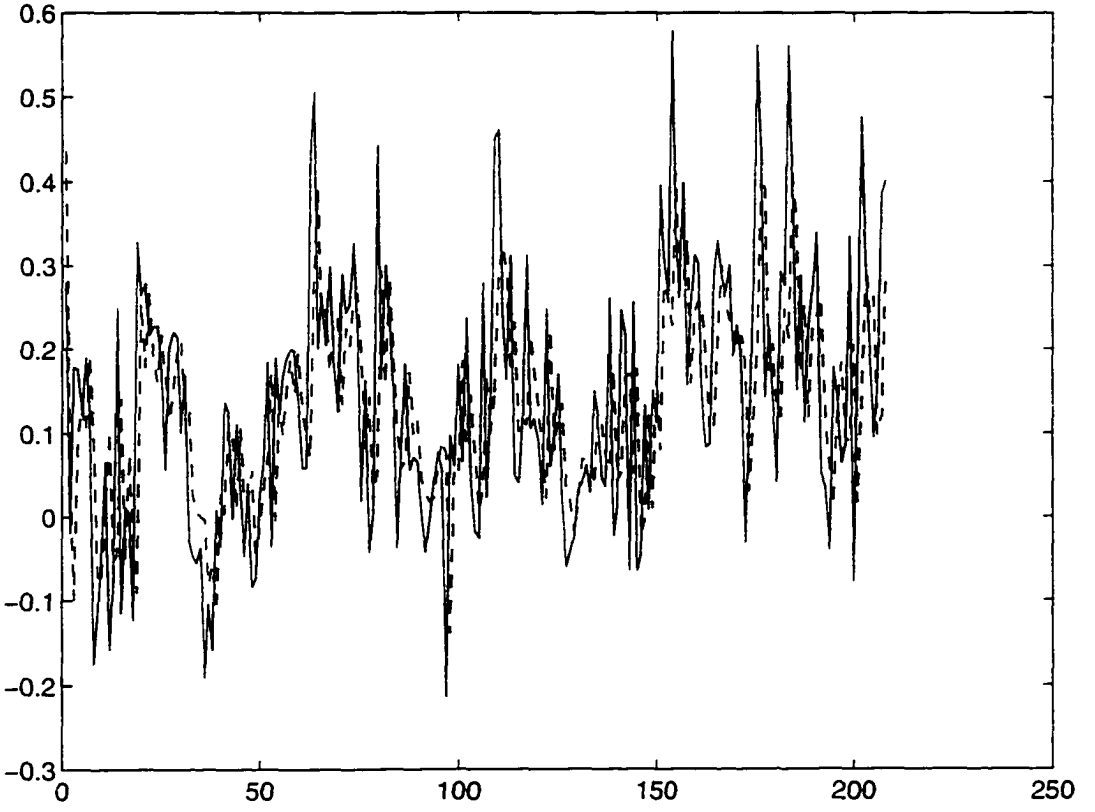

Figure 52 Prediction Results of UPenn's five movies-2

4.5 An Integrated System

With the understanding of the significance of the low frequency components in the long-term traffic characterization in multimedia communication for the bandwidth allocation, and the high frequency components in the short-term behavior for buffer allocation, and by exploring the research work in areas of ATM resource management as well as the technologies in Wavelet-based Multiscale Signal Decomposition and Neural-Network-based Adaptive Learning Mechanisms, we developed a

novel mechanism for the effective dynamic bandwidth analyses and further utilized a recurrent neural network with adaptive learning rate to predict the traffic bandwidth requirement.

The integrated solution consists of :

- Subsystem for extracting the Effective Dynamic Bandwidth;
- Recurrent (time-delayed) neural network for EDB Prediction;
- Bandwidth allocation unit;
- Quality of Service maintainer which calculates the Cell-Loss-Rate based on the actual traffic volume and the system generated bandwidth;
- Progressive NN training system which retrains the NN weights in real-time by taking the actual traffic volume, the calculated bandwidth, the QoS requirement, the actual QoS parameters, and the bandwidth truth value.

The integrated system for such implementation is presented in Figure-53.

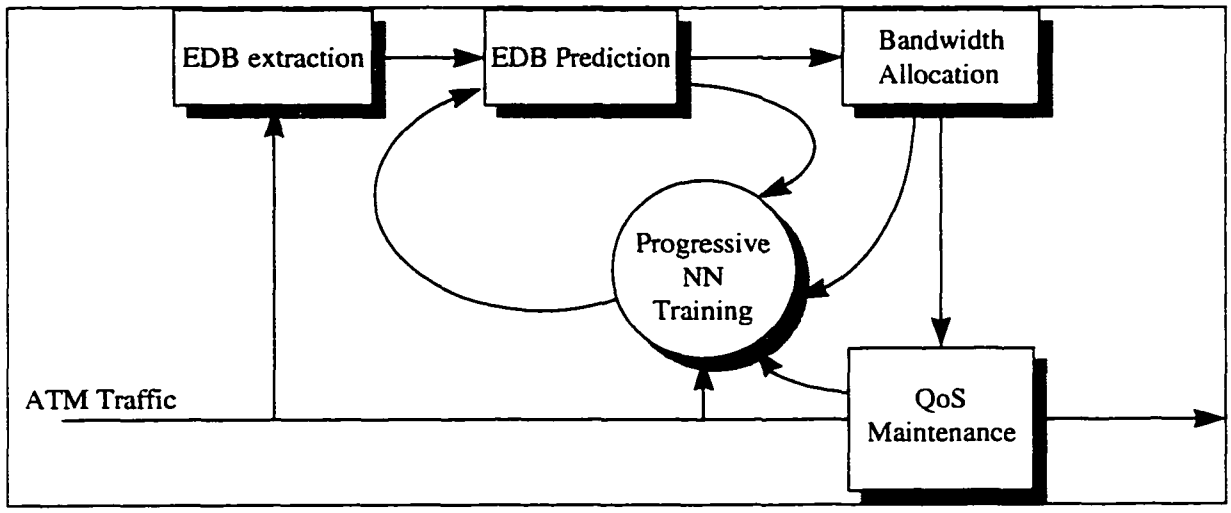

Figure 53 System Figure for the Integrated Solution

The system simulations have been conducted based on this integrated system solution which is evolved from the ATM simulator described in Section 3.5.2. A sample output of the traffic bandwidth allocation system is illustrated in Figure-54.

In our simulation system for the integrated solution, the bounded buffers are utilized at the switch node. In contrast to the unbounded buffers in Figure-39, the cell loss occurs when the local buffered traffic exceeds the buffer size, i.e. buffer overflow. The local buffer size is determined by the product of the maximum cell delay and the bandwidth to ensure the satisfaction of the service attributes provided in Table-2. The cell loss rate (CLR) can be calculated by counting the traffic cell overflow following the definitions described in Section 2.2.

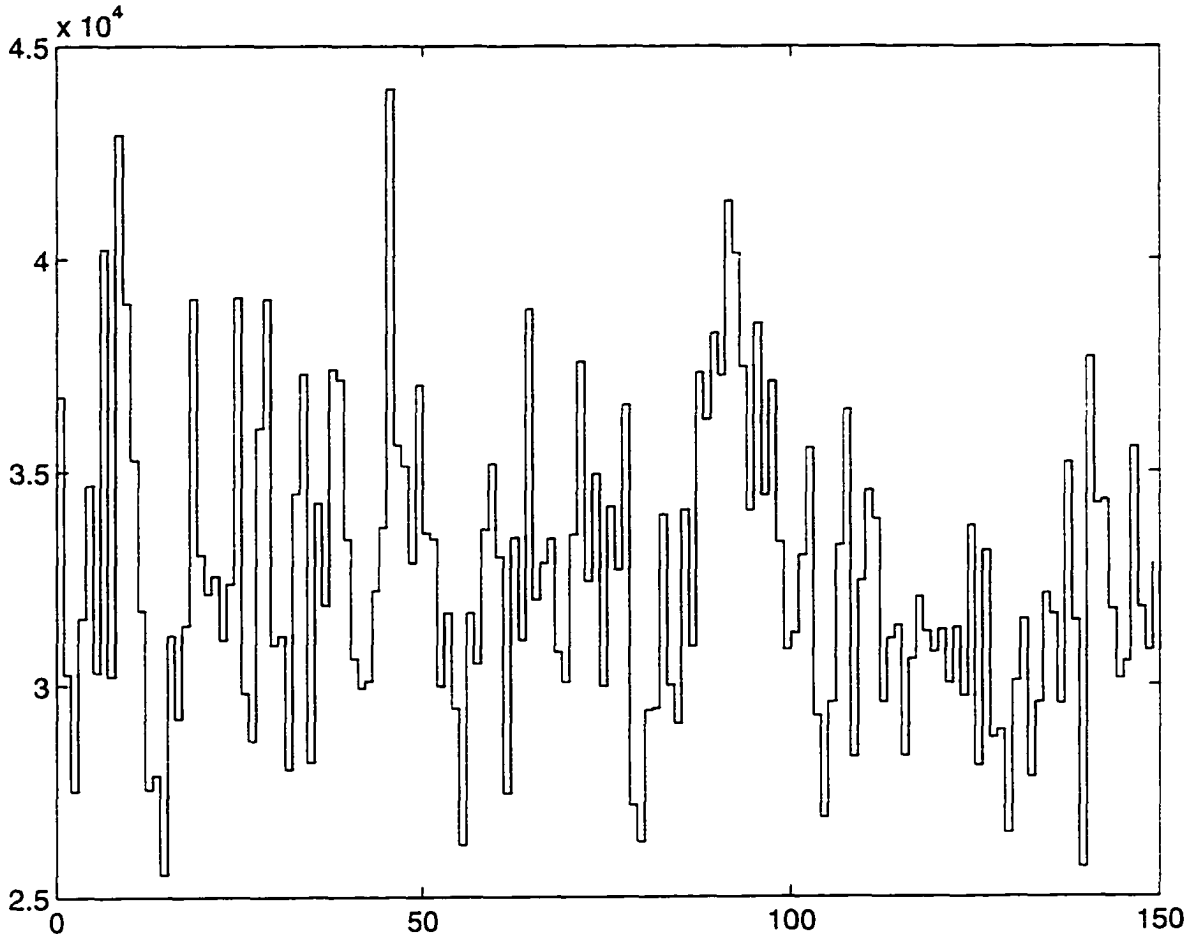

Figure 54 A Sample Output of the Traffic Bandwidth Allocation Subsystem

The Utilization/Delay relationships from our simulations are reported in Figures 55, 56, and 57, where the bandwidth is updated at every 4, 8, 16, and 32 GOP's and maintaining CLR = 10^{-7}.

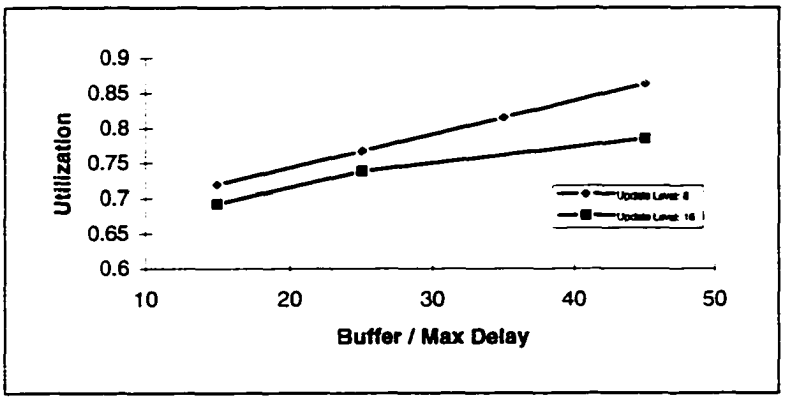

Figure 55 Utilization Rate with Different Bandwidth Update Level for Single Source While CLR = 10^{-7}

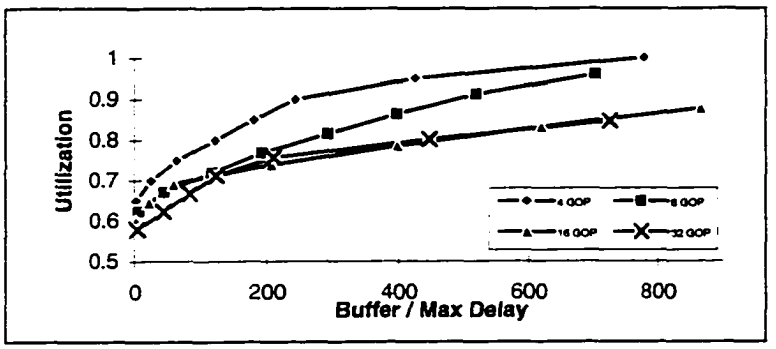

Figure 56 Utilization Rate with Different Bandwidth Update Level for 4 Sources While CLR = 10^{-7}

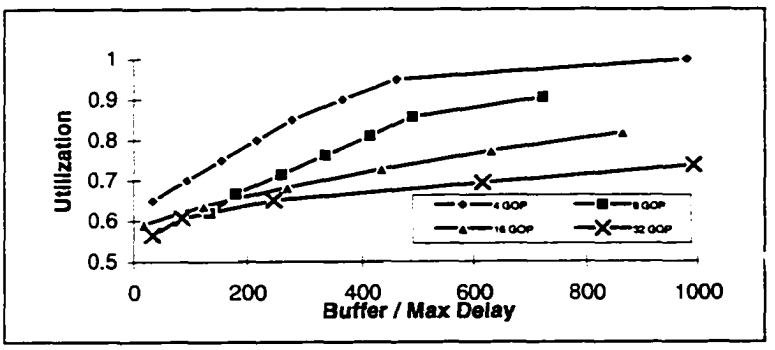

Figure 57 Utilization Rate with Different Bandwidth Update Level for 6 Sources While CLR = 10^{-7}

The utilization rate represents the transmission efficiency $\frac{S(t)}{\hat{B}(t)}$.

Based on (16), *B(t)* contains the details components, i.e., traffic fluctuations. The utilization rate tends to be high when experiencing less traffic fluctuation, and low with high traffic fluctuation. The utilization rate also goes higher when B(t) is updated more frequently. However, utilization rates achieved by our integrated system consistently outperform the results of 0.62 to 0.74 as reported in [54].

The next four figures represent the general requirement for the maximum delay (buffer size) in order to meet the Cell Loss Rate requirements with multiple achievable utilization rates.

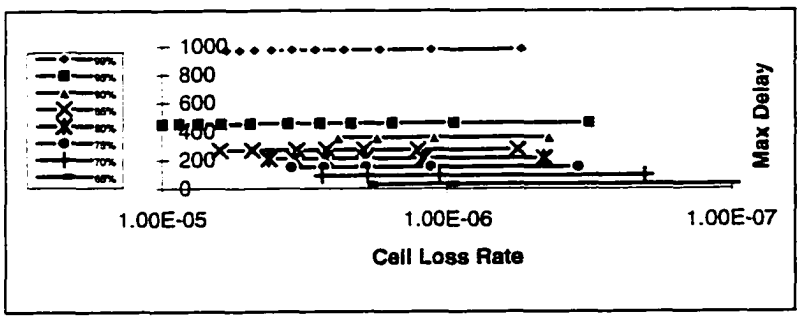

Figure 58 Cell Loss Rate While Update Level at 2

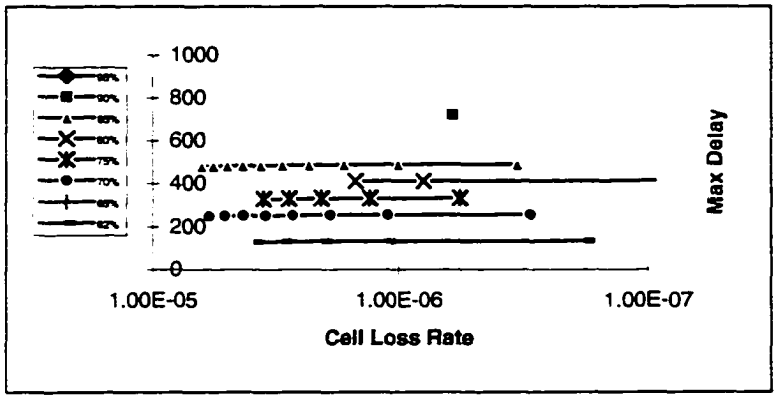

Figure 59 Cell Loss Rate While Update Level at 3

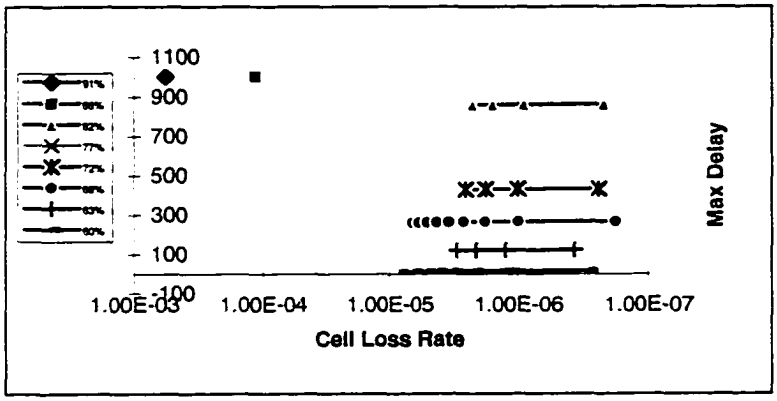

Figure 60 Cell Loss Rate While Update Level at 4

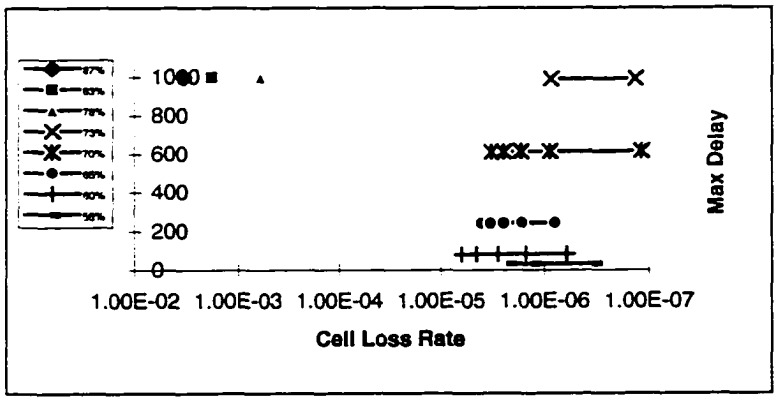

Figure 61 Cell Loss Rate While Update Level at 5

Obviously, the more frequently the bandwidth is updated, the higher the bandwidth utilization rate could be achieved from the above figures. However, our system could achieve a much higher bandwidth utilization rate than the ones reported in [54] and [56], while CLR is at 10^{-6} to 10^{-7} level.

The major contributions of our work are summarized as:

- The prerequisite knowledge of the traffic bandwidth dynamics is not necessary in our system. This is in contrast to the proposal in [54], where the pre-calculated spectrum is a prerequisite for the filter design for extracting the bandwidth components before its predictions take place.
- The utilization rates at 73% ~ 87% as illustrated in Figures 55, 56 and 57, are achieved in the integrated system we developed, which outperform the previously reported results in [54], ranging from 62% to 74%. Our system also demonstrated better than 90% utilization rate when multiple traffic sources are multiplexed.
- The ATM traffic trace does not need to be separable into I, P, B groups in our proposal. However, the proposal in [56] is only applicable to pre-separated frame groups of MPEG traffic.

- While the bandwidth is updated every 8 seconds, our neural network is re-trained progressively. The traffic prediction resulted from our system demonstrated less than 1% error rate, which is significantly better than the results reported in [56], i.e. 6.7% for groups of B frames, 2.3% for groups of P frames, and 1.0% for groups of I frames.
- Furthermore, it has been demonstrated that our algorithms are independent from the traffic type, such as VBR or ABR and the traffic fluctuations, because the multiscale analysis is capable of grouping them into the appropriate categories automatically based on the update levels at ATM switch nodes.

We also acknowledge a recent research report on Wavelet analysis of video traffic[89], which signals the public research interests in area of Wavelet for traffic analysis. That work provided a unified approach to model both the long-range and short-range dependence in video traffic simultaneously, and it attempted to model the wavelet coefficients by independent or Markov models. However, it is limited to the video traffic analysis and it did not address the practical solution in the ATM traffic management. Our research work directly contributes to this area. Our

research not only provides the traffic bandwidth analyses utilizing the multiscale techniques, but also demonstrates an integrated novel solution for the ATM bandwidth resource management with our research and development in Wavelet, Prediction and Learning.

5. Conclusions and Discussions

There is a great desire and need to support various applications with varied Quality-of-Service (QOS) requirements on an integrated network infrastructure in the future. Asynchronous Transfer Mode (ATM) based networks have attracted much attention as they promise support for a wide variety of applications and QOS requirement, including web browsing, bulk data transfer, telephony, and video conferencing. This promise of a seamless and efficient transport infrastructure, however, depends on effective traffic management techniques, which in turn motivates the work being carried out in our research.

Reviews are provided in various aspects of traffic management in ATM, which includes the ATM service classes, policing, connection admission control, scheduling, resource management, flow control and traffic shaping among others, and a brief outline of the framework, major issues and select solutions are highlighted. We have found that for non-stationary signals, especially for the ATM traffic volume, the assumptions

for statistical models may not be satisfied, the results based on those models will not be valid. Hence, the traditional traffic studies based on the statistical model have their limitations. However, the analyses of the ATM traffic at multiple time scales become more meaningful with the aid of Wavelet signal decomposition,

With the understanding of the significance of the low frequency components in the long-term traffic characterization in multimedia communication for the bandwidth allocation, and the high frequency components in the short-term behavior for buffer allocation, and by exploring the research work in areas of ATM resource management as well as the technologies in Wavelet-based Multiscale Signal Decomposition and Neural-Network-based Adaptive Learning Mechanisms, we have developed a novel mechanism for the effective dynamic bandwidth analyses and further utilized a recurrent neural network with adaptive learning rate to predict the traffic bandwidth requirement.

Our research work not only identified an effective representation of the dynamic bandwidth, but also further developed an integrated solution of the dynamic traffic bandwidth prediction contributing to real-time end to end ATM QoS maintenance, which is robust not only to the existing services but also adaptive to the future new service establishments.

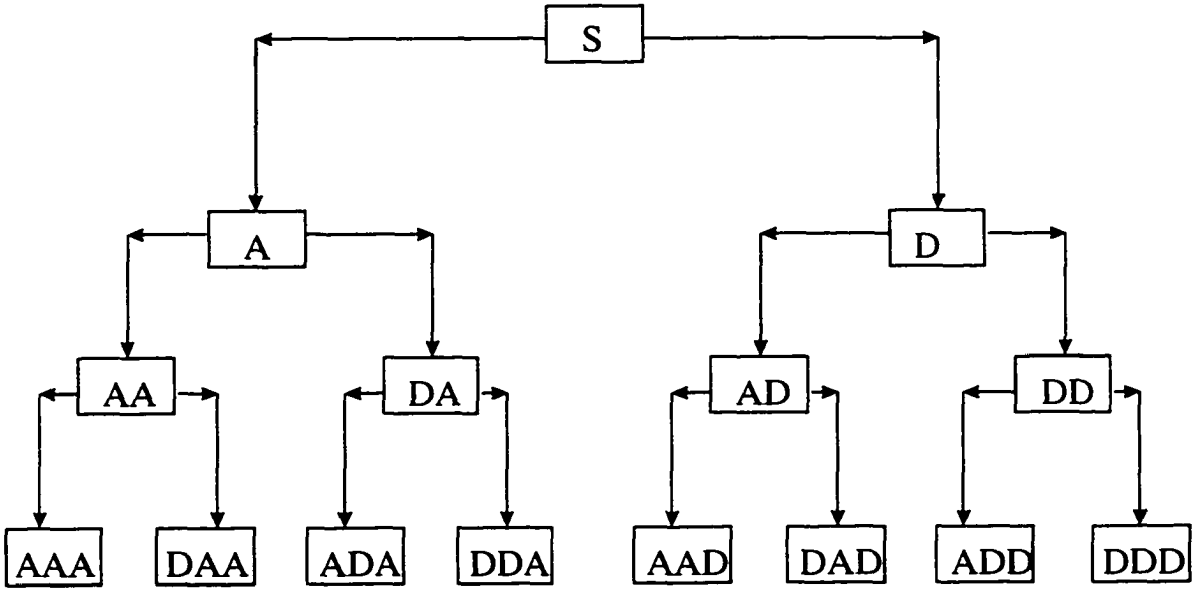

Figure 62 Decomposition with Wavelet Packets

Future research work remains in the areas such as the studies on the relationships between the wavelet decomposition parameters and the multiple cell-loss-rate requirements. Other subjects such as the correlation between the coefficients in Wavelet Packets and the buffer resource allocation as well as the studies on the integration involving multiple ATM nodes and the realtime implementation [86] will be of interest.

6. Appendix A: Test Dataset

Two groups of the MPEG data have been utilized to evaluate the algorithms effectiveness on VBR traffic. They are: the Bellcore's MPEG data for Single VBR/MPEG source, and the Upenn's MPEG data for Multiplexed VBR/MPEG sources. In addition, Bellcore's LAN traffic data is used for verifying the ABR traffic.

6.1 Single VBR Source

The Bellcore's dataset, Variable Bit Rate Video Bandwidth Trace Using MPEG Code, by Mark W. Garrett [71] and Antonio Fernandez, is used to examine the effectiveness of the Wavelet decomposition schemes and traffic predictions.

This dataset is available over the Internet via anonymous ftp in "/pub/vbr.video.trace/MPEG.data" from the host of "ftp.bellcore.com". This data set represents the bandwidth output of a variable bit rate (VBR) video coder which conforms to the MPEG-I standard. The source material

contains quite a diverse mixture of material ranging from low complexity/motion scenes to those with very high action. The data file consists of 174,138 integers, representing the number of bits per video frame. Only the original film frames are coded (i.e., 24 per second). The movie length is approximately 2 hours. The original video was captured as 408 lines by 508 pels, and then interpolated and filtered to standard CIF frame size, which is 240X352 (Luminance - Y), 120X176 (Crominance - U & V). The sequence of MPEG I, P and B frames used is IBBPBBPBBPBB IBB..., so there are 12 frames in a Group of Pictures (GOP) [72].

6.2 Multiple VBR Sources

The frame size traces of MPEG-I encoded video of five movies at University of Pensyvania, are used to verify the Effective Dynamic Bandwidth with Multiplexed VBR/MPEG sources.

These Upenn's datasets are publicly available on Internet from "http://www.seas.upenn.edu/~reisslei/video.html". They were originally obtained via anonymous ftp from "ftp://ftp-info3.informatik.uni-wuerzburg.de/pub/MPEG/", at University of Wuerzburg, Germany[82].

The encoder parameters are

- Encoder Input: 384 x 288 pel

- Color Format: YUV (4:1:1, resolution of 8 bits)
- Quantization Values: I = 10, P = 14, B = 18
- Pattern: IBBPBBPBBPBB
- GOP Size: 12
- Motion Vector Search: 'Logarithmic' / 'Simple'
- Frame Rate: 25 frames/second
- Number of Frames per Sequence: 40,000

The available traces include:

1. "The Silence of the Lambs"
2. "James Bond: Goldfinger"
3. "Mr. Bean"
4. Soccer
5. "Terminator II"

Each movie is treated as a separate traffic source. They are multiplexed together. The multiplexed traffic volume of every 12 frames, a GOP interval, is used in this simulation.

6.3 ABR Source

The Bellcore's LAN traffic data by Will Leland, (wel@bellcore.com), is used to verify the Effective Dynamic Bandwidth with ABR/LAN source.

This dataset is available on internet from http://town.hall.org /Archives/pub/ITA/. The files are ASCII-format tracing data, consisting of one 20-byte line per Ethernet packet arrival. Each line contains a floating-point time stamp and an integer for the Ethernet data length in bytes. The effective resolution is roughly 10 microseconds. The length field does not include the Ethernet preamble, header, or CRC; however, the Ethernet protocol forces all packets to have at least the minimum size of 64 bytes and at most the maximum size of 1518 bytes. 99.5% of the encapsulated packets carried by the Ethernet PDUs were IP. The records include all complete packets, but do not include any fragments or collisions.

The LAN traffic data file being used in our tests is "pAug.TL", which consists of the first 1 million arrivals (about 3142.82 seconds) of the day-long trace started at 11:25 a.m., 29 August 1989.

7. References

[1] ITU-T Recommendation I.121, Integrated Services Digital Network (ISDN) - General Structure and Service Capabilities - Broadband Aspects of ISDN, 1991

[2] ITU-T Recommendation I.371, Integrated Services Digital Network (ISDN) - Overall Network Aspects and Functions - Traffic Control and Congestion Control in B-ISDN, March 1993

[3] ITU-T Recommendation I.356, Integrated Services Digital Network (ISDN) - Overall Network Aspects and Functions - B-ISDN ATM Layer Cell Transfer Performance, Nov. 1993, and revised draft, May 1996

[4] M.Prycker, "Asynchronous Transfer Mode - solution to broadband ISDN", 3rd Edition, Prentice Hall, p36, p51, 1995

[5] D.LeGall, "MPEG: A Video Compression Standard for Multimedia Applications", Commun. of the ACM, pp47-58, April 1991

[6] P.Pancha, and M.E.Zarki, "MPEG Coding for Variable Bit Rate Video Transmission", IEEE Communications Magazine, pp54-66, May 1994

[7] M.Nomura, T.Fujii, and N.Ohta, "Basic characteristics of variable rate video coding in ATM environment", IEEE JSAC, pp752-760, June 1989

[8] R.Grunenfelder, J.P.Cosmas, S.Manthorpe, and A.Odinma-Okkafor, "Characterization of Video Codecs as Autoregressive Moving Average Processes and Related Queuing System Performance", IEEE JSAC, pp284-293, April, 1991

[9] R.M.Rodriguez-Dagnino, M.R.K.Khansari, and A.Leon-Garicia, "Prediction of Bit Rate Sequences of Encoded Video Signals", IEEE JSAC, pp305-314, April 1991

[10] J.Y.Hui, and E.Karasan, "A Thermodynamic Theory of Broadband Networks with Application to Dynamic Routing", IEEE JSAC, pp991-1003, August 1995

[11] N.G.Duffield, J.T.Lewis, N.O'Connell, R.Russell, and F.Toomey, "Entropy of ATM Traffic Streams: A Tool for Estimating QoS Parameters", IEEE JSAC, pp981-990, August 1995

[12] H.J.Fowley, and W.E.Leland, "Local Area Network Traffic Characteristics, with Implications for Broadband Network Congestion Management", IEEE JSAC, pp1139-1149, Sept. 1991

[13] D.N.C.Tse, R.G.Gallager, and J.N.Tsitsiklis, "Statistical Multiplexing of Multiple Time-Scale Markov Streams", IEEE JSAC, pp1028-1038, August 1995

[14] V.G.Kulkarni, L.Gun, and P.F.Chimento, "Effective Bandwidth Vectors for Multiclass Traffic Multiplexed in a Partitioned Buffer", IEEE JSAC, pp1039-1047, Aug, 1995

[15] R.Guerin, H,Ahmadi, and M.Naghshineh, "Equivalent Capacity and Its Application to Bandwidth Allocation in High-Speed Networks", IEEE JSAC, pp968-981, Sept, 1991

[16] G.deVeciana, G.Kesidis, and J.Walrand, "Resource Management in Wide-Area ATM Networks Using Effective Bandwidths", IEEE JSAC, pp1081-1090, Sept. 1995

[17] G.L.Choudhury, and D.M.Lucantoni, "Squeezing the Most out of ATM", IEEE Trans. on Communications, pp203-217, Feb. 1996

[18] R.J.Gibbens, F.P.Kelly, and P.B.Key, "A Decision-Theoretic Approach to Call Admission Control in ATM Networks", IEEE JSAC, pp1101-1114, August 1995

[19] A.Elwalid, D.Mitra, and R.H.Wentworth, "A New Approach for Allocating Buffers and Bandwidth to Heterogeneous Regulated Traffic in an ATM Node", IEEE JSAC, pp1115-1127, August 1995

[20] E.P.Rathgeb, "Modeling and Performance Comparison of Policing Mechanisms for ATM Networks", IEEE JSAC, pp325-334, April 1991

[21] M.Butto, E.Cavallero, and A.Tonietti, "Effectiveness of the 'Leaky Bucket' Policing Mechanism in ATM Networks", IEEE JSAC, pp335-342, April 1991

[22] H.J.Kushner, "Analysis of controlled multiplexing systems via numerical stochastic control methods", IEEE JSAC, pp1207-1218, Sept. 1995

[23] R.Ting, et al, "Studies on ATM Quality of Service Improvement", Proceedings of 1996 International Conference on Information Infrastructure (ICII'96), pp492-495, April, 1996

[24] F.Kishino, K.Manabe, Y.Hayashi, and H.Yasuda, "Variable Bit-Rate Coding of Video Signals for ATM Networks", IEEE JSAC, pp801-806, June 1989

[25] A.Hiramatsu, "Integration of ATM Call Admission Control and Link Capacity Control of Distributed Neural Networks", IEEE JSAC, pp1131-1138, Sept. 1991

[26] Simon Haykin, Neural Networks - A Comprehensive Foundation, 1994, Macmillan Publishing Company

[27] Howard Demuth, Neural Network Toolbox, 1995, The MathWorks, Inc.

[28] R.Ting, et al, "Neural-Net Applications in Character Recognition", book chapter, in press

[29] R.Ting, et al, "Address Block Location Using a Neural Network Hardware", book chapter, in press

[30] L.Bottou, and V.N.Vapnik, "Local Learning Algorithms", Neural Computation, pp888-900, 1992

[31] V.N.Vapnik, and L.Bottou, "Local Algorithms for Pattern Recognition and Dependencies Estimation", Neural Computation, pp893-909, 1993

[32] Y.LeCun, etc. "Learning Algorithms for Classification: A Comparison on Handwritten Digit Recognition", Neural Networks: The Statistical Mechanics Perspective, pp261-276, World Scientific, 1995

[33] D.Wrege, E.Knightly, H.Zhang, and J.Liebeherr, "Deterministic Delay Bounds for VBR Video in Packet-Switching Networks: Fundamental Limits and Practical Trade-Offs", IEEE/ACM Trans. on Networking, pp.352-362, June 1996

[34] I.W.Habib, R.J.T.Morris, H.Saito, and B.Pehrson, "Computational and Artificial Intelligence in High Speed Networks", IEEE JSAC, pp133-135, Feb. 1997

[35] D. Heyman, and T.V.Lakshman, "What are the implications of Long-Range dependence for VBR-Video traffic Engineering?", IEEE/ACM Trans. on Networking, pp.301-317, June 1996

[36] A.A. Tarraf, I.W. Habib, and T.N. Saadawi, "A Novel Neural Network Traffic Enforcement Mechanism For ATM Networks", IEEE JSAC, pp1089-1096, Aug. 1994

[37] S.Chong, and S.Li, "(σ,ρ)-Characterization Based Connection Control for Guaranteed Services in High Speed Networks", IEEE Infocom'95, pp.835-844, April 1995

[38] T.Poggio, and F.Girosi, "Networks for Approximation and Learning", Proc. of the IEEE, pp1481-1497, Sept. 1990

[39] E.S.Yu, and C.Y.R.Chen, "Traffic Prediction Using Neural Networks", IEEE GlobeCom'93, pp991-995, 1993

[40] T.Sikora, "The MPEG-4 Video Standard Verification Model", IEEE Trans. Circuits and Systems for Video Technology, pp.19-31, Feb. 1997

[41] A.Hiramatsu, "Training Techniques for Neural Network Applications in ATM", IEEE Communications Magazine, pp58-67, October 1995

[42] E.Nordstrom, J.Carlstrom, O.Gallmo, and L.Asplund, "Neural Netwroks for Adaptive Traffic Control in ATM Networks", IEEE Communications Magazine, pp43-49, October 1995

[43] J.E.Neves, M.J.Leitao, L.B.Almeida, "Neural Networks in B-ISDN Flow Control: ATM Traffic Prediction or Network Modeling?", IEEE Communications Magazine, pp50-57, October 1995

[44] Y.Park, and G.Lee, "Applications of Neural Networks in High-Speed Communication Networks", IEEE Communications Magazine, pp68-75, October 1995

[45] A.Farago, J.Biro, T.Henk, and M.Boda, "Analog Neural Optimization for ATM Resource Management", IEEE JSAC, pp156-164, Feb. 1997

[46] P.K.Campbell, A.Christiansen, M.Dale, H.L.Ferra, A.Kowalczyk, and J.Szymasnski, "Experiments with Simple Neural Networks for Real-Time Control", IEEE JSAC, pp165-178, Feb. 1997

[47] K.Uehara, and K.Hirota, "Fuzzy Connection Admission Control for ATM Networks Based on Possibility Distribution of Cell Loss Ratio", IEEE JSAC, pp179-190, Feb. 1997

[48] E.Gelenbe, X.Mang, and R.Onvural, "Bandwidth Allocation and Call Admission Control in High-Speed Networks", IEEE Communications Magazine, pp122-129, May 1997

[49] R.Bolla, F.Davoli, and M.Marchese, "Bandwidth Allocation and Admission Control in ATM Networks with Service Separation", IEEE Communications Magazine, pp130-137, May 1997

[50] K.Liu, D.W.Petr, V.Frost, H.Zhu, C.Braun, and W.L.Edwards, "Design and Analysis of a Bandwidth Management Framework for ATM-Based Broadband ISDN", IEEE Communications Magazine, pp138-145, May 1997

[51] H.Saito, "Dynamic Resource Allocation in ATM Networks", IEEE Communications Magazine, pp146-153, May 1997

[52] C.Douligeris, and G.Develekos, "Neuro-Fuzzy Control in ATM Networks", IEEE Communications Magazine, pp154-162, May 1997

[53] S.Q.Li, S.Chong, and C.Hwang, "Link Capacity Allocation and Network Control by Filtered Input Rate in High-Speed Networks", IEEE/ACM Trans. on Networking, pp.10-25, Feb. 1995

[54] S.Chong, S.Li, and J.Ghosh, "Dynamic Bandwidth Allocation for Efficient Transport of Real-Time VBR Video over ATM", Proc. IEEE Infocom'94, pp81-90, 1994

[55] J.Ghosh, and Y.Shin, "Efficient High-order Neural Networks for Classification and Function Approximation", Int. Journal of Neural Systems, pp323-350, Vol.3,No.4, 1992

[56] P.R. Chang, and J.T. Hu, "Optimal Nonlinear Adaptive Prediction and Modeling of MPEG Video in ATM Networks Using Pipelined Recurrent Neural Networks", IEEE JSAC, Vol 15, No 6, pp1087-1100, August 1997

[57] S.Haykin, and L.Li, "Nonlinear Adaptive Prediction of Nonstationary Signals", IEEE Trans. On Signal Processing, pp526-535, Feb. 1995

[58] Alan V. Oppenheim & Ronald W. Schafer, Digital Signal Processing, 1975, Prentice-Hall

[59] Alan V. Oppenheim & Ronald W. Schafer, Discrete-Time Signal Processing, 1989, Prentice-Hall

[60] S.G.Mallat, "A Theory for Multiresolution Signal Decomposition: The Wavelet Representation", IEEE Trans. on Pattern Analysis and Machine Intelligence, pp674-693, July 1989

[61] Martin Vetterli & Jelena Kovacevic, Wavelets and Subband Coding, 1995, Prentice Hall

[62] Ingrid Daubechies, Ten lectures on Wavelets, 1992, Capital City Press

[63] Mary Beth Ruskai, etc., Wavelets and their Applications, 1992, Jones and Bartlett Publishers

[64] Leo Cohen, Time-Frequency Analysis, 1995, Prentice-Hall

[65] Michel Misiti, etc., Wavelet Toolbox, 1996, The MathWorks, Inc.

[66] W.E. Leland, M.S. Taqqu, W. Willinger, and D.V. Wilson, "On The Self-Similar Nature Of Ethernet Traffic (Extended Version)", IEEE/ACM Transactions On Networking, Vol.2,No.1, pp1-15, Feb.1994

[67] R.F.Rey, etc., Engineering and Operations in the Bell System, 2nd Edition, 1984, Bell Laboratories

[68] Herbert Taub & Donald L. Schilling, Principles of Communication System, 2nd Edition, 1986, McGraw-Hill

[69] Mischa Schwartz, Telecommunication Networks: Protocols, Modeling and Analysis, 1987, Addison-Wesley

[70] Uyless Black, data Networks Concepts, Theory, and Practice, 1989, Prentice Hall

[71] M. W. Garrett, "Contributions Toward Real-Time Services on Packet Switched Networks", Ph.D. Thesis, Chapter 4, Columbia University, May 1993

[72] A. Wong, C-T Chen, D. J. LeGall, F.-C. Jeng and K. M. Uz, "MCPIC: A Video Coding Algorithm for Transmission and Storage Applications", IEEE Commun. Mag., Vol 28, No 11, pp. 24--32, November 1990

[73] R.Gusella, "Characterizing the Variability of Arrival Processes with Indexes of Dispersion", IEEE JSAC, pp203-211, Feb.1991

[74] A.R.Reibman, and A.W.Berger, "Traffic Descriptors for VBR Video Teleconferencing Over ATM Networks", IEEE/ACM Trans. on Networking, pp329-339, June 1995

[75] M.Nomura, T.Fujii, and N.Ohta, "Basic Characteristics of Variable Rate Video Coding in ATM Environment", IEEE JSAC, pp752-760, June 1989

[76] R.Kishimoto, Y.Ogata, and F.Inumaru, "Generation Interval Distribution Characteristics of Packetized Variable Rate Video Coding Data Streams in an ATM Network", IEEE JSAC, pp833-841, June 1989

[77] P.Sen, B.Maglaris, N.Rikli, and D.Anastassiou, "Models for Packet Switching of Variable-Bit-Rate Video Sources", IEEE JSAC, pp865-869, June 1989

[78] Athansios Papoulis, Probability, Random Variables, and Stochastic Processes, 3rd Edition, 1991, McGraw-Hill

[79] ITU-T Recommendation I.412, Integrated Services Digital Network (ISDN) - ISDN User-Network Interfaces - ISDN User-Network Interfaces - Interface Structures and Access Capabilities, 1988

[80] The ATM Forum Technical Committee, ATM User-Network Interface Specification, Version 3.1, Sept. 1994

[81] The ATM Forum Technical Committee, Traffic Management Specification, Version 4.0, April 1996

[82] Oliver Rose, "Statistical properties of MPEG video traffic and their impact on traffic modeling in ATM systems". Proceedings of the 20th Annual Conference on Local Computer Networks, Minneapolis, MN, pp397-406, 1995

[83] S.Shioda and H.Saito, "Real-time Cell Loss Ratio Estimation and Its Applications to ATM Traffic Controls", INFOCOM'97, pp.1072-1079, 1997

[84] I.Reljin, "Neural Network Based Cell Scheduling in ATM Node", IEEE Communications Letters, pp.78-80, March 1998

[85] Q.Zhu, J.Kneuer, and W.Huang, "A Study of Temporal Behavior Between the Packet Loss Bursts in Packet Switched Systems", IEEE ICC'95, pp.1930-1936, June 1995

[86] F.Chiussi, J.Kneuer, and V.Kumar, "Low-Cost Scalable Switching Solutions for Broadband Networking: The ATLANTA Architecture and Chipset", IEEE Comm. Mag. pp.44-53, December 1997

[87] A.Grossmann, R.Kronland-Martinet, and J.Morlet, "Reading and Understanding Continuous Wavelet Transforms", Wavelets, Time-Frequency Methods and Phase Space, Springer-Verlag, Berlin, 1989

[88] A. Grossmann, and J. Morlet, "Decomposition of Hardy Functions into Square Integrable Wavelets of Constant Shape", SIAM Journal of Math. Anal. Pp.723-736, July 1984

[89] S.Ma, and C.Ji, "Modeling Video Traffic in the Wavelet Domain", IEEE Infocom'98, pp.201-208, April 1998

[90] N.Ahmed, T.Natarajan, and K.R.Rao, "Discrete Cosine Transform", IEEE Trans. On Computers, pp.88-93, Jan.1974

[91] S.Kirkpatrick, C.D.Gelatt, and M.P.Vecchi, "Optimization by Simulated Annealing", Science 220, pp.671-680, 1983

[92] D.H.Ackley, G.E.Hinton, and T.J.Sejnowski, "A Learning Algorithm for Boltzmann Machines", Cognitive Science 9, pp.147-169, 1985

[93] R.Ting, R.Jeremiah, J.Barba, J.Kneuer, I.Habib, and J.Zhu, "A Bandwidth and Multiscale Analysis of ATM Traffic", Proc. of SCS/IEEE SPECTS'98, pp.18-24, July 1998

[94] R.Jeremial, R.Ting, and J.Barba, "A Bidimensional Multiresolution Analysis Without Edge Effects", submitted to IEEE Trans. on Image Processing

[95] R.Jeremial, R.Ting, and J.Barba, "A Generalized DCTWM Model", submitted to IEEE Trans. on Circuit and Systems for Video Technology

[96] R.Ting, et al., "A Novel Adaptive System for VBR/ABR Traffic Analysis and Bandwidth Reservation with Wavelet-based Multiscale Decomposition", submitted to IEEE JSAC