DUBLIN CITY UNIVERSITY
SCHOOL OF ELECTRONIC ENGINEERING

A Review of Connection Admission Control Algorithms for ATM Networks

A Thesis Submitted for the award of a MEng.
Electronic Engineering.

Maureen Curran
BA, BAI

Supervised by **Dr. M. Collier**

11 September, 2002

I hereby certify that the material, which I now submit for assessment on the program of study leading to the award of Masters in Electronic Engineering is entirely my own work and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the text of my work.

Signed: Maureen Curran **ID Number: 9597 1360**

Candidate

Date: 12th September 2002

Acknowledgements

I dedicate this work to the loving memory of my mother, Dr. Elizabeth Bastible-Curran, a wonderful scholar and physician, and to my family who provided constant help and support. I would like to thank Dr. Martin Collier for his guidance and advice throughout the project, and all who provided encouragement and help.

Abstract

The emergence of high-speed networks such as those with ATM integrates large numbers of services with a wide range of characteristics. Admission control is a prime instrument for controlling congestion in the network. As part of connection services to an ATM system, the Connection Admission Control (CAC) algorithm decides if another call or connection can be admitted to the Broadband Network. The main task of the CAC is to ensure that the broadband resources will not saturate or overflow within a very small probability. It limits the connections and guarantees Quality of Service for the new connection. The algorithm for connection admission is crucial in determining bandwidth utilisation efficiency. With statistical multiplexing more calls can be allocated on a network link, while still maintaining the Quality of Service specified by the connection with traffic parameters and type of service.

A number of algorithms for admission control for Broadband Services with ATM Networks are described and compared for performance under different traffic loads. There is a general description of the ATM Network as an introduction. Issues to do with source distributions and traffic models are explored in Chapter 2. Chapter 3 provides an extensive presentation of the CAC algorithms for ATM Broadband Networks. The ideas about the Effective Bandwidth are reviewed in Chapter 4, and a different approach to admission control using online measurement is presented in Chapter 5. Chapter 6 has the numerical evaluation of four of the key algorithms, with simulations. Finally Chapter 7 has conclusions of the findings and explores some possibilities for further work.

Table Of Contents

Table of Figures

Chapter 1

Introduction

1.1 Introduction

This is a report on the research work undertaken to study Connection Admission Control (CAC) algorithms for Broadband Asynchronous Transfer Mode (ATM) networks *[1] -[9]*. There is numerical evaluation performed with experimental trial simulations of four important algorithms. The simulations were designed and developed to highlight key aspects of the algorithms. The results provide some interesting conclusions about the algorithms and their features. There is a description of other research work concerning a wide range of the algorithms.

The thesis is organized as follows. This Chapter outlines some of the standards and recommendations for ATM technology. Techniques are described in Chapter 2 for the modeling of ATM networks. There is a report on the algorithms as described in the current literature in Chapters 3, 4 and 5. The experimental work performed by the author is found in Chapter 6. The purpose of these numerical evaluations and simulations is to look at the properties of the key algorithms in greater detail. The conclusions are discussed in Chapter 7, with indications for the possibilities of further work. There is extra detailed information about the Effective Bandwidth concepts in the Appendix.

1.2 Resource Allocation in ATM Networks

The exploration of algorithms for connection admission reveals many interesting implications for resource allocation in ATM networks *[1]-[8]*. The purpose of the CAC Algorithm is to ensure that connections will be admitted provided that the probability is very small that network resources will saturate or overflow. The CAC algorithm plays a vital role in the management of these resources. The most effective solutions are achieved with the efficient use of bandwidth allocation. The algorithm for connection admission control is crucial in determining bandwidth utilization efficiency.

The network consists of shared resources such as bandwidth and internal buffering capacities. The resources are reserved along the path between the source and destination nodes of each call. As part of the congestion control strategy, the network uses the CAC algorithm.

Traffic control is the set of actions taken by the network to avoid congestion.

The primary role of traffic control is to achieve the pre-defined network performance objectives while meeting the requirements of Quality of Service (QoS). Traffic control is based on a combination of the Usage Parameter Control (UPC) procedure *[1] [7][8]* and the CAC algorithm, to monitor the network for congestion. The UPC monitors the user-network interface to ensure the accepted rate is not exceeded, while the CAC algorithm ensures that resources are available for a new connection.

The UPC uses the Generic Cell Rate Algorithm (GCRA) or 'leaky bucket' *[7][8]* mechanism to check ATM cell flow levels. The policing and monitoring mechanism also uses explicit feedback to sources to assure that capacity is fairly allocated. The connection is *compliant* as long as the proportion of non-conforming cells does not exceed thresholds established on the connection by the traffic contract. The CAC and UPC procedures take the Connection Traffic Descriptor and requested Quality of Service to set up a compliant connection, described in the coming sections.

1.3 Different Approaches for Connection Admission Control

There are two main approaches to admission control. The parameter-based approach computes the amount of network resources required to support a set of calls with pre-defined traffic characteristics. The second approach is the Measurement-Based approach, which relies on the measurement of actual traffic in making admission decisions. The evaluation of an algorithm for CAC depends on how well it fulfils its primary role of ensuring that service commitments are not violated. Other evaluation criteria are network utilization and implementation and operational costs.

Measurement-Based algorithms *[10]-[13]* have no prior knowledge of the traffic statistics and make the admission decision based on the current state of the network only. In contrast to the other algorithms which look at the characteristics of source traffic and represent them as parameters, Measurement-Based algorithms make decisions on a monitored amount of traffic on the network. It is found that it can achieve the same performance as that of an optimal scheme based on the knowledge of traffic statistics.

1.4 Broadband ISDN and ATM

The adjective 'broadband' refers to the high capacity of networks available to support digitised communications. It enables them to transport large amounts of information such as real-time video. Asynchronous Transfer Mode (ATM) *[1] - [4][8][9][15]* is a protocol used in broadband networks. This technology has as its basis the ATM cell, a small packet of data of 53 bytes in length. The short length facilitates transport of real-time services. This universal network follows the standards and recommendations devised by the international network governing bodies, the International Telecommunications Union (ITU, formerly CCITT) and the industry established group, the ATM Forum *[7][8]*.

1.5 ATM Protocol Architecture

ATM is a streamlined packet-switching protocol *[1]-[4]*, with reduced overhead in processing of ATM cells. ATM operates at high data rates, ranging from 155.52Mbps to 10Gbps. The support of multiple line rates is a key advantage of ATM and allows for seamless inter-working of LANs and WANs. The protocol can be implemented in a variety of ways to allow the integration of legacy systems, while improving overall network performance. The basic protocol stack for ATM is illustrated in *Figure 1.1*.

The physical layer standards give the specification of the transmission medium and the signal-encoding scheme. The ATM layer is common to all services and packet transfer capabilities. The ATM Adaptation Layer (AAL) is service dependent and facilitates the support of transport layer protocols such as TCP. The protocol reference model also indicates three separate planes, a user plane to transport user information with flow control and error control, a control plane for call establishment and connection control, and a management plane for co-ordination between planes and overall management functions.

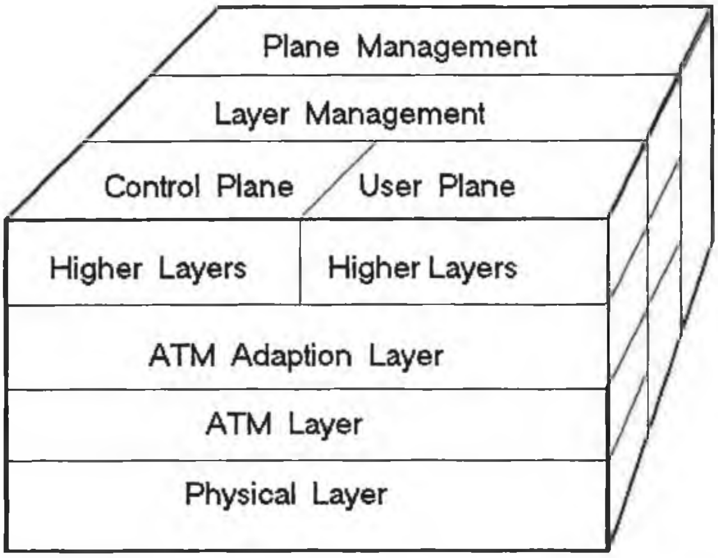

Figure 1.1 ***The B-ISDN ATM Protocol Reference Model***

1.6 ATM Connections

The ATM connections are viewed abstractly as logical connections. These 'logical' connections are referred to as 'virtual', as Virtual Channel Connections (VCCs) or Virtual Channels (VCs). They allow different allocations of bandwidth depending on the service. A VCC is set up between two end users through the network. A variable-rate, full-duplex flow of fixed size cells can be exchanged over the connection and regulated in different ways depending on the exact service provided.

The concept of a virtual path (VP) and virtual path connections (VPCs) is used for grouping and planning VC connections. A number of VCCs grouped together form a VPC. The advantages of using virtual paths mean that network architectures are simplified with reduced processing and connection set-up time. The VPCs may be established by prior agreement for a semi-permanent connection, or it may be customer controlled instead of network controlled. B-ISDN Recommendation I.150 specifies methods for providing the establishment /release facility for VCCs.

1.7 The Requested Quality of Service Class

The requested Quality of Service (QoS) class is negotiated during connection establishment. The ATM network is designed to transfer many different types of traffic simultaneously, including real-time flows such as voice, video and bursty TCP flows. The way each stream of cells is handled is defined by its Quality of Service category and depends on the requirements of the application and the characteristics of the traffic flow. For example, real-time video traffic must be delivered within a minimum variation in delay.

The following service categories have been defined by ATM Forum:

Real-time service:

- Constant Bit Rate (CBR)
- Real-time Variable Bit Rate (rt-VBR)

Non-real-time service:

- Non-real-time Variable Bit Rate (nrt-VBR)
- Available Bit Rate (ABR)
- Unspecified Bit Rate (UBR)
- Guaranteed Frame Rate (GFR)

QoS is evaluated in terms of Cell Loss Ratio (CLR), Cell Transfer Delay (CTD), Cell Delay Variation (CDV), and Minimum Cell Rate (MCR) when applicable. The Peak Cell Rate (PCR), Sustained Cell Rate (SCR) and Burst Tolerance (BT) are known as Source Traffic Descriptors. In order to simplify network management, a given number of parameter combinations have been identified and grouped into classes, called service classes.

The ATM Adaptation Layer (AAL) defines four classes of service in the ITU-T recommendation I.362 as follows:

- **Class A** has a time relation between source and destination. The bit rate is constant and the service is connection-oriented. An example is voice or fixed bit rate video.

- **Class B** also has a time relation between source and destination. The bit rate is variable and the service is connection-oriented. An example is variable bit rate video and audio.

- **Class C** does not have a time relation between source and destination. The bit rate is variable and the service is connection-oriented. An example is connection-oriented data transfer and signaling.

- **Class D** is connectionless. An example is Switched Multimegabit Data Service (SMDS).

1.8 Guaranteed Frame Rate

Guaranteed Frame Rate (GFR) has recently been approved by the ATM Forum *[16]*. It is an important service category that supports TCP/IP traffic for ATM. It provides bandwidth guarantees while being as easy to use as the Unspecified Bit Rate (UBR) service category. Like UBR, it allows the end system to transmit cells at the line rate of their ATM adapter. The GFR is different to UBR, as it requires the network elements to discard AAL frames when congestion occurs. Another difference is that GFR allows the user to reserve bandwidth. This means the user is guaranteed that transmitting at a minimum rate will be without losses.

1.8.1 The GFR Traffic Contract

The GFR traffic contract is composed of four main parameters:

- Peak Cell Rate (PCR)
- Minimum Cell Rate (MCR)
- Maximum Burst Size (MBS)
- Maximum Frame Size (MFS)

PCR is the maximum rate and is often set at the line rate of the ATM adapter of the end system. The MFS is the largest size of AAL5 frame that the end systems can send. The MBS defines the maximum burstiness allowed for the traffic with minimum guaranteed bandwidth.

1.9 The Quality of Service Parameters

The ATM Forum defines the following QoS parameters:

- Peak-to-peak Cell Delay Variation (CDV)
- Maximum Cell Transfer Delay (maxCTD)
- Cell Loss Ratio (CLR)

1.10 The Traffic Contract and the CAC Algorithm

At the connection setup stage a contract is established between the user and the network. The user specifies the source traffic descriptors and the desired Quality of Service. Based on these parameters, the CAC decides whether to accept or reject a connection. A connection request is accepted only when sufficient resources are available to satisfy the QoS requirements of both existing and new connections. If the request is accepted, the network contracts to meet these QoS objectives as long as the user complies with the traffic parameters declared.

When a new connection is requested, the user must specify the service required for the connection. A connection request must include of the following information about the connection:

- Service category (CBR, rt-VBR, nrt-VBR, ABR, UBR, GFR)
- Connection Traffic Descriptor
- Requested and accepted value of each QoS parameter (peak-to-peak CDV, maxCTD, CLR)

By accepting the connection request (i.e. providing the user with the connection requested) the network forms the traffic contract with the user for that connection.

1.11 Traffic Parameters

The CAC represents a set of actions taken by the network at call setup phase in order to accept or reject the connection. Their values established at connection set-up are called traffic parameters. The parameters are held in what is known as the Traffic Descriptor, which also has parameters for Quality of Service. A traffic contract between the source and the network is negotiated. The parameters for traffic and Quality of Service are allocated at connection setup. The values of these parameters depend on the type of service and link capacity required for that connection.

1.11.1 Traffic Parameter Specification

The traffic characteristics of connections are described by a set of standardized traffic parameters. Traffic parameters are a specification of a particular traffic aspect, e.g.

- Peak Cell Rate (PCR)
- Minimum Cell Rate (MCR)
- Average Burst Duration

Different services specify different values for the Peak Cell Rate (PCR) and a Minimum Cell Rate (MCR) required for that connection a connection. The network resources are allocated so that all connections receive at least their MCR capacity. The remaining unused capacity may then shared in a fair and controlled fashion among all the sources *[1]-[4][7][8]*.

1.11.2 The Traffic Contract Specification

The traffic contract negotiated during connection establishment has the following key components:

- The Connection Traffic Descriptor
- The requested QoS class
- Definition of a compliant connection

1.11.3 The Connection Traffic Descriptor

The Connection Traffic Descriptor is made up of the Source Traffic Descriptor and Cell Delay Variation tolerance. The Source Traffic Descriptor consists of a set of parameters, which indicate the agreed traffic settings for the User Network Interface (UNI) agreed with the network when setting up the connection. The Source Traffic Descriptor parameters are:

- Peak Cell Rate (PCR)
- Sustainable Cell Rate (SCR)
- Burst Tolerance (τ) or Maximum Burst Size (MBS)
- Minimum Cell Rate (MCR)

1.12 Summary

The CAC algorithm works with other congestion control procedures and routing algorithms to ensure that traffic congestion is minimised and that sufficient network resources are available to support the connection. The criteria for admission are that the Quality of Service standards required by the particular type of connection can be met with the network resources available, without compromising existing connections. The outcome of the CAC process is the traffic contract at the UNI, which includes the definition of a compliant connection to ensure that the requested Quality of Service is achieved.

Bandwidth is a fundamental network resource, and its efficient allocation to new connections is part of the admission control process. CAC algorithms present a range of possibilities to increase bandwidth utilisation with statistical multiplexing. This means that more connections can be allocated than the available bandwidth for their combined peak rates, because the likelihood that the peak rate occurs for traffic from all connections at the same time is small. This likelihood or probability is evaluated, and must be within Quality of Service definitions so that the connection still complies with the traffic contract. There are a number of issues concerned with how to meet conflicting goals of the CAC algorithm. These are to maximize bandwidth utilization through efficient statistical multiplexing while still ensuring that each connection has the QoS agreed by the traffic contract at connection setup time.

Chapter 2

ATM Network Modeling

2.1 Introduction

There are a variety of methods and approaches used in the modeling of ATM networks *[1]-[9][14][15][17]-[19]*. This Chapter describes theory relevant to the analysis and modeling of CAC algorithms. First there is an overview of stochastic processes to represent the streams of traffic in the network. Different forms of traffic modeling are then described. The concepts of queuing models and fluid flow approximation are presented. Finally there is a discussion of timescale analysis of traffic *[6][20][21]* from high level to cell level detail. The stochastic traffic models included in the next sections are 'ON-OFF' bursty sources, Markov modulated sources, and Self-similar traffic models.

2.1.1 Stochastic Processes

The traffic on the connections to be multiplexed together at the ATM switch is represented by stochastic processes *[5][6][14][15][17][18]*. A stochastic process is a parameterised family of random variables, $\{X\ (t),\ t \in T\}$, where the parameter t is usually time, T is the index set. If T is a countable set it means the process is a discrete parameter process. Otherwise it is a continuous parameter process. The set of random variables $X(t)$ have a state space, which may be discrete or continuous. The state space of a process is the set of all possible values of the random variables. Each of these values is called the 'state' of the process. The state space or phase space of the process *[10]* is the set S of $X_0 ,X_{1,} X_2 \ldots X_n$ a sequence of n random variables whose ranges are contained in S.

The properties of a stochastic process can be used to represent a cell arrival process by characterising the inter-arrival time distribution. The inter-arrival time distribution is the probability of an arrival in a given time interval. The mean and variance of this distribution, and its multivariate probability mass functions (pmfs) can be found for this stochastic process *[18]*.

2.1.2 Markov Processes

A Markov process *[8][9][12]-[14]* is one where the present state of the process determines the future of the process, and full knowledge of its past is not required. A Markov process is called a Markov chain if its state space is discrete.

- A **Discrete Time Markov Chain (DTMC)** is a process which makes transitions from one state to another at well defined instants t_n. The DTMC is fully determined when the one-step transition probabilities are known. These are the set of numbers P_{ij}, representing the probability of transition from state i to state j. These can be arranged into a one-step transition probability matrix $P = (P_{ij})$ where:

$$\sum_{j=0}^{\infty} P_{ij} = 1 \ \forall \text{i.}$$

P is called a stochastic matrix and each row elements sum to 1.

- A **Continuous-Time Markov Chain (CTMC)** is a Markovian chain where transitions from state i to state j occur in continuous time, and this requires some extra equations. In addition to the transition probability matrix P, there is a transition density matrix Q, also called the *infinitesimal generator* of the Markov chain *[9]*.

$$Q(t) = \lim_{\Delta t \to 0} \frac{P(\Delta t) - 1}{\Delta t}$$

This means that the elements q_{ij} of matrix Q have a probability of $\Delta t\, q_{ij}$ of moving from state i to state j in interval Δt.

2.3 Traffic Models

Traffic models *[1]-[9][14][15][17][18]* provide a means of evaluation for the appraisal of flows in telecommunications networks. ATM networks need to provide performance guarantees to their connections. To estimate if a new connection is to be admitted, the flows in the network can be represented by various traffic models. Traffic models are divided into two classes, short-range and long-range dependent. Examples of short-range dependent models are Markov processes and Regression models. They have a correlation structure that is significant for small time lags. Long-range dependent models such as Fractal Autoregressive Integrated Moving Average (F-ARIMA) and Fractal Brownian motion have significant correlations for longer time lags.

2.3.1 Markov and Embedded Markov Models

The activities of a source can often be modeled by a finite number of states, where a set of random variables $\{X_n\}$ forms a discrete Markov chain *[14][17][18]*, and where the probability of the next value $\{X_{n+1}\}$ depends only on the current state. In a simple Markov model, each state transition represents a new arrival. Therefore their inter-arrival times are exponentially distributed (for CTMC) or have arbitrary distributions for semi-Markov processes, with an embedded discrete time Markov chain.

2.3.1.1 'ON-OFF' Source Models

The 'ON-OFF' source is widely used to represent bursty traffic sources in source characterisation for traffic modeling *[1][3][5][6][19]*. The information is sent as a series of 'ON' and 'OFF' periods, see *Figure 2.1.* The information is transmitted at peak rate for the 'ON' period, and none is transmitted in the 'OFF' period. The geometric distribution is used if the network is modelled as a discrete-time system.

The source may switch from the 'ON' to the 'OFF' state according to a CTMC with two states. The information emission process is a two-state Markov modulated Poisson Process (MMPP), with zero rates in one state. Because the sojourn time in the state of a continuous Markov chain is exponentially distributed, the burst ('ON') and silence ('OFF') times or sojourns of the source are exponentially distributed. The discrete analogue of this source type is a Discrete-Time Markov Chain (DTMC) with two states and has transitions occurring at periodic instances $t_n = n\Delta t$. In modeling ATM networks Δt is chosen as the duration of a timeslot.

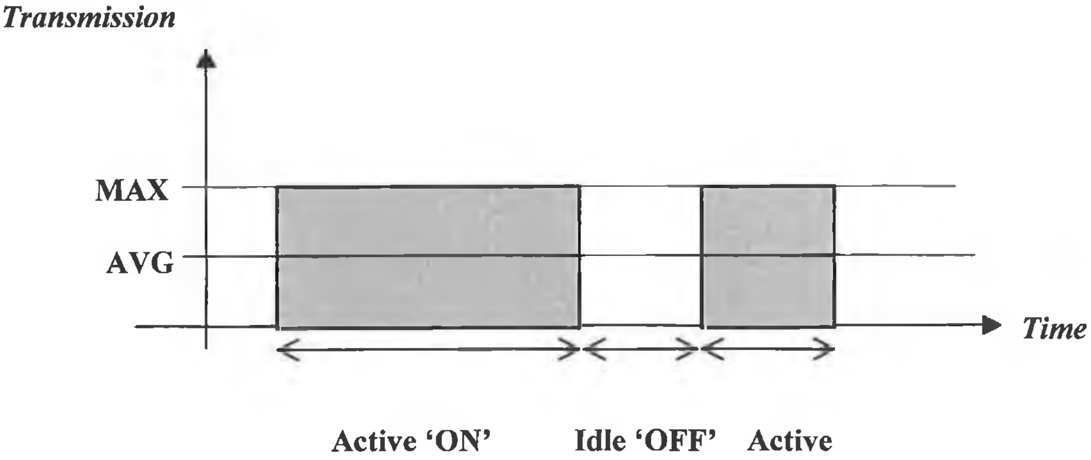

Figure 2.1 *Bursty 'ON-OFF' Sources*

2.3.2 Markov Modulated Poisson Process (MMPP) Models

The Markov modulated process *[5][9][14][17][18]* is a generalisation of the 'ON-OFF' process, which has two states, to allow $m \geq 2$ states. When in state i the source emit at a rate r_i and then switches to another state j at rate r_j. The embedded process consisting of the changes of state is assumed to be a Markov chain, so the arrival process is called a Markov modulated Rate Process *(MMRP*). MMPP can be used to model traffic integration from different source types. The arrival of cells from one type of source in state k is assumed to be Poisson with rate λ_d, while another type can also be Poisson with rate λ_k. The resulting state s_k will be $\lambda_d + \lambda_k$. The performance measures such as queuing distribution and the moments of the delay distribution are obtained using *MMPP/G/1* queue analysis *[15]*.

2.3.3 Self-Similar and Long-Range-Dependent Processes

The term self-similar or 'fractal' can be applied to traffic that looks "the same" on all time-scales, with the important characteristic that it has long-range dependence, or the existence of correlation over a broad range of time scales *[12][13][20][22][23]*. For a stationary process only a "lag" $j\text{-}i = k$ is relevant. The definition for a stochastic process X_k is a process with mean $E\{X_k\} = \overline{X}$ and autocorrelation function $r(i) = C(i)/\mathrm{var}\{X_k\}$, where $C(i)$ is the autocovariance. The processes $X_k^{(m)}(m = 1,2,...)$ are constructed out of X_k as:

$$X_k^{(m)} = (\sum_{n=0}^{m-1} X_{km+n}) \,/\, m$$

i.e. by averaging over non-overlapping blocks of size m. The processes $X_k^{(m)}$ have a mean x and autocorrelation function $r_m(i)$ *[5]*. The process is called second-order self-similar if $r_m\,(i){=}r(i)\; for\;\; m,i \to \infty$.

An important parameter of a long-range-dependent process is the Hurst parameter, $H = 1 - \beta/2$ where $0<\beta<1$. Given a set of experimental data a_k $(k{=}1,2,...,n)$ with sample mean:

$$\overline{a}_n = \sum{}_{k=1}^{n} a_k \Big/ n$$

and sample variance:

$$S^2(n) = \sum_{k=1}^{n} [a_k - \overline{a}_n^2] \big/ (n-1)$$

define the rescaled adjusted range *(R/S)*, where *R* is the autocorrelation and *S* is autocovariance:

$$\frac{R(n)}{S(n)} = \frac{1}{S(n)}[\max(0, W_1, W_2, ...W_n) - \min(0, W_1, W_2, ...W_n)]$$

whereby $W_k = (a_1, a_2, ..., a_k) - kE[a(n)]$.

The quantities W_k measure the deviation of the process from the 'expected value'. $R(n)$ measures the values of this deviation. A value for $H = 0.5$ and greater implies that the process is self-similar.

2.4 Queuing Models and Analytical Solution Methods

In the various models described in the previous sections, sources are represented by the arrival processes and the network by buffered systems that queue the traffic at various nodes and switches *[5][6][14][15][17][18]*. The queuing systems are represented using Kendell notation *[15]* to summarise the type of arrival process, service time distribution and system capacity in a letter and number notation. The impact of burstiness or congestion is seen in terms of buffer overflow probability or Cell Loss Probability, and this is an important Quality of Service criterion for admission control algorithms.

The analytical solution methods used to find the equilibrium distributions of buffer occupancy and waiting times are an important aspect of modeling *[6]*. The three main methods of solution are matrix analytical method, probability generating functions and the fluid flow approximation method. The fluid flow approximation method is described next.

2.4.1 Fluid Flow Approximation

The simple 'First-In-First-Out' (FIFO) queue with exponentially distributed 'ON-OFF' sources can be used to analyse a statistical multiplexer fed by bursty sources *[6][15][19]*. The differential equations describe the contents of the buffer. They can be modeled by assuming the filling process to be a Markov process and the service time to be constant. A knowledge of the probability of exceeding the buffer capacity is important for admission control algorithms, as the Cell Loss Probability (CLP) can be represented by this. It is the admission criterion for many of the algorithms.

Let $\{Y(t), t \geq 0\}$ be a CTMC that takes the values $\{0,1,\ldots,N\}$ and let the infinitesimal generator be the matrix $\boldsymbol{Q}$ with elements q_{ij}. When the Markov chain is in state j the fluid arrives with a rate a_j. The buffer drains at a constant rate c, and so the net rate of change of the buffer contents is r_j:

$$r_j = a_j - c$$

Let $X(t)$ denote the buffer contents at time t. $X(t)$ is a continuous random variable satisfying $0 \leq X(t) \leq K$, where K is the buffer size. The equilibrium overflow probability of the buffer beyond level x is:

$$G(x) = \lim_{t \to \infty} \Pr[X(t) > x]$$

The contents of the buffer is a queue represented by the bivariate stochastic process $[X(t), Y(t)]$ with a joint pdf-pmf $F_j\,(x,t)$:

$$F_j(x,t) = \Pr[X(t) \leq x, Y(t) = j]$$

for $(j = 0 \ldots N)$.

F_j (x,t) is the probability that at time *t*, the buffer is filled to at most level *x* and the modulator is in phase *j*. When the system has reached equilibrium:

$$F_j(x) = \lim_{t \to \infty} F_j(x,t) = \Pr[\, X \le x, Y = j\,]$$

the *(N+1)* -dimensional row vector ***F(x)*** is:

$$\mathbf{F}(x) = [F_0(x), F_1(x), \ldots, F_N(x)]$$

which allows the equilibrium overflow probability to be expressed as:

$$G(x) = \Pr[\, X > x\,] = 1 - \mathbf{F}(x).\mathbf{1}$$

where $\mathbf{1} = (1,1,\ldots,1)^T$, where ^{T} is for transpose and '**.**' for scalar product.

The time evolution of $F_j(x,t)$ is governed by the following equation:

$$F_j(x,t+\Delta t) = \sum_{i,i\neq j} q_{ij}\Delta t F_i(x - r_{ij}\Delta t, t) + \left(1 - \sum_{i,i\neq j} q_{ij}\Delta t \right) F_j(x - r_j\Delta t, t) + o(\Delta t)$$

The probability that at time *t* + *Δt* the buffer is filled to at most *x* and that the modulator is in state *j* consists of two terms. The net rate of change is r_{ij}. Firstly, in order to progress to be in state *j* from state *i* at $t + \Delta t$, it undergoes the transition from *i* to *j*, which happens with probability q_{ij} *Δt,* and the buffer contents changes by $x - r_{ij}$ in the interval *Δt.* The second term is similar without the phase transition of the modulator.

$F_j(x,t)$ is subtracted from both sides, and it is divided by *Δt* while letting *Δt→0.* The properties of the infinitesimal generator means the following is obtained:

$$\frac{\partial F_j}{\partial t}(x,t) + r_j \frac{\partial F_j}{\partial x}(x,t) = \sum_i q_{ij} F_i(x,t)$$

for the equilibrium solution, $\delta F_j / \delta t = 0$, resulting in a set of equations:

$$r_j \frac{\partial F_j}{\partial x}(x) = \sum_{i=0}^{N} q_{ij} F_i(x)$$

for *(j = 0, ...n)*. This equation has the rate matrix ***R***, and **Q** is the transition density matrix or infinitesimal generator of the Markov chain *[5][14]*, and it may be written as:

$$\frac{\partial}{\partial x} \mathbf{F}(x).\mathbf{R} = \mathbf{F}(x).\mathbf{Q}$$

This is a linear first-order differential equation, with a solution that is a linear combination of exponentials. The solution is:

$$\mathbf{F}(x) = \mathbf{\Phi} e^{\mathbf{QR}^{-1}x}$$

where Φ is a constant row vector. The exponentials are of the form:

$$e^{z_i x}$$

where z_i, $i = 0...N$ are the eigenvalues of:

$$\mathbf{QR}^{-1}$$

To solve fluid-flow models, the eigenvalues and vectors are found. It is possible to find closed-form expressions of the eigenvalues and eigenvectors for homogeneous 'ON-OFF' sources.

2.5 Timescale Analysis

The issue of timescale analysis is an important one as the estimation of requirements differs according to the timescale *[6][19]* or the time duration under consideration. The timescales usually considered are:

- **Cell or Packet level**, i.e. the inter-arrival time between cells, in microseconds.

- **Burst Level**, i.e. the cell arrival groups that occur as a 'burst' of traffic, in milliseconds.

- **Call or connection level**, i.e. average time for the VC connection setup, in minutes.

The ATM network is a network of queues, and the consideration of delay is closely correlated with the buffer sizing along the links. Size of buffers can be categorised according to timescales, cell scale buffers deal with congestion at the cell level, i.e. simultaneous arrival from different sources. Larger buffers at burst scale can accommodate burst traffic, such as a data file transfer, thus increasing delays but decreasing Cell Loss Probability. The review of resource allocation in *[3][4][19][20]* uses timescale analysis to evaluate congestion at different levels with respect to integrated traffic of different services. Congestion is measured in terms of the blocking probabilities at each level, i.e. cell blocking, burst blocking and call blocking.

2.6 Statistical Multiplexing

This section presents some essential ideas of statistical multiplexing [14][18][45]. Statistical multiplexing results in the allocation of a bandwidth less than that required for PCR of a connection source. It is based on the idea that there is a probability that all sources are not transmitting all together all the time. The following sections give an explanatory example of statistical multiplexing as a background to the algorithms that have been presented.

2.6.1 The ATM Multiplexer

The ATM multiplexer *[9][20][51][66]* is described as a buffer and a high-speed link. *Figure 2.2* illustrates the buffer receiving the cells generated by establishing a new connection, with excess cells lost or delayed. The cell loss and delay are found from the QoS requirements and the admission policy or CAC ensures that these requirements are met.

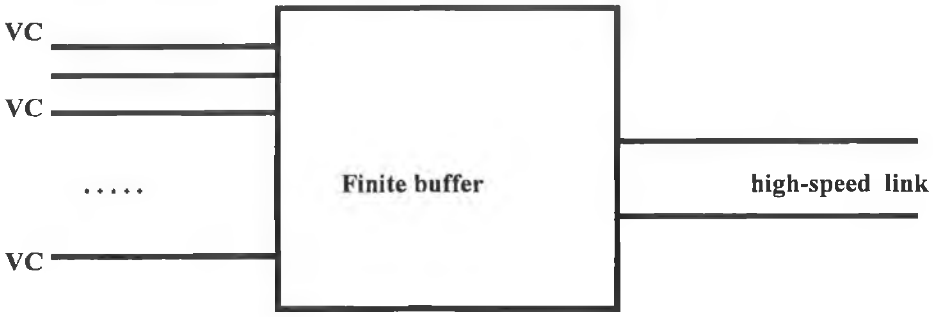

Figure 2.2 An ATM Multiplexer Model

The ideas of virtual connections 'VC' are presented in Section 1.6. The peak rate of the *VC* is defined as follows. If the *VC* generates cells with the minimum spacing of *1/T* cells per second, then its *1/T x 53 x 8 bits* per second. The units of *C* the capacity of a high-speed link can be in bits per second. A buffer is required at the interface between the incoming cell streams and the high-speed link in order to limit the effect of cell scale congestion or burst scale congestion.

2.6.2 Statistical Multiplexing of Connections

Statistical multiplexing allows for the allocation of a bandwidth less than that required for PCR of a connection source. The allocated amount of the shared link is less than that of the peak rate, and so the overall capacity is used more efficiently. The statistical gain is therefore the ratio of the number of accepted connections using multiplexing to those accepted using peak bit rate allocation.

The peak demand of all multiplexed connections may exceed link capacity, but this will only occur with only a small probability. This probability must be less than the maximum value specified by the Quality of Service requirements. The network must be able to determine in real-time how much bandwidth to allocate for statistical multiplexing.

2.6.3 An Example of Multiplexing

If there are a number of 'ON-OFF' connections multiplexed together at a bufferless switch, if each stream is seen as a continuous flow of cells (fluid-flow model) the aggregate bit rate distribution can be computed. This is done by a convolution of the bit rate distribution of each connection, assuming all connections are independent. Under the previous assumptions, the Cell Loss Probability (CLP) is accurate, but does not meet real-time requirements, but as an example highlights the features of multiplexing.

Consider two types of classes of traffic:

Type 1:	Peak Cell Rate 10Mbps, Mean Cell Rate 2Mbps
Type 2:	Peak Cell Rate 2Mbps, Mean Cell Rate 1Mbps

The total link capacity is 150Mbps, the diagram in *Figure 2.3* show the solid line that represents the maximum numbers of sources from each type that can be accepted by the network to comply with the requested QoS.

If there are 35 *Type 1* sources, each source has a Bandwidth of 4.28Mbps (or 150/35), in the same way 120 *Type 2* sources have a Bandwidth equal to 1.25Mbps. If 50 *Type 2* sources are multiplexed together, a maximum of 19 *Type 1* sources can be multiplexed at the same time on the common link.

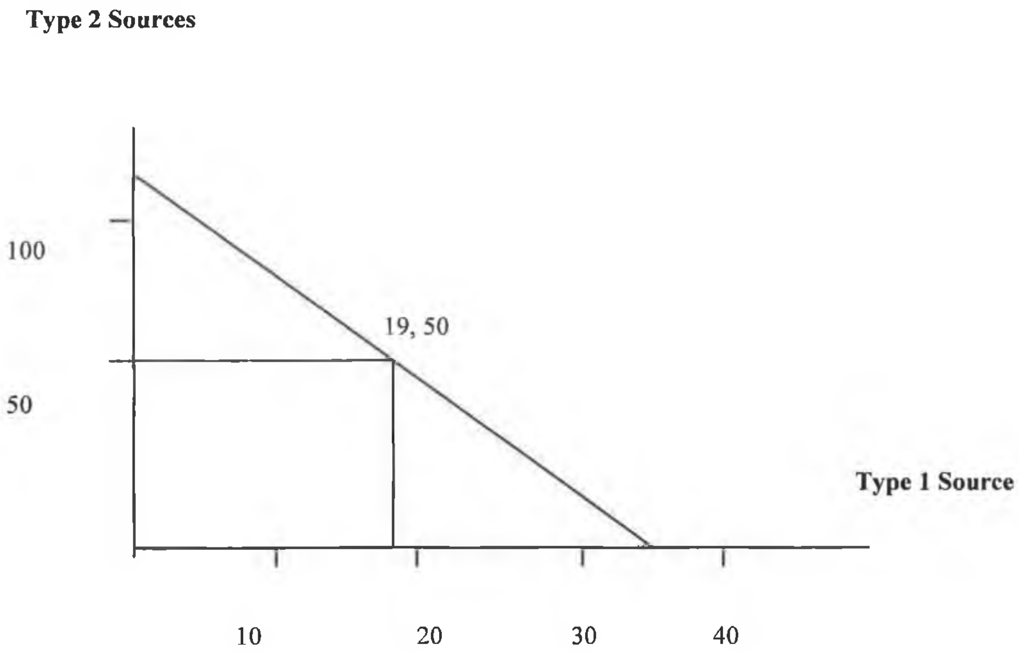

Figure 2.3 Multiplexing of Two Sources

2.7 Summary

The review of Traffic Modeling and the characterisation of sources should provide a background of understanding for the following Chapters in which the CAC algorithms are examined. The probability of cell loss must be within a certain value to achieve the required Quality of Service. The sources can be represented by a few parameters, and the loss probabilities can be calculated easily for each algorithm. Markov models with 'ON-OFF' sources represent bursty traffic such as video. A new area of interest is that of Self-similarity found in traffic traces. Issues of different timescales can be considered, ranging from cell level to connection level. The numerical modeling of a number of key algorithms uses the fluid flow approximation model with FIFO queuing in Chapter 6. These algorithms are described in the next Chapters.

Chapter 3

Connection Admission Control Algorithms

3.1 Introduction

This Chapter gives descriptions of the various CAC algorithms presented in the literature *[9]-[13][19]-[66]*. The descriptions provide a basis for the numerical evaluations of the algorithms in Chapter 6. The algorithms are presented in detail in this Chapter. The following algorithms were chosen:

- The Convolution Algorithm
- The Chernoff Bound Algorithm
- The Gaussian Approximation Algorithm
- Algorithms for Timescale Analysis
- A Decision-Theoretic Approach Algorithm
- A Dynamic CAC Based on the Arrivals Distribution
- Algorithms using Neural Networks, Fuzzy Logic and Artificial Intelligence Techniques
- Algorithms with Prioritised Traffic Types
- Algorithms Based on Simulation and Reinforcement Learning

The algorithms depend on a wide range of fundamental principles. The Convolution, Chernoff Bound and the Gaussian Approximation Algorithm are based on mathematical approximations. The Convolution Algorithm uses the bufferless fluid flow model to find the aggregate source rate. There is an estimation of Cell Loss Probability as the encapsulating Quality of Service requirement. The Chernoff Bound Algorithm uses a similar approach. The Chernoff Bound Algorithm can be used together with large deviations theory. This is explored in the next Chapter for Effective Bandwidth Algorithms. The Normal distribution is one of the most important distributions in probability theory. It is found from the strong law of large numbers and the Central Limit Theorem *[14][17][18]*. The Gaussian Approximation Algorithm uses this estimate to find the blocking probability for network traffic. It can be combined with the Chernoff Bound Algorithm for sharper estimates. The numerical evaluations in Chapter 6 examine these algorithms with the Effective Bandwidth Algorithm.

Next are algorithms concerned with Timescale Analysis with computations at cell, burst, and call level. Then there are algorithms using Baysian decision theory with the 'ON-OFF' model. There are algorithms focusing on the dynamics of the network traffic flow. They are quite a different approach as they use the arrivals distributions to estimate the CAC. They require large storage for implementation, but have many advantages such as flexibility and error estimation.

The areas of Artificial Intelligence, fuzzy logic and neural networks are represented, and finally there are Priority Algorithms and those based on simulations and Reinforced Learning (RL) techniques. The algorithms based on Effective Bandwidth are presented in the next Chapter, and then there are Measurement-Based algorithms in Chapter 5.

3.2 The Convolution Algorithm

The convolution algorithm *[9][23]-[29]* is a very accurate scheme for bufferless models. The connection admission control decision is based on the measure of Cell Loss Probability (CLP). The algorithm gives very accurate estimation, but there is a high cost in terms of accumulated calculations and storage for real-time implementation.

3.2.1 The Bufferless Fluid Flow Model

In the fluid flow model *[5][6][29]* the traffic sources are multiplexed together in a 'fluid flow'. The aggregate source rate is used to find an estimation of the CLP. The convolution algorithm uses the peak cell rate *max* and the average cell rate *avg* and burst duration as parameters. The bufferless fluid flow traffic model is suitable for estimations of bursty traffic, such as video sources. The sources have active and idle periods, known as 'ON-OFF' sources, see *Figure 2.1.* The sources have active periods when cells are generated at a constant rate *max*, the peak rate, with no cells generated in the idle period. The average rate *avg* is found from calculations.

For 'ON-OFF' sources (Section 2.3.1.1), under the bufferless fluid flow model, the probability that a connection is in an active or burst state is *avg/max*, the probability that it is idle is *1 - avg/max*. For *N* existing calls, let $f_i(x)$ represent the probability density function (pdf) of the traffic generated by call *i*:

$$f_i(x) = \begin{cases} \quad {avg_i}/{\max_i} & if \quad x = \max_i \\ 1 - \quad {avg_i}/{\max_i} & if \quad x = 0 \end{cases}$$

The density function of the aggregate traffic *[17][18]* generated by *N* existing calls, denoted by $q(x)$, is equal to the convolution of $f_1, f_2, \ldots f_N$:

$$q(x) = (f_1 * f_2 * \ldots * f_N)(x)$$

The computation cost becomes considerable as indicated by the above formula, so the algorithm does not fulfill the real-time requirement of the CAC function. Approximations that may be used to overcome this difficulty are described in the coming sections. In the bufferless fluid flow model, if the aggregate peak rate *R* is smaller than the link capacity, i.e. $R \leq C$ then the cell loss is assumed never to occur. There is a buffer with the *M/D/1* queuing model *[28][29]* that accommodates the short-term fluctuations caused by simultaneous cell arrivals from different connections. The buffer has a length of 100-200 cells so that it is small enough to prevent excess delay.

3.2.2 Cell Loss Probability Estimation

The bufferless fluid flow model means cells are discarded when the instantaneous total traffic load *R* exceeds the link capacity *C*. *R* is defined by a load with *n* active sources i.e. *n*max.* Cell Loss Probability found from the M/D/1 queuing model should be less than the quality estimate '*Virtual Cell Loss Probability'.* It is found with the fluid flow model that has a small buffer used to accommodate the minor fluctuations, *[9][24]-[29].*

Virtual Cell Loss Probability is the ratio of excess traffic and traffic load ρ *[24].* The *Virtual Cell Loss Probability (pv)* is defined where *N* denotes the number of sources multiplexed in the link:

$$pv = \frac{OF}{\rho}$$

$$OF = \sum_{(n.\max - C = 0)}^{n=N} p(n)(n.\max - C)$$

$$\rho = N.\,avg$$

where:

$$p(n) = \binom{N}{n}\left(\frac{avg}{\max}\right)^{n}\left(1-\frac{avg}{\max}\right)^{N-n}$$

with *p(n)* as the probability that *n* out of *N* sources are active

3.2.3 Enhancements to the Convolution Algorithm

A virtual bandwidth technique is described to replace the convolution in *[26]*, and a fast implementation for it with a 'real-time' computation algorithm. The computational algorithm is extended to obtain a close upper bound on cell loss probabilities. To reduce the calculations accumulated, a Multi-nominal Distribution Function (MDF) is described by a study in *[27].* The performance of the convolution approach is improved by application of the MDF to store groups of the same source types. It evaluates the complexity in terms of processor capacity and the memory required to do the calculations. The statistical multiplexing gain is found from the probability distribution density function of the individual sources.

The general state probabilities are evaluated by convolution of partial results obtained from groups of sources. The transmission rates already established at a given moment are found, with the probability that the sources will continue at those rates. These can be represented as vectors - a system status vector SV, and a source status vector SV_j, both having the same two fields representing the rate and probability.

To calculate the bandwidth requirements of the superposition of several sources, this approach is based on the convolution expression:

$$P(\,Y\; + \;New = b\;) \quad = \quad \sum_{k=0}^{b} P(Y = b - k)P(New = k)$$

where Y is the bandwidth requirement of the already established connections. New is the bandwidth requirement of a new connection, and b denotes the instantaneous required bandwidth. The convolution approach obtains a probability density function for the offered system load, expressed as the probability that all traffic sources together are emitting at a given rate. When the connection terminates, the state of the system must be updated.

With implementation, the bandwidth now occupied may be obtained by deconvolution. Other implementation approaches are the Fast Fourier Transform and the binary tree implementation *[29]*.

3.3 The Chernoff Bound Algorithm

The Chernoff Bound *[9][20]-[30]* is used as a measure of the limit of probability that is tolerated for the bandwidth to exceed link capacity. The notion of capacity of the network is measured for a given Quality of Service (QoS) guarantee. This is to allow for very small loss probability and to extract multiplexing gains from the statistical independence of the traffic processes.

Defining the Chernoff Bound, *[9][17][18]* let X_i be the bandwidth required by connection *i*, *C* is the link capacity and *exp(-γ)* is the given probability of overflow, the following inequality must hold:

$$\text{Prob}\{\sum_i X_i \geq C\} \leq \exp(-\gamma)$$

By definition, the moment generating function of a random variable X_i *[17][18]* is:

$$\phi_i(\text{s}) = \text{E}[\,\exp(\text{s}\,X_i)\,]$$

If the connections are independent, the Chernoff Bound allows us to write:

$$\text{Prob}\{\sum_i X_i \geq C\} \leq \exp\ \{\inf_s\,[\sum_i \ln\{\phi_i(s)\} - sC\,]\}$$

A connection is excepted if the right side of the inequality is less than *exp(-γ)*.

The algorithm seeks to find a minimum value for the expression in the square brackets in the above inequality. The moment generating functions of random variables that represent different users or sources need to be found. The calculations for the expression in square brackets are determined numerically as the number of traffic classes increases.

3.3.1 Statistical Multiplexing and the Chernoff Bound Algorithm

In *[21]*, let the i_{th} virtual circuit of class *j* be represented by $u_{ji}(t)$ denoting the utilised bandwidth. The $u_{ji}(t)$ is an 'ON-OFF' process, with values for the utilised bandwidth $e_{0,j}$ and *0* for 'ON' and 'OFF' respectively. Since we assume statistical independence of the traffic sources, the processes $u_{ji}(t)$ *(i=1,2,...K)* of the same source class have identical templates, and differ only in their phase, i.e.

$$u_{ji}(t) = u_j(t + \theta_{ji})$$

where $u_j(t)$ is a deterministic, periodic 'ON-OFF' function with period of T_j, where $\omega_j = \Pr(u_{ji} = e_{0,j})$ and $1 - \omega_j = \Pr(u_{ji} = 0)$, while the phases θ_{ji} are independent random variables uniformly distributed in the interval T_j.

The performance measure is the loss probability P_{loss}:

$$P_{loss} = \Pr(U>C)$$

where K_j is the number of virtual calls of class *j*, the total instantaneous load is:

$$U = \sum_{j=1}^{J} \sum_{i=1}^{Kj} u_{ji}$$

P_{loss} is the fraction of time that the aggregate demand for bandwidth from all the sources exceeds the total bandwidth, the Quality of Service requirement is:

$$P_{loss} \le L$$

where L is a small number, such as 10^{-6}.

3.3.2 The Chernoff Bound and Admissible Set

The estimation of P_{loss} is by the Chernoff Bound for this algorithm. The sources have been characterised by stationary random processes $u_{ij}(t)$. This denotes the utilised bandwidth of the virtual circuit for each source. It provides a simple single resource loss model from which calculations for adding additional sources to the overall capacity *C* may be estimated.

The instantaneous loads $u_{ij}(t)$ are independent, non-negative random variables (denoting the utilised bandwidth of the i_{th} virtual circuit of class *j*) with moment generating functions:

$$M_j(s) = E[\exp(s\,u_{ij})] = \int_0^{\infty} e^{sx} \partial W_j(x)$$

where:

$$W_j(x) = \Pr(u_{ij} \le x).$$

Chernoff's Bound *[17][18]* gives:

$$\log P_{loss} \le -F_{\mathbf{K}}(s^*)$$

where:

$$F_{\mathbf{K}}(s) = sC - \sum_{j=1}^{J} K_j \log M_j(s)$$

and:

$$F_{\mathbf{K}}(s^*) = \sup F_K(s) \qquad \text{for } s \ge 0.$$

If $C \to \infty$ and $K_j/C=O(1)$ then from the probabilities of large deviations for sums of independent random variables *[31]:*

$$\log P_{loss} = -F_{\mathbf{K}}(s^*)[1 + O(\log C / C)]$$

hence the asymptotic large deviations approximation is:

$$P_{loss} \approx \exp(-F_{\mathbf{K}} (s^*))$$

To avoid trivialities, the stability condition is assumed:

$$\sum_{j=1}^{J} K_j E(u_{ji}) < C$$

and:

$$\lim_{s \to \infty} \sum_{j=1}^{J} K_j \frac{M_j'(s)}{M_j(s)} > C$$

where the prime denotes a derivative.

The function $F_k'(s)$:

$$F_{\mathbf{K}}'(s) = C - \sum_{j=1}^{J} K_j \frac{M_j'(s)}{M_j(s)}$$

The function $F_{\mathbf{K}}(s)$ is a strictly concave function with a unique maximum at $s = s^*$, which is the positive root of the above equation $F_{\mathbf{K}}'(s) = 0$.

In the case of binomially distributed u_{ji}, where $\omega_j = \Pr(u_{ji} = e_{a,j})$ and $1-\omega_j = \Pr(u_{ji} = 0)$ then:

$$F_K(s) = sC - \sum_{j=1}^{J} K_j \log\{1-\omega_j + \omega_j \exp(se_{a,j})\}$$

and s^* is obtained by solving the equation:

$$\sum_{j=1}^{J} \frac{K_j \omega_j e_{0,j} \exp(s e_{0,j})}{1 - \omega_j + \omega_j \exp(s e_{0,j})} = C$$

In the single-class case, i.e. *J = 1* the resulting expressions give simple guides to the numerical evaluation simulations of the algorithm (see Chapter 6).

With $a = (C/e_0)/K$:

$$s^* = \frac{1}{e_0} \log \left[\frac{a}{1 - a} . \frac{1 - \omega}{\omega} \right]$$

$$F_K(s^*) = K\left[a \log\left(\frac{a}{\omega} \right) + (1 - a) \log\left(\frac{1 - a}{1 - \omega} \right) \right]$$

This expression is used to obtain $K_{\max}$ (or maximum number of sources) which is the value of *K* for:

$$F_K(s^*) = \log(1/L)$$

where *L* is the Quality of Service requirement representing Cell Loss Probability.

3.3.3 The Chernoff Bound and The Burstiness Parameter

In *[30]* there is a similar expression as that in the previous section found using *p*, a 'burstiness parameter'. The value for $1/p$ is the peak to mean ratio of the load produced by a source or call. The instantaneous load on the resource at time *t* is: $S_n(t) = X_1(t) + X_2(t) + ... + X_{n(t)}$ and is assumed to have a binomial distribution with the random variables $P\{X_i(t) = 1\} = p$ and $P\{X_i(t) = 0\} = 1 - p$.

The Chernoff Bound for a binomial random variable is:

$$P\{S_n > C\} = P\{S_n > na\} \leq \exp(-nK(a, p))$$

where: $$K(a,p) = a \log \frac{a}{p} + (1-a) \log \frac{1-a}{1-p}$$

and $a = C/n$.

The use of large deviation approximation based on the Chernoff Bound to estimate loss probability in *[21]* is combined with Effective Bandwidth estimation for admission control. This method uses Chernoff Bound for bufferless networks to analyse resources in buffered networks. The VBR traffic is modeled with 'ON-OFF' sources and a fluid model. The traffic is divided into two classes, one for which statistical multiplexing is effective and the other where it is not. For statistically multiplexed sources, Effective Bandwidth is found where there is an admissible set as defined by *[30]* (see Section 2.6 for explanations of statistical multiplexing). The main disadvantages are that the moment generating functions of the different sources are required. It can be difficult to determine the optimal values s^* to minimise the expression.

3.4 The Gaussian Approximation Algorithm

The aggregate traffic rate for a number of traffic sources is assumed to have a Gaussian distribution. The algorithm *[5][6][11][17]-[20][31][33]* relies on the Central Limit Theorem. This states that the aggregate traffic converges to a Gaussian distribution as the number of connections approaches infinity. It is not a conservative approach and may be too optimistic. Hence it may not be as accurate for bursty traffic. First in this section there is background theory to explain the algorithm. Then its behavior is described with the *M/D/1* model for the output buffer with Poisson sources and a mixture of source types for connections.

3.4.1 The M/D/1 Tail Distributions and Blocking Probability

The algorithm in *[20]* uses the Gaussian Approximation in its estimation of blocking probability at burst level for heterogeneous traffic. For the offered traffic

W(t) and carried traffic $W'(t)$, the difference between offered and carried traffic is the blocked traffic. This indicates the possible losses and need to be within acceptable limits. By looking at the output queue length distribution given by the *M/D/1* formula, we can use the Central Limit Theorem *[14][18]* and large deviations theory to approximate the tail probability $P(\,W(t) > C)$, where *C* is the link capacity.

First, the tail distribution of offered traffic for Poisson traffic is found. The distribution for offered traffic *W* is related to the carried traffic W'. Using the moment generating function for *W*, large deviation theory is applied to obtain good approximations for the tail distribution of *W*. Then there is the computation of *W* with mixed Poisson traffic and continuous varying traffic such as compressed video. It is computed using numerical methods for fast evaluation of congestion for mixed traffic types.

The log moment generating function of the tail distribution: $q(x) = p(\,W(t) = \omega)$ gives the mean and variance of *W(t)*. It can then be substituted into the Gaussian Approximation formula to find the distribution density $p(W = \omega)$:

$$p(W = \omega) \approx \frac{1}{\sqrt{2\pi Var\,(W\,)}}\,e^{-(\omega - E(W\,))^2/2Var\,(W\,)}$$

This approximation is used to find the estimation of blocking probability at burst level for heterogeneous traffic. The blocked traffic needs to be within the Quality of Service requirements for loss probability *[7][8]*. The output queue length distribution is given by the *M/D/1* formula. The blocking probabilities are found by relating the degree of queue saturation with service and arrival rates, and using the steady-state equations.

Let *K* be a set of calls (or sources) assigned to the link with total bandwidth *C*. The offered load is:

$$W(t) = \sum_k R_k(t)$$

where the arrival process for call k is $R_k(t)$ and the offered load is the sum of the instantaneous bit rates.

The following will find the mean and variance of *W(t),* the offered traffic or load. The offered traffic has the following recursive relation for computing the tail distribution, taking the expectation of x over interval *[0,y]:*

$$Q(y) = P(W(t) < y) \qquad \textit{for offered traffic } W(t)$$

For Poisson traffic *W(t)* is modeled by jumps of different amplitudes $a_i > 0$, which arrive at Poisson rate γ_i and last for duration b_i. This random duration can be represented with an associated λ.

$$\int_0^y x\,\partial Q(x) = \sum_i \gamma_i a_i b_i Q(y - a_i)$$

Differentiating with respect to y gives the marginal distribution *q(y):*

$$yq(y) = \sum_i \gamma_i a_i b_i q(y - a_i)$$

This is called a Poisson shot noise process *[14][18].* To improve the efficiency of computing *P(W(t)* $\leq$ *x)*, which is too large to be practical, large deviations theory is used to find the tail distribution of the Poisson shot noise process *W* in the next section. The blocking probability for the lossy system is obtained by the following relationship between *W* and W'. For $x \leq C$, the total bandwidth is:

$$P(W'(t) \leq x) = \frac{1}{1 - P(W(t) > C)} P(W(t) \leq x)$$

The blocking probability from the relationship between *W* and *W'* is:

$$P(W^{/}(t) > C - a_i) = \frac{1}{1 - P(W(t) > C)} P(W(t) > C - a_i) - P(W(t) > C)$$

To find the characteristic functions of *W*, the log moment generating function of *q(ω) = p(W(t) = ω)* for the Poisson shot noise process, is defined as:

$$\mu_W(s) = \log_E \Psi_W(s) = \log_e \int_0^\infty q(\omega) e^{s\omega} d\omega$$

Using the marginal distribution *q(y):*

$$yq(y) = \sum_i \gamma_i a_i b_i q(y - a_i)$$

then: $$\mu_W(s) = \sum_i \gamma_i b_i (e^{sa_i} - 1)$$

The mean and variance of *W(t)* are obtained by differentiating $\Psi_\omega(s)$:

$$E(W(t)) = \Psi_W^{/}(0) = \sum_i \gamma_i a_i b_i$$

$$Var(W(t)) = \Psi_W^{//}(0) - \Psi_W^{/2}(0) = \sum_i \gamma_i a_i^2 b_i$$

3.4.2 Applying Large Deviations Approximations

Having found the mean and variance of *W, p(W(t))* can be computed by the Gaussian Approximation *[20][31]:*

$$p(W = \omega) \approx \frac{1}{\sqrt{2\pi Var(W)}} e^{-(\omega - E(W))^2 / 2Var(W)}$$

This approximation is not very accurate at ω more than the standard deviation from the mean. To develop sharper estimates, the Chernoff Bound *[18][31]* can be applied:

$$P(W(t) > y) \le e^{-(s^* y - \mu_W(s^*))}$$

where s^* satisfies the equation for the first derivative of $\mu_w(s)$:

$$y = \mu'_W(s) = \sum_i \gamma_i a_i b_i e^{sa_i}$$

The above bound can be sharpened by the theory of large deviations *[11][19][30][31]*, which is concerned with the sum of a large number of random variables. The new result is improved by a factor of $1/s^*\sqrt{2\pi\mu''_W(s^*)}$ such that:

$$P(\omega < Y) \approx F(s^* \sqrt{2\pi\mu''_W(s^*)})e^{-(s^* y - \mu w(s^*))}$$

3.4.3 A Mixture of Heterogeneous Traffic Sources

To evaluate a mixture of Poisson and non-Poisson traffic, there are the estimates for $P(W > y)$ derived in the last section such as:

$$P(W < Y) \approx F(s^* \sqrt{2\pi\mu''_W(s^*)})e^{-(s^* y - \mu_W(s^*))}$$

these remain true for other $R_k(t)$ such as VBR or compressed video sources. Suppose there is the steady-state probability p_i for $R_k(t)=a_{i.}$. Thus $\mu_k(s)$ for call k is given by:

$$\mu_k(s) = \log_e \sum_i p_i e^{sa_i}$$

the log moment generating function is given by:

$$\mu_k(s) = \sum_k \mu_k(s)$$

With a mixture of Poisson and variable rate traffic, there is a mix of log moment generating functions for $\mu_k(s)$. To compute in real-time the following approximation is used:

$$P(w < Y) \approx F(s * \sqrt{2\pi \, \mu_W^{//} (s*)}) \, e^{-(s* y - \mu \varpi \, (s*))}$$

where: $\mu_{W'}(s^*) = y.$

Expanding the individual $\mu_W(s)$ by Taylor's series can be computed for each call type. The series expansion of $\mu_k(s)$ $=\Sigma_k(s)$ $\mu_k(s)$ is given by first expanding $\Sigma_i \, p_i e^{sa_i}$ as a series, then expanding the log of the resulting series:

$$\mu_k(s) = \log_e(1 + d_1 s + d_2 s^2 + \ldots)$$

$$= c_1 s + c_2 s^2 + c_3{}^3 + \ldots$$

in which: $$c_i = d_i - \frac{1}{i} \sum_{j=1}^{i-1} j \, d_{i-j} - c_j$$

With these pre-computed coefficients, it is easy to obtain the series expansion of $\mu_W(s)$ as well as its first two derivatives in real-time.

3.4.4 Algorithm Implementation

The log moment generating function of the tail distribution: $q(x) = p(\, W(t) = \omega)$ gives the mean and variance of $W(t)$. They can then be substituted into the Gaussian approximation formula to find the distribution density $p(W = \omega)$:

$$p(W = \omega) \approx \frac{1}{\sqrt{2\pi Var(W)}} e^{-(\omega - E(W))^2/2Var(W)}$$

3.5 Algorithms with Timescale Analysis

Mitre, Reiman, and Wang *[32]* combine cell and call level for dynamic admission control to obtain efficient resource sharing. The model is a single bufferless link with multiple call classes, each source behaves as an 'ON-OFF' fluid source while in the system. The optimisation problem is that given a maximum cell loss, a CAC is designed to maximise the revenue due to carried traffic. This problem is too computationally intensive, so timescale decomposition is used to simplify it. The reduced state optimisation problem is then numerically feasible.

An admission control algorithm for the combination of different types of traffic was presented by Hui *[20]*. It is called the multilayer bandwidth allocation algorithm. This is one of the earliest and most important papers to establish the ideas of timescale. A CAC algorithm for heterogeneous source types providing different services is designed by analysis of traffic with different characteristics. The evaluation of congestion occurs at different timescale levels (See Chapter 2), packet or cell level, burst level and call level. The acceptable bounds are chosen based on the blocking probabilities at each level.

The multilayer bandwidth allocation scheme allows a call to join a group forming a trunk. The admissible region is calculated as the probability of call blocking, depending on the call arrival and holding times. The multilayer refers to the computations of probabilities at packet or cell level, burst level and call level. For call k, the packet arrivals process at the switch input is $R_{l,k}(t)=u$, the channel rate at time t.

Each level l chooses a subset of the level above to allocate resources, if it does not cause blocking of the level below. Thus the burst level allocation of a call checks within the call bandwidths to see if more resources can be allocated within the trunk of calls. Allocation also depends on if the packets in the level below will not congest. The algorithm in *[20]* defines a request packet which checks to see if the bandwidth request can be met for a call or burst. The summation of allocated resources is computed so that the connection admission will cause the over-allocation of resources.

The offered traffic at level *l* to the output of resource Γ_l at time *t* is the sum of all the traffic of all sources *k*, which is $W_l(t) = \sum R_{l,k}(t)$. The carried traffic is $W_l^{'}(t)$ as the loading of output resources, so the difference between offered and carried traffic is the blocked traffic. This needs to be within acceptable limits. The blocking probabilities at cell, burst and call levels can be found from relating the degree of queue saturation with service and arrival rates and then using the steady-state equations.

The traffic model is a two state fluid flow model, with the data source behavior is described by: *idle state-> tx at 0 bit rate->burst state->tx at peak rate*. Hence the peak rate and distributions of burst and idle periods completely describe the traffic statistics of the connection, represented by R_{peak}, σ the utilisation factor (the fraction of time the source is active), and *b*, the mean of the burst period. The source metric vector is (R_{peak}, σ, b). The model can be extended to non-exponential burst/idle periods by the standard moment matching approximations in Section V-A of *[20]*. The admissible call region for a class of traffic defined as an *n* dimensional space of f_i where the burst blocking probability is acceptably small. It was found to be a concave region with the boundary becoming more linear as the trunk capacity increases.

3.6 A Decision-Theoretic Approach Algorithm

This CAC algorithm is based on Bayesian decision theory where the acceptance of a connection is if the current load is less than a pre-calculated threshold *[30]*. This methodology allows for explicit treatment of the trade-off between cell loss and call rejection. It also allows for the consequences of estimation error. The use of timescale analysis as described by Hui *[20]* in the previous section is used as a basis, to look at the call level, burst level and cell level congestion problems. A separation of timescales provides the framework for analysis, buffering is assumed to allow for cell delay variation. A bufferless model is used at burst level.

An offered call is accepted based on a simple threshold value. The threshold implements a robust estimation procedure, where the decision-theoretic framework facilitates the trade-off between the benefits of accepting the call (earned revenue, customer satisfaction) and the drawbacks (inability to reach QoS targets). The use of Bayesian theory at burst level allows the Quality of Service requirements of a source to be met. The model used is the basic 'ON-OFF' model with an unbuffered capacity *C*. The call loss probabilities are estimated first. Assuming a prior distribution for burstiness parameter is available, different choices of this distribution give different amounts of uncertainty. This uncertainty is combined with additional information from measurements of load. They are integrated by Bayesian formulations to trade off between utilisation and cell loss.

The scheme in *[30]* is also extended for multiple call types, and a call need only specify its peak rate and Cell Delay Variation (CDV) tolerance. It is found *[30]* that Bayesian decision theory "provides a coherent and general framework within which the several trade-offs involved may be effected".

3.7 A Dynamic CAC Based on the Arrivals Distribution

When the number of classes of calls is large it can mean a variety of QoS requirements need to be met. To help this process the algorithm in *[25]* for a dynamic CAC uses the distribution of the number of cells arriving during a fixed interval. The call acceptance is based on the online evaluation of the upper bound of cell loss probability. The call acceptance is derived from this distribution and from the traffic parameters provided by the source at connection setup.

Other CAC algorithms for a wide range of classes of call require a large storage table for traffic parameter values and analysis of QoS performance. The table of values may be based on simulation or analysis. A particular arrival process is assumed, such as an interrupted Poisson process, with or without output buffers. The advantage of the dynamic CAC approach in *[25]* is that it is independent of the classification of calls and arrival process modeling. It also tolerates policing errors using the cell flow measurement. It concentrates on the Cell Loss Probability as buffer-sizing dimensioning is used to satisfy the delay requirement. So a new connection is admitted if it is less than the upper bound on Cell Loss Probability from the distribution of the arriving cells, as estimated using a formula in Section II of *[25]*. The implementation of the algorithm uses an estimated load state vector to represent the probability distributions. Numerical examples are given to demonstrate the use of different types of traffic such as voice and video. This algorithm presents an interesting idea of measurement, a completely different approach to admission control that is developed further in Chapter 5.

3.8 Algorithms using Neural Networks, Fuzzy Logic and Artificial Intelligence Techniques

Neural networks and fuzzy logic have been proposed *[46]-[56]* as a basis for connection admission control. They attempt to predict the statistical behavior of the multiplexed sources. From this prediction they are able to forecast the cell loss rate. The decision to accept or reject the incoming connection can be made based on the accumulated intelligence by the neural network. The disadvantage is that the techniques may not be not fast enough to deal with traffic in real-time. Schemes to integrate various traffic controlling functionalities such as link capacity allocation, flow routing and network management can be achieved by a distributed system of neural networks and intelligence in the network.

3.8.1 Introduction to Artificial Neural Networks

Artificial neural network systems, or neural networks *[46][47]*, are physical cellular systems that can acquire, store and utilise experimental knowledge. The knowledge is in the form of stable states or mappings embedded in networks that can be recalled in response to the presentation of cues. The basic processing elements of neural networks are called artificial neurons or nodes. Neurons perform as summing or non-linear mapping functions. They can also be perceived as threshold units that fire when their total input exceeds a certain bias level. Neurons usually operate in parallel and are configured in regular architectures. They are often organised in layers, and feedback connections may exist within the layer and towards adjacent layers. Each connection strength is expressed by a numerical value called a weight, which can be modified.

Neural networks can be distinguished by their architecture *[46]* and their learning modes. They have the unique ability to be taught or trained, and learn new associations, patterns and functional dependencies. Learning corresponds to parameter changes, and in this neural networks seem to differ from the programming of a more traditional machine. Instead they select the best architecture, specify characteristics of the neurons and initial weights and chose the training mode of the network. Appropriate inputs are then applied to the network so that it can acquire knowledge for the environment. The knowledge is assimilated and can be recalled later by the user.

Fundamental concepts and models of artificial neural networks are based on their biological counterpart the human neuron consisting of a linking mechanism via synapses. Neurons are linked together in a variety of groupings, depending on functionality and may have layered architecture and feedback mechanisms. Models of neural networks are defined in terms of their inter-connections. Neurons are connected through weights allowing a variety of sequences of delay or lag factors. The elementary feedback network has input and output neurons represented by vectors and connected by weights, which denote the source and destination nodes respectively. The processing done by the network is a non-linear mapping of input to output influenced by the values of the weight, this type of network has no feedback connections. A feedback network is achieved by connecting neuron outputs to their inputs to enable control of the output with a suitable time lag.

Another important concept is that of neural processing. The process of computation of a given output performed by the network for a given input is known as recall. Recall is to retrieve information stored as a content of the node. We can assume the network stores a set of patterns, and the input associated with the pattern is a process called auto-association. Classification is another form of neural computation, where a set of input patterns is divided into classes or categories. The classifier responds to an input pattern, and recalls information regarding the categorisation.

3.8.2 Integrated ATM Traffic Control using Neural Networks

In *[49]* the integration of link capacity control and call admission control is achieved via a distribution of neural networks. This system is particularly effective for multimedia call services with unknown traffic characteristics. An adaptive control method using neural networks is proposed that learns the relation between offered traffic and service quality. Non-linear functions for link capacity and their assignment are optimised with the integration of adaptable neural networks for connection admission control.

A three-layered neural network is able to approximate the shape of an arbitrary non-linear function by precisely adjusting connection strengths, called weights, between neurons *[46][47]*. An algorithm, with back-propagation according to a set of correct input and output data, does the adjustment of the weights from the target function. The three layers of the neural network are an input layer, a hidden layer and an output layer. Each layer consists of a group of neurons, and the output of a neuron in one layer is the input in the next layer. In the operating phase, the user sets the values of the input neurons, and the network produces output values. In the training phase, the user simultaneously sets desired input and output values, and then the weight values are modified according to the following learning equation:

$$w_i(t+1) = w_i(t) - c[y_t - f(x_t)]\frac{\partial f(x_t)}{\partial w_i}$$

where $w_i(t)$ is one of the weights in the cycle *t*, $f(x_t)$ is the neural network for the output for input x_t and y_t is the corresponding desired output, with *c* a positive constant called the learning constant. In online training a pattern table is used in combination with back-propagation. The pattern table contains a number of observed values from running systems. These are then randomly selected from the pattern table to be used as input and output value pairs during the training phase. The diagrammatic representation of this neural network is in *Figure 3.1*, the call input is *a*, the corresponding weighting factor is *w* and the output is *q*.

Another example is found in *[51]* which uses a back-propagation feedforward neural network. It partitions the bandwidth among a set of users and approximates the admission control for each user. The output link bandwidth is dynamically assigned between isochronous (guaranteed bandwidth) and asynchronous traffic types. Investigation of the use of neural networks and stochastic approximation algorithms for admission control and bandwidth allocation is done for a hybrid multiplexer serving multiple users with different traffic types. The neural network controller is for a two-level hierarchical system where bandwidth is allocated among a number of user sites that independently perform admission control.

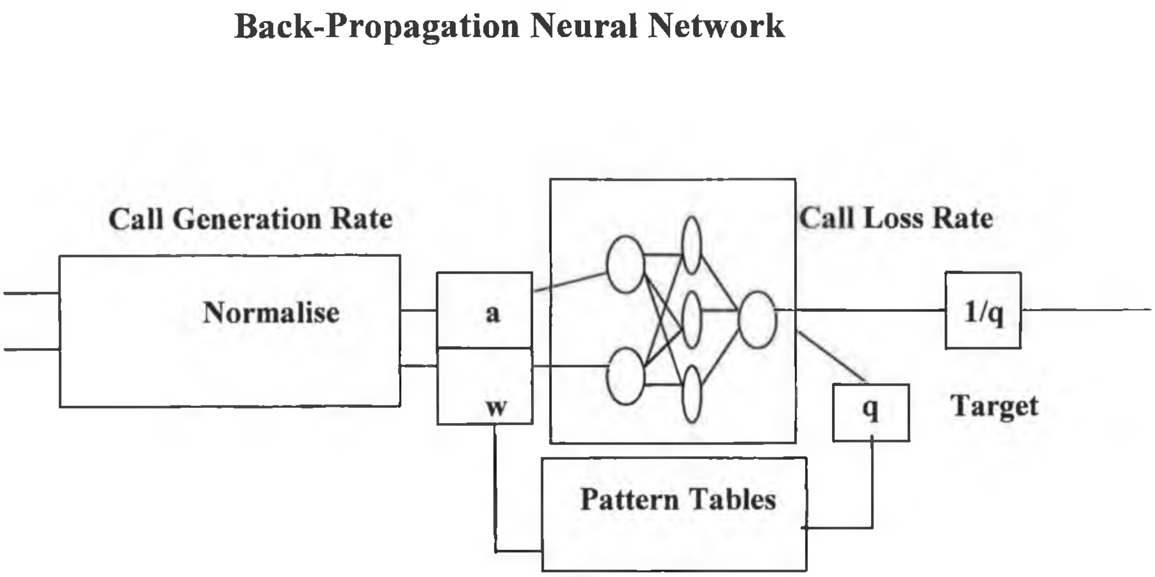

Figure 3.1 Neural Network for Call Loss Rate Estimation

3.8.3 Training Strategy

In a distributed system consisting of a number of neural networks *[49]*, each neural network is trained independently. The networks are then trained simultaneously to shorten the length of time this requires if they were to be trained separately. The initial weights are important as they determine when the training period is likely to converge. The best weight values can't be known prior to installation, so first random weights are used for the initial period of off-line training *[49]*. The converged weights from off-line training are used as the initial weights for the online training. The network is then trained to control the real target system, and the weights are gradually improved to achieve more efficient control.

The neural networks in *[48]* uses the ability to model parameters at the synaptic level rather than threshold level. The 'pRAM' neural network learns to approximate a real-valued function from a given set of training patterns and their corresponding desired outputs. The output is accumulated and a memory update rule uses a reinforcement technique to generate rewards and penalties. To improve on the real-time application requirement, the training rate for the neural network is adjusted not by dependency on output error, but on the values of the input variables.

3.8.4 Integrated Call Admission Control Using Neural Networks

Call admission control and link capacity assignment are integrated in *[49]* to provide an efficient control system, with greater potential for optimisation. Neural networks decide to accept or reject a call setup request for each output link. The neural network for link capacity control learns the results of call admission and decides the optimum link capacity assignment. The neural networks co-operate to learn and so improve overall network performance.

Call admission control decides whether to accept or reject a setup request according to declared traffic characteristics and the required Quality of Service. When a node receives a call setup request, it categorises the call into bit-rate class according to cell emission characteristic parameters, to satisfy the following condition:

$$Q(n_1,\dots n_i,\dots n_K\,;v) \le Q_{req}$$

where n_i represents the connected call of bit-rate call i *(i=1,...K), K* is the number of bit-rate classes, and v denotes the capacity of the output link. Q is the service quality estimation function, and Q_{req} indicates the required service quality. The initial network design determined the call admission according to maximum bit rate. Then online training of neural networks improves the call admission boundary. The neural network call loss estimation adapts to the changes in this boundary. With 'adaptive' control the call loss rate is much smaller and near constant for all traffic conditions. This is in contrast to non-adaptive control using a fixed neural network, with a larger call loss rate and changes in offered traffic.

3.8.5 Call Admission Boundaries

The call admission boundaries derived by a neural network *[49]* are the boundaries between acceptance and rejection. The neural network finds this boundary from the data observed from the operating network. The example given is for two classes, each representing a different bit rate, with maximum bit rates v_m of $v_{m1} = 10, v_{m2} = 20$ and average bit rates v_a of $v_{a1} = 2, v_{a2} = 1$ respectively. These are given by the source traffic characteristics, and the service quality parameter is the cell loss rate. These sources are called class 1 and class 2, with the weights initially set to random values then the neural network is trained for 10, 000 seconds. The value of v is varied from 500 to 1000 according to the cosine function to simulate burstiness in the traffic. *Figure 3.2* gives an illustration of the cell admission boundary as the link capacity is trained by the network, and demonstrates the effectiveness of the integration with connection admission control.

3.8.6 A Decision Hyperplane Using Neural Networks

Neural networks have a self-learning capability, which can be utilised to characterise the relationship between input traffic and the system performance. The neural network in *[12]* uses a power-spectral-density *[5]* to contain the correlation behavior of the input process and uses it to evaluate system performance. Under the Quality of Service constraint, a decision hyperplane is constructed for connection admission control, according to the parameters of the power spectrum. The learning capabilities of the neural network adjust the optimum location of the boundary between these two decision spaces.

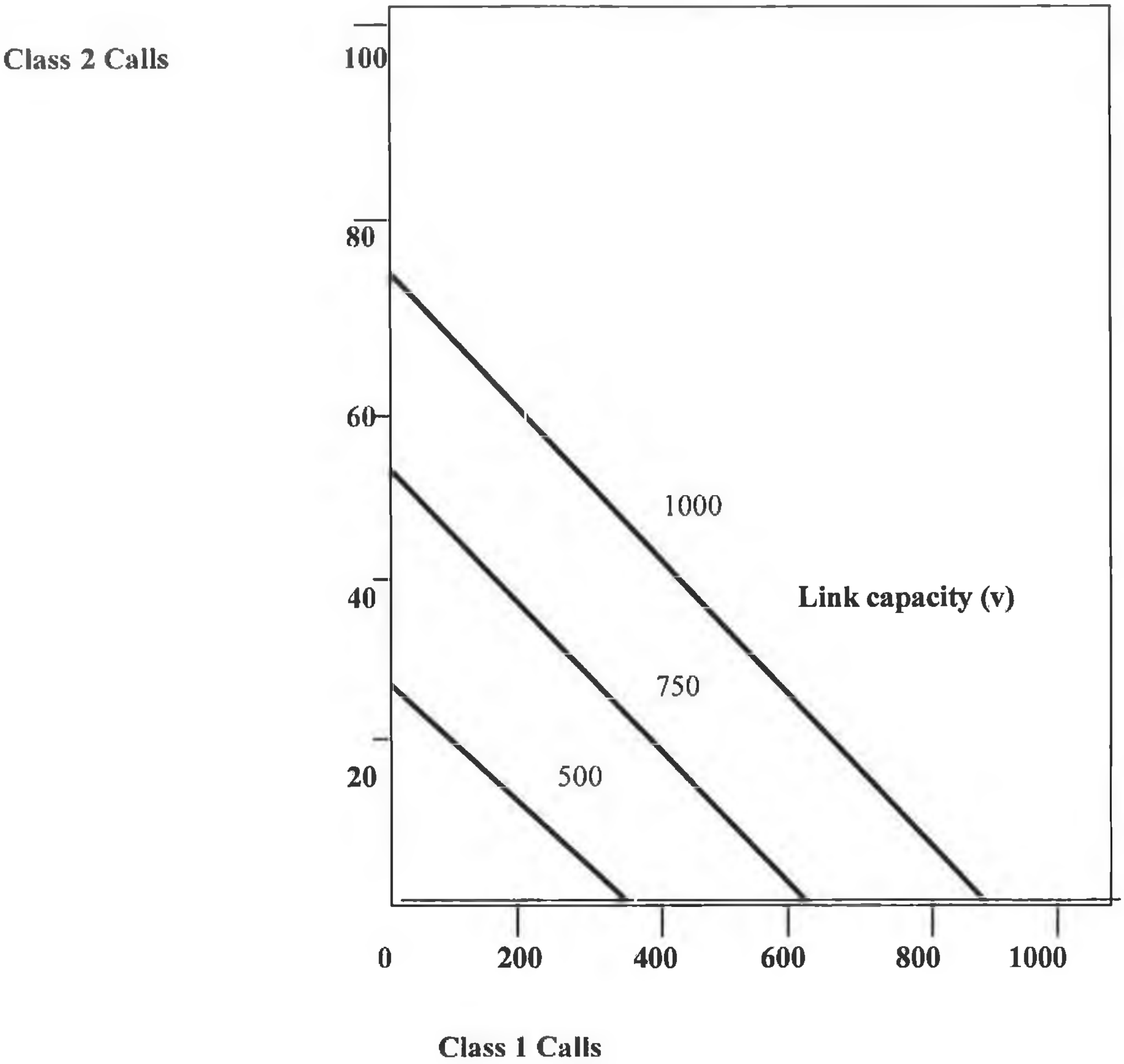

Figure 3.2 Call Admission Boundary derived by a Neural Network

The study in *[12]* looks at the performance for the frequency domain of the input traffic in comparison to many approaches with time-domain analysis. The power-spectral-density in the frequency domain is the Fourier Transform of the auto-correlation function *[5]* of the input process, capturing the correlation and burstiness features of the input process in the time-domain. The decision hyperplane uses the constraint of Quality of Service for its construction according to the parameters of the power spectrum. The sample space is split into two, one for 'accept' and one for 'reject'. When a new call is connected it is admitted to the 'accept' sample space.

3.8.7 Fuzzy Logic and Connection Admission Control

In *[56]* a fuzzy inference method is proposed in order to effectively estimate the probability distribution of CLR from its observed data. The method used is based on a weighted average of fuzzy sets. Fuzzy rules for the fuzzy inference are tuned automatically by a learning algorithm, energy functions are considered for this algorithm. A dynamic energy function is proposed, and the upper bound of the allowed Cell Loss Ratio (CLR) can be estimated. The fuzzy inference method based on a weighted average of fuzzy sets is proposed rather than conventional fuzzy inference, which is found to estimate an excessively high CLR. The estimation scheme is provided with a learning mechanism, the fuzzy rules are adjusted automatically by a learning algorithm with the observed data. The possibility distribution of the CLR is inferred from these fuzzy rules.

The relationship between CLR and the CAC algorithm is often non-linear, and the average learning provides an average of dispersion of maximum values. The estimation of the probability distribution of the CLR is needed to guarantee the allowed CLR for the CAC algorithm. The fuzzy inference approach has the 'then-part' of each fuzzy rule that gives the probability distribution of CLR. This is the distribution for the number of connections covered by the 'if-part' of the fuzzy rule. The transmission rate is classified into a number of classes, which also means that other parameters such as burstiness are taken into account. The fuzzy sets in each fuzzy rule are automatically extracted and tuned by a learning algorithm. Finally there is also a real-time compensation for CLR estimation errors to improve accuracy.

Fuzzy logic can be used in combination with other approaches. The CAC algorithm in *[58]* computes the Equivalent Bandwidth required to support each class of connections dynamically. It is based on online traffic statistics, declared traffic parameters and a fuzzy logic controller. Gaussian and diffusion approximations are used to characterise the aggregate traffic stream. Fuzzy logic control combines the model and measurement results to estimate the Equivalent Bandwidth in real time. It is shown that system utilisation is improved by the tuning of the fuzzy logic controller to combine the traffic characteristics deduced from the parameters and traffic measurements.

3.8.8 Multiple Quality of Service Requirements and Connection Admission Control with a Neurocomputing Controller

The papers *[55][56]* use neural fuzzy logic, proposing a neurocomputing call admission control algorithm to calculate the bandwidth requirements of multimedia traffic with multiple Quality of Service requirements. The algorithm uses a neural network and the online measurements of traffic rather than traffic parameters for estimations. The controller is a hierarchical structure of small size parallel neural network units. Each unit is a feedforward back-propagation neural network that has been trained to learn the complex non-linear function relating the different traffic patterns and Quality of Service. The controller allows for different classes of traffic with different Quality of Service requirements. The units can then be trained for different traffic classes for a specific traffic pattern, hence simplifying the design. The use of online traffic data allows for a swifter response to traffic congestion. Results show an improvement in accuracy of estimation over conventional methods based on mathematical or simulation analysis.

3.9 Algorithms with Prioritised Traffic Types

By classifying service types according to different priorities, it is possible to integrate a variety of services with different Quality of Service requirements for cell loss and delay. There have been a number of interesting studies addressing this topic *[56][59]-[61]*. More recent work in *[10]* combines this approach with Measurement-Based algorithms, allowing for a more sophisticated resource allocation scheme.

3.9.1 A Congestion Control Framework for Priority Traffic

A congestion control framework proposed in *[56]* describes an ‘express’ service for real-time traffic with bandwidth allocation at peak rate, and another class called ‘first class’ which has a guaranteed rate less than peak rate for allocation when congestion occurs. Thus statistical multiplexing is only used in the non-real-time traffic allocation. The integration of services in this way means that QoS performance requirements of Cell Loss Probability and end-to-end cell delay for both types of services can be met.

The CAC reserves bandwidth for an incoming call according to either peak rate for express services or a congestion parameter γ(between 0 and 1) for guaranteed bit rate < peak rate. The call is accepted if:

$$\frac{\sum M}{W} \leq p$$

$M =$ *total bandwidth reserved for the local access network*

$p =$ *the allowed utilisation level for Cell Loss Probability*

W = transmission capacity of the link

A connection can send *hW/f cells/sec*, where *h* is the cells in a logical frame length *f*. Congestion control occurs by a buffering mechanism in the router, which has a buffer for non-real-time first class service traffic. If it reaches a threshold a congestion indicator cell is sent back to the source to throttle the transmission back to guaranteed rates.

The disadvantage of this framework and CAC algorithm is that it requires extra hardware implementation at the sources and multiplexers to respond to the congestion indicator cells. It does not use statistical multiplexing fully and hence does not attempt to achieve maximum network utilisation. It uses the discrete-time Markov chain traffic model to derive Cell Loss Probability and cell delays, which may not be the most suitable model for bursty traffic. It does provide an overall framework for all types of traffic, and a system for calculation of buffer sizing and link utilisation, with a CAC algorithm that is simple and practical.

3.9.2 New Models for Admission Control of Priority Traffic

A study of M/D/1 queuing models in *[59]* produces approximations for Cell Loss Probability, the admissible load and buffer length. It can be used for expressions in traffic for both time and space priority cells. The analysis focuses on 'express rate' or priority cells and provides partial buffer sharing for both types of traffic. Time priorities are assigned to the cells, and approximate the effect of high priority cells by the use of random interrupts on the queue to give a new formula for CLP.

The use of separate buffers for priority traffic is proposed in *[60]* and can be contrasted with a shared buffer scheme in *[61]*. Both providing highly effective solutions to Multi-class QoS services with different levels of priority. The study in *[60]* provided separate CAC algorithms for each queue type, with Measurement-Based admission control for the lower priority traffic or best-effort service. The study in *[61]* points out that a cell scheduler is required to allocate separate queues, and proposes a shared queue system instead. The queue space is divided into multiple subspaces and is allocated to different classes depending on traffic levels with a class acceptance function. Their work refers to the Effective Bandwidth vectors found by Elwalid and Mitre *[34]* described in the next Chapter.

3.10 Algorithms Based on Simulation and Reinforcement Learning

Simulation and Reinforcement Learning algorithm have their bandwidth assignment for each class of service based on simulation results *[62]-[64]*. The sources with different services are grouped according to traffic descriptors, and the bandwidth assigned is derived from the mean of each. The second step is to consider the traffic with a mix of several classes. The assigned bandwidth is then found using previous simulations. So the performance measures for heterogeneous traffic are evaluated using the results obtained for homogenous traffic. The CAC algorithm uses the simulation results to have a set of values in order to decide if it can accept a call, in each individual class.

The evaluation of QoS performance (e.g. burst-level blocking probability) can be found from simulation results. It is confirmed with analyses, using traffic parameters such as peak rate (PCR) and average rate (ACR), burstiness (PCR/ACR) and average durations of bursts. Assumptions made such as a particular arrival process (for example an interrupted Poisson process, with or without buffers) need to be considered regarding suitability when representing the type of traffic controlled by the CAC algorithm.

In *[63][64]* the CAC policies are derived from solutions to Neuro-Dynamic programming. This is a simulation-based approximate dynamic programming methodology for producing near optimal solutions for large-scale dynamic programming problems. Neuro-Dynamic programming is also called Reinforcement Learning (RL). In *[63]* the CAC problem is naturally formulated as an average reward dynamic programming problem with a very large state space. So the CAC policy is essentially a problem of revenue maximisation. The computational requirements may be too slow for online use however, unless a smaller set of tunable parameters is used.

In *[64]* Reinforcement Learning (RL) is used to solve an adaptive admission control problem. The network revenue is to be maximised while meeting the Quality of Service constraints. This is formulated as a semi-Markov decision process with RL providing the solution. RL is better than model-based algorithms as it does not require explicit state transition models. These have such a large number of states that the algorithms become infeasible. The network accepts or rejects the call depending on a description given in terms of bandwidth as a function of time. The network measures QoS metrics. An example is the fraction of time that the total bandwidth exceeds the network bandwidth, called the capacity constraint. Another QoS metric is the call-level blocking probability. When offered traffic needs to be reduced to meet the capacity constraint, it is done according to a fairness constraint. The revenue is maximised subject to these QoS constraints.

The RL methodology in *[64]* means learning the optimal policy using a 'Q-learning' algorithm. This means that when a call arrives the Q-value of accepting the call and the Q-value of rejecting the call is determined. If rejection has the higher value, the call is rejected, otherwise if acceptance has the higher value, the call is accepted. The Q-value is learned from a value function that is updated when there is a transition from one state to another, due to an action in a particular length of time for a stepwise learning rate. Q-learning does not require explicit state transition models and the initial values can be arbitrary. The capacity and fairness QoS constraints are incorporated into the RL solution to maximise revenue.

3.11 Summary

The Chapter has presented a wide range of algorithms based on a variety of ideas. The algorithms can be catagorised as those based on a mathematical approximation such as the Convolution Algorithm, the Chernoff Bound Algorithm and the Gaussian Approximation Algorithm. They use Fluid Flow Analysis, Probability theory and the Central Limit Theorem to formulate the basis for an admission algorithm. They can be used in combination with large deviations theory and with other algorithms to provide sharper estimates.

As a different approach, there are algorithms with admission control using timescale decomposition. This means the optimisation problem of admission control is simplified and so the algorithm becomes easier to implement. Baysian decision theory is another basis for algorithms. It provides a pre-calculated threshold as shown by the Baysian formulations, these formulations trade off between utilisation and cell loss. A dynamic CAC based on the arrivals distribution acts as an introduction to the ideas of Measurement-Based algorithms. Then there are the Priority algorithms, those with Artificial Intelligence, and finally algorithms based on Reinforcement Learning. These represent areas of further research as they prove to be highly adaptable forms of admission control.

Chapter 4

Effective Bandwidth Algorithms

4.1 Introduction

The theory of large deviations *[31]* provides a unified basis for statistical mechanics, information theory and queuing theory. The theory of Effective Bandwidth is developed from this. The Effective Bandwidth of a source is the minimum amount of bandwidth required to satisfy its QoS constraint. Chang and Thomas *[35]* develop the theory from the laws of thermodynamics and the entropy function. The source is compared to a constant rate fluid, with a tail distribution of the queue length in the network. The theory of large deviations finds that the probability density function may be used to derive the 'energy' and 'entropy' functions of the source. By solving for the dominant exponent in its integral, an approximation of queue length distribution can be made. This corresponds to finding the minimum action path in classical mechanics.

The theory of Effective Bandwidth is extended to yield approximations for a network of local nodes and sources. This is achieved by close examination of how buffers build up. The approach by Gibbs *[31][35]* in statistical mechanics provides a solution. By specifying when the average energy the distribution of the coordinates from a uniform distribution to the Boltzmann distribution may be found. Similarly we look for the most likely distribution of a source given that the buffer builds up. Section VI in *[35]* establishes a connection between the entropy function and the relative entropy rate (the Kullback-Leibler distance) defined in Information Theory.

This Chapter explains the important concepts of Effective Bandwidth and equivalent capacity *[11][21][33]-[44]*. They are used as a basis for several admission control algorithms. The various traffic models and source characterisations are examined. The algorithms have been found to be very efficient in comparison to other types in terms of network resource allocation.

4.2 Defining Effective Bandwidth

The *'Effective Bandwidth'* allocation needs to meet Quality of Service requirements for connections while targeting good link utilisation. It is less than the bandwidth required for the peak rate of the source. This is possible since the likelihood that all sources will transmit simultaneously at peak rate is very low. Effective Bandwidth theory allows for the derivation of bandwidth allocation techniques for connection admission control from the behavior of individual and aggregate sources.

The concept of Effective Bandwidth is used to describe the utilisation of network resources in terms of the statistical characteristics of the sources, and their Quality of Service requirements. It provides a measure associated with the source for performance guarantees expressed in terms of cell loss or delay, and so the CAC algorithm is reduced to a consideration of whether the sum of Effective Bandwidths is less than a threshold value.

Kelly and Gibbens *[36][37]* state the definition of Effective Bandwidth of a source as depending on two parameters, the space and time scaling. The choice of these time scales depends on the characteristics of the resource, capacity, buffer size, traffic model, etc. The Effective Bandwidth is given by the statistical descriptor:

$$\alpha(s,t) = \frac{1}{st}\log E[e^{sX[\tau,\tau+t]}]$$

where s is the space scale (in bytes or cells) and t is the time scale (in seconds). $X[\tau,\ \tau{+}t]$ is the workload arriving at a resource in time period $[\tau,\ \tau{+}t]$ and the expectation is taken over the distribution of random periods. This means that $\alpha(s,t)$ lies between the mean and peak arrival rates of the source measured over an interval *t*. Hence improved link utilization results if the Effective Bandwidth can be allocated instead of the peak rate bandwidth requirement.

The definition of Effective Bandwidth for *X[0,t]* is the amount of work that arrives from a source in the interval *[0,t]*. Assuming that *X[0,t]* has stationary increments, the Effective Bandwidth of the source is defined as:

$$\alpha(s,t) = \frac{1}{st}\log E[e^{sX[0,t]}] \qquad \text{for } 0 < s,t < \infty$$

with properties as described in the Appendix. The scales of time and space are determined by the source and Quality of Service required, and by the capacity of buffer lengths. Kelly *[36]* derives the *α (s,t)* Effective Bandwidth descriptors for different source traffic models - Bernoulli bufferless models, periodic models, fluid models and fractal Brownian motion input models. They lead to admissible regions that give the time and space scales, *s* and *t*, for these sources.

4.3 Effective Bandwidth and General 'ON-OFF' Sources

Let the source alternate between long periods in an 'ON' state with an Effective Bandwidth $\alpha_1(s,t)$ and long periods in an 'OFF' state where it produces no workload. If *p* is the proportion of time spent in the 'ON' state, for small values of *t* compared with the periods spent in the 'ON' or 'OFF' states, then:

$$\mathrm{E}[e^{X[\tau,\tau+t]}] = \mathrm{E}[e^{X[\tau,\tau+t]} \mid \text{Source is 'ON'}]\,\mathrm{p} + \mathrm{E}[e^{X[\tau,\tau+t]} \mid \text{Source is 'OFF'}]\,(1\text{-p})$$

$$= \mathrm{E}[e^{sX_1[\tau,\tau+t]}]\mathrm{p} + \mathrm{E}[e^{s0}](1\text{-p})$$

where $X_1[\tau,\tau+t]$ is the work generator for [τ,τ+t] by the 'ON' source. By definition of $\alpha_1(s,t)$: $\mathrm{E}[e^{X_1[\tau,\tau+t]}] = \alpha_1(s,t)$

hence: $\mathrm{E}[e^{sX_1[\tau,\tau+t]}] = \mathrm{p}\,\alpha_1(s,t) + 1\text{-p}$

and so: $\alpha(s,t) = \dfrac{1}{st}\log\left[1 + p(e^{(st\,\alpha_1(s,t))} - 1)\right]$

The 'ON' periods may at a finer time scale appear as a periodic source, with bursts having a structure, so the definition of Effective Bandwidth depends on the range of *s* and *t*.

4.4 Multiplexing Models

The arrivals process is assumed in *[36]* to be the aggregation of the sources:

$$X[0,t] = \sum_{j=1}^{J}\sum_{i=1}^{n_j} X_{ji}[0,t]$$

The $(X_{ji}[0,t])_{ji}$ are independent processes with stationary increments whose distributions may depend on *j* but not on *i*, and the resource such as the switch has to cope with the aggregate arriving stream of work. The number of sources of type *j* is n_j, and the Effective Bandwidth $\alpha_j(s,t)$ for a source of type *j* is thus:

$$\alpha(s,t) = \sum_{j=1}^{n_j} n_j\,\alpha_j(s,t)$$

The point of looking at multiplexing models is to figure out the constraints that exist, and to see if the sum of Effective Bandwidths for n_j number of sources is within the acceptance region for resource and Quality of Service requirements. The acceptance region is defined by a set of vectors $(n_1, n_2, \ldots n_j)$, for which a given performance in terms of queuing delay or buffer overflow is guaranteed.

The constraints are (s^*, t^*, C^*) with the relationship:

$$\sum_{j=1}^{J} n_j \alpha_j (s^*, t^*) \leq C^*$$

The choices of values for the constraints *(s*, t*, C*)* and the acceptance region vectors are described for different types of multiplexing models in the Appendix.

4.5 Connection Acceptance Control for 'ON-OFF' Sources and Charging Mechanisms

Kelly *[36]* proposes using a charging mechanism and CAC algorithm based on a combination of prior declarations and empirical averages, let:

$$Z = E[e^{sX[\tau,\tau+t]}]$$

and so the Effective Bandwidth of the source is:

$$\alpha(z) = \frac{1}{st} \log E[Z]$$

Before admission of a call, the network requires the user to specify a value z, and then charges an amount *f(z:Z)* per unit time, where Z is estimated by an empirical averaging. The user is assumed to select the value z' for minimising the expected cost per unit time. The tariff *f(z;Z)* should be chosen so as to allow the network to estimate the number of users from the estimate of z from z' so that f(z' ;Z) is proportional to α(z). Kelly shows that the appropriate function is:

$$\mathrm{f(z;Z)} = \mathrm{a(z)} + \mathrm{b(z)Z}$$

defined as a tangent to the curve *α(Z)* at the point $Z = z$.

An 'ON-OFF' source produces a workload of constant rate h when in an 'ON' state, and none in an 'OFF' state. Let M and h represent the mean and peak rates in the equation for Effective Bandwidth so that $\alpha_1(s,t) = h$, and $p = M/h$. If h is fixed then:

$$Z = 1 + \frac{M}{h}[e^{sth} - 1]$$

When z is evaluated using the above formula and M is replaced by m the tariff *f(z;Z)* becomes : f(z;Z) = a(z) + b(z)Z = a[m,h] + b[m,h] M

It is the tangent to the function:

$$\alpha[M, h] = \frac{1}{st}\log\left[1 + \frac{M}{h}[e^{sth} - 1]\right]$$

at the point $M = m$. The interpretation is that for a tariff, the user is free to choose a value m, and then incur a charge of *a[m,h]* per unit time, and a charge of *b[m,h]* per unit volume carried.

The admission control algorithm associated with the above tariffs is as follows. Suppose that a resource has accepted connection times *1,2, ... i* and that (a_i, b_i) are the coefficients $(a(z_i), b(z_i))$ chosen at connection time. The resource measures the load $X_i[\tau, \tau+1]$ produced by connection i over a time t, let $Y_t = exp(s\ X_i[\tau, \tau+1])$. The effective load on the resource is then defined to be:

$$\sum_{t=1}^{I} (a_t + b_t Y_t)$$

The new connection is accepted if the calculated effective load is below or above a threshold value.

4.5.1 'ON-OFF' Sources

For h_i the fixed peak rate of connection i then (a_i,b_i) for the coefficients $(a[m_i,h_{i]},b[m_i,h_i])$ chosen by the user, and the measured load from the connection i is $M_i = X_i[\tau,\tau+t]/t$. Then the effective load on the resource becomes:

$$\sum_{i=1}^{I}(a_i + b_i M_i)$$

This is compare with the threshold value to determine connection acceptance.

4.6 Effective Bandwidth and Equivalent Capacity

The nature of Effective Bandwidth for statistically multiplexed sources is examined in order to assess the allocation of bandwidth for a connection to meet Quality of Service requirements. A unified metric is proposed for the representation of Effective Bandwidth of individual connections and also the aggregate multiplexed connections *[33][35]*. A computationally simple approximate expression of the 'equivalent capacity' is made from this metric. The model used to characterise the connection is significant. The approach in *[33]* is to combine two approximations, one that represents the sources with a fluid flow model, and a second approximation that focuses on the distribution of stationary bit rate of the link. The first approximation is to estimate where the impact of individual connections is critical, the second to represent bandwidth requirements when the effects of statistical multiplexing is significant. So the two approximations complement each other and are also computationally simple. This allows for real-time implementation.

The bit rate generated by a number of multiplexed connections is represented by a continuous flow of bits. It varies with intensity according to the state of the underlying continuous-time Markov chain. This Markov chain is obtained from the superposition of the sources associated with each connection. The aggregate bit rate offered to a buffer is emptied at a constant rate of c. Guerin et al *[33]* determines the smallest value C (equivalent capacity) of c such that the overflow probability (representing QoS) is smaller than ε. The determination of the equivalent capacity C requires that first an expression is found giving the distribution of the buffer contents as a function of the connection characteristics and the service rate. This expression is then inverted to determine the value of the service rate, which ensures an overflow probability of ε or smaller for the available buffer size. The value of the overflow probability is the equivalent capacity.

4.7 Effective Bandwidth of General Markovian Traffic Sources

Elwalid and Mitra *[21][34]* show that the Effective Bandwidth of a Markovian source is the maximal real eigenvalue of a matrix. It is derived from the source parameters, network resources and service requirements, with dimension equal to the number of source states. Two sets of results are obtained, one for Markov modulated fluid sources with a fluid model, and also results for queues and point processes, where the sources are Markov modulated Poisson or phase renewal processes. They add to the results for 'ON-OFF' fluid sources, as described in the last sections. Effective Bandwidth is based on source characteristics and call acceptance criteria, and so can be used as a basis for call admission. Its value is bounded between the peak and mean rates.

The model of statistical multiplexing is made up of fluid sources, each source being characterised by $(\boldsymbol{M},\lambda)$ where $\boldsymbol{M}$ is the infinitesimal generator of the controlling Markov chain. The source generates fluid at a constant rate λ_s, when in state s. The mean source rate is λ_m and the peak source rate is λ_p. The multiplexing buffer is serviced by a channel of constant capacity, c.

Let *G(B)* denote the stationary distribution $\Pr[X \geq B]$ where *X* represents the random buffer content and *G(B)* is the overflow probability for buffer size *B*. For a given *B* and *p*, let the service requirement be $\{G(B) \leq p\}$, which is also the admission criterion. The value of *p* is small, e.g. 10^{-6}.

First consider a multiplexing system with only one source, $(\boldsymbol{M},\lambda)$. The asymptotic regime is where $p \to 0$ and $B \to \infty$ so that $\log p/B \to \zeta \in [-\infty, 0]$, and the admission criterion is satisfied if $e<c$ and violated if $e>c$, where *e* is the Effective Bandwidth.

The Effective Bandwidth is the maximal real eigenvalue of the matrix $[\mathbf{\Lambda} - \frac{1}{\zeta}\mathbf{M}]$, where $\Lambda = diag(\lambda)$. The Effective Bandwidth *e* depends on $(\boldsymbol{M}, \lambda)$, and on the buffer and overflow probability only through ζ. Next, the single source considered is in fact an aggregate of *K* arbitrary sources, $(\boldsymbol{M}_k,\lambda_k)$ $(1 \leq k \leq K)$. The result obtained is very simple, the Effective Bandwidth becomes $e = \sum e_k$, where e_k is the Effective Bandwidth of a single source in the system.

The results carry over to the framework of queues and point processes. The source characterisation differs only in that λ_s is the rate of the Poisson stream that is generated by the source in state *s*. The Effective Bandwidth of a single source $(\boldsymbol{M},\lambda)$ in the multiplexing stream is now the maximal real eigenvalue:

$$[\frac{1}{e^{\zeta}}\mathbf{\Lambda} - \frac{1}{1-e^{\zeta}}\mathbf{M}]$$

For the fluid model the Effective Bandwidth decreases monotonically with increasing ζ from λ_p at $\zeta = -\infty$ to λ_m at $\zeta = 0$.

4.7.1 Call Admission with Heterogeneous Classes of Sources

The following observation on Effective Bandwidth *[34]* is useful in its estimation from measurements. A source is supplied by a buffer serviced by a channel of variable capacity c. The Effective Bandwidth e is the value for c for which the asymptotic slope of *log G(x)* = ζ.

For call admission with heterogeneous classes of sources, the condition is:

$$A(B,p) = \{\mathbf{K} = K_1,...K_j : G_{\mathbf{K}}(B) \leq p \ \} \cong \sum e_j K_j < c$$

the asymptotic result is that $A(B,p)$ is essentially the constraint $\sum e_j K_j < c$, where e_j is the Effective Bandwidth of a single source of class j. The approximation from the asymptotic result, $\sum e_j K_j < c$, is the acceptance set in real, non-asymptotic cases.

4.7.2 Mathematical Development of the Inverse Eigenvalue Problem

The mathematical development *[21][34]* is in two stages:

1. Analysis of a single source: This is an inverse eigenvalue problem. The growth of properties of a maximal real eigenvalue occurs with respect to a parameter in the problem. This is due to the convex behavior of the maximal real eigenvalue of essentially non-negative matrices with respect to all diagonal elements.

2. The algebraic decompositions which give the additive form of the Effective Bandwidth of several sources; decompositions based on Kronecker representations.

This section covers basic background facts about the statistical multiplexing system in three parts. First there is a description of a standard eigenvalue problem to compute the spectral expansion of the systems stationary distribution. The second part broadens the scope of the eigenvalue problem by introducing the parameter, channel capacity. The eigenvalues are viewed as functions of this channel capacity. Then the inverse problem is described which is also an eigenvalue problem. Finally there are some facts about essentially non-negative matrices, and the maximal real eigenvalues are presented. These are critical for the analytical development of this algorithm.

4.7.3 The Statistical Multiplexing System

The model of statistical multiplexing *[21][34]* consists of a buffer supplied by independent Markov modulated fluid sources. It is serviced by a channel of constant capacity, i.e. of rate *c*. The sources are described by lumping them into a single Markov modulated fluid source with state space *S* and irreducible generator ***M***. The source generates fluid at a constant rate λ_s, when in state *s* (s∈S). Let $\lambda = \{\lambda_s \mid s \in S\}$. So the aggregate source is characterised by *(**M**,λ)*. Let the rate matrix $\Lambda = diag\ (\lambda)$.

Let Σ denote the stationary aggregate-source state and *X* the buffer content. Let the stationary source distribution of the multiplexing system be denoted by $\pi\ (x)$ where $\{\lambda_s \mid s \in S\}$ and:

$$\pi_S(x) = \Pr(\Sigma = s, X \leq x) \qquad (s \in S, 0 \leq x \leq \infty)$$

The governing system of differential equations is:

$$\frac{\partial \pi(x)}{\partial x}\mathbf{D} = \boldsymbol{\pi}(x)\mathbf{M} \qquad (0 \leq x \leq \infty)$$

where $\boldsymbol{D}=\Lambda - c\boldsymbol{I}$ and $\boldsymbol{I}$ are the identity matrixes and the diagonal element $Dss = (\lambda_s - c)$ is the drift, or rate of change in the buffer content when the source is in state s. Hence we call $\boldsymbol{D}$ the drift matrix.

The stationary probability vector for the aggregate source is denoted by $\boldsymbol{w}$; hence $\boldsymbol{wM} = \boldsymbol{0}$ and $<\boldsymbol{w},\boldsymbol{1}> = 1$ where 1 is the vector in which all elements are unity. The ergodicity condition is $\lambda_m < c$, the mean source rate is $\lambda_m = \langle \lambda, w \rangle$. The peak source rate is $\lambda_p = \max_s \lambda_s$. It is assumed that $c<$ *peak rate.*

Since the spectral state distribution is a bounded solution, it has the spectral representation:

$$\boldsymbol{\pi}(x) = \sum_{i:\operatorname{Re} z_i < 0} a_i \boldsymbol{\Phi}_i e^{z_i x} + \mathbf{w}$$

where (z_i, Φ_i) is the eigenvalue/eigenvector pair. Such pairs are solutions to the eigenvalue problem:

$$z\ \boldsymbol{\Phi D} = \boldsymbol{\Phi M}$$

The eigenvalues with real negative parts are indexed as:

$$0 > Re\, z_1 \geq Re\, z_2 \geq Re\, z_3 \geq \ldots$$

If z_1 is real and $z_1 > Re\ z_i$ for all i>1, then z_1 is called the dominant eigenvalue.

In the spectral expansion, the coefficients $\{a_i\}$ are obtained by solving a system of linear equations that are obtained from the following boundary conditions:

$$D_{ss} > 0 \Rightarrow \pi(s,0) = 0$$

The number of such conditions exactly equals the number of eigenvalues for negative real parts.

The stationary buffer overflow distribution is given by *G(x)*, i.e.

$$G(x) = \Pr(X \leq x)$$

$$= 1 - < \pi(x), 1 >$$

$$= \sum_{i \geq 1} a_i < \Phi_i, 1 > e^{z_i x}$$

if z_1 is the dominant eigenvalue, then:

$$G(x) \approx a_1 \langle \Phi_1, 1 \rangle e^{z_1 x} \qquad as \quad x \rightarrow \infty$$

note that: $$z_i = \lim_{x \to \infty} \frac{\log G(x)}{x}$$

Plots of log *G(x)* versus *x* approach linearity as *x* increases and the slope approaches z_i.

4.7.4 The Inverse Eigenvalue Problem

Consider the eigenvalue problem *[21][34]* as before:

$$z\, \Phi(\Lambda - cI) = \Phi M$$

The scope of the problem is extended by considering *c* to be a variable parameter and the eigenvalues to be functions of *c*, *z(c)*. The inverse problem requires *c* to be found for a given *z*. This is done with an inverse eigenvalue problem, with *c* = *g(z)*:

$$g(z)\ \Phi = \Phi A(z)$$

where $A(z) = \Lambda - 1/z\ \boldsymbol{M}$. This means $g(z)$ is an eigenvalue of the matrix $A(z)$ in which z is a parameter. The inverse eigenvalue problem, its maximal real eigenvalue and behavior of this eigenvalue as a function of z is central to the mathematical development for the algorithm.

4.7.5 Essentially Non-Negative Matrices

A real matrix with non-negative elements off the main diagonal is called essentially non-negative. The matrix $A(z)$ is essentially nonnegative *[21][34]*, for real and negative z. Since $\boldsymbol{M}$ is irreducible, so is $A(z)$. By adding $\sigma\mathbf{I}$ to $A(z)$

where: $$\sigma > \left[\max_i \left(\frac{1}{z} M_{ii} - \lambda_i\right)\right]^+$$

a nonnegative matrix is obtained whose eigenvalues are those of $A(z)$ shifted by σ.

4.8 A Single Source: Monotonicity of Eigenvalues and Effective Bandwidth

The single source studied next is an aggregate of many lower-order sources. The properties of the eigenvalues, monotonicity and convexity are established in *[34]* and the asymptotic view of the admission control problem is introduced, as well as proving that the Effective Bandwidth of the source as the maximal real eigenvalue.

The Effective Bandwidth is monotonically increasing and convex function of all state-dependent rates of the source. A corollary in *[34]* shows that the coupling of state transitions of two sources with identical generators for their controlling Markov chains and proportional rate vectors, the effect is to increase the Effective Bandwidth.

4.9 Multiple Markov Modulated Sources and Admission Control

The results for single sources are extended to multiple multiplexed systems with several sources as follows. The asymptotic regime of buffer overflow probability of 10^{-6} is specified by scaling, to arrive at the following natural asymptotic regime, by letting $B \rightarrow \infty$, and $p \rightarrow 0$ so that:

$$\log p = \zeta B \; + O(1)$$

Let the admission criterion be $G(B) \leq p$ hence $\log p/B \rightarrow \zeta$. where $\zeta \in [\, -\infty, \, 0 \,]$. The following characterises K sources that satisfy the admission criterion in this asymptotic regime. If there are K sources:

$$(\mathbf{M}^{(k)}, \boldsymbol{\lambda}^{(k)}) \qquad \text{for } \; (1 \leq k \leq K)$$

Let the admission criterion be $G(B) < p$. Suppose $B \rightarrow \infty$, and $p \rightarrow 0$ as letting $\log B/p = \zeta \in [\, -\infty, \, 0 \,]$.

If $\sum_k g_1^{(k)}(\zeta) < c$, then the admission criterion is satisfied. Here $g_1(\varsigma)$ is the maximal real eigenvalue of:

$$\mathbf{A}^{(\mathbf{k})}(\zeta) = [\boldsymbol{\Lambda}^{(\mathbf{k})} - \frac{1}{\zeta}\mathbf{M}^{(\mathbf{k})}]$$

4.10 Summary

The algorithms of Effective Bandwidth are of great importance and have been the focus of much of the research in the area of connection admission control and resource allocation. They demonstrate the application of large deviations theory *[26]* and its approximations for bandwidth allocation. Kelly *[36]* derives the Effective Bandwidth descriptors for different source models and the CAC algorithm for the 'ON-OFF' source model, while Elwalid and Mitra *[21][34]* show how the Effective Bandwidth for Markovian source models in general is the maximal real eigenvalue of a matrix derived from source parameters. Admission control policies can be found from these approaches, note that extra background theory is found in the Appendix. The numerical evaluation in Chapter 6 will provide a basis for comparison of the Effective Bandwidth algorithms with others, and will clarify the advantages of this method.

Chapter 5

The Measurement Approach

5.1 Introduction

The estimation of bandwidth requirements may be approached in two ways. One approach is to assume a parametric model of the traffic and the parameters for the connection to be added, as in Chapters Three and Four. These parameters are found from information declared by the connection when it requests admission, or measurements made on the traffic generated by the connection, or a combination of both. Once the detailed model is completed, the estimate can be calculated. The problems with this approach are that unless online measurement is employed, the application is required to deliver a detailed self-characterisation before it has transmitted any traffic. Then the network still has to fit suitable parameters to a model that adequately describe the traffic source, given such a characterisation. This may be difficult and the solution may contain redundant information. Also a new traffic type may mean a complex modeling process in advance of transmission.

An alternative approach used by Measurement-Based algorithms *[10]-[13][67]-[74]* is to measure the bandwidth requirement directly. This avoids the problem of requiring new traffic types to specify a parameterised model in advance and removes the estimation of redundant information. The important advantage of this approach is that it requires very little declared information on the part of the application. Measurement-Based Admission Control (MBAC) algorithms study the performance of a scheme that has no prior knowledge of the traffic statistics and makes the admission decision on the current state of the network only. In contrast to the other algorithms, which look at the characteristics of source traffic and represent them as parameters, Measurement-Based algorithms make decisions on a monitored amount of traffic on the network. This means that the information about the behavior of the cells at a given moment is measured and this information is used to make a decision. The following sections present a range of Measurement-Based algorithms and examine their behavior by simulations with different traffic sources. The advantages of the measurement approach can be seen in contrast to other types of algorithms, particularly those with long-range dependent (LRD) traffic. They provide an exciting new approach to admission control and opportunities for further research.

5.2 Measurement-Based Algorithms

There are several different algorithms with measurement as described in *[12][13][68]-[74]*. Through the use of analysis and simulations the performance of Measurement-Based algorithms is explored, and the dynamics of the system analysed.

5.2.1 The 'Certainty Equivalent' Controller Algorithm

The first algorithm is called 'certainty equivalent' controller *[12]*. This is an admission controller that assumes that the measured statistics are the true statistics of the calls, and uses this information to make decisions. The two performance measures that are of interest are the steady-state probability of the event that the system overloads, and the expected fraction of the bandwidth utilised. The success of the admission control scheme is evaluated by how well it meets the Quality of Service requirement.

A Measurement-Based algorithm accepts or rejects a call based on the observed past history of calls that are currently in the system and have possibly terminated. There may be no prior knowledge of the sources but measurements of traffic flow are taken and measurement errors are also to be considered. The analysis in *[12]* estimates the statistics of the calls from observing their past empirical behavior. The scheme has a number of calls $n_k(t)$ that are currently generating data at rate c_k, for each k $(k=1,...,K)$. This gives the empirical distribution $\{\tilde{\Pi}_k\}$ of bandwidth requirements for a typical call, and a distribution:

$$\Pi_k = \frac{n_k(t)}{x(t)}$$

where *x(t)* is the number of calls currently in the system at time *t*. The idea is to use $\{\tilde{\Pi}_k\}$ to estimate the distribution of $\{\Pi_k\}$ which is the bandwidth requirements for the duration of the call. The admission control scheme is of a 'certainty equivalent' type, the controller assumes that the measured values are the true parameters. The performance is studied with fluid approximation and large deviations analysis. An acceptance region and rejection region is used to clarify the boundary of call acceptance. The main theoretical result is that a memoryless 'certainty equivalent' control can achieve the performance of the optimal scheme with knowledge of traffic statistics. The conclusion from the simulation studies is that the scheme works well only with large link capacities. For small link capacities it makes too many admission mistakes due to measurement errors.

5.2.2 The Aggregate Traffic Envelope Algorithm

To continue the idea of 'certainty equivalence' from the last section, the paper *[72]* describes a framework with an adaptive Measurement-Based aggregate traffic envelope. It is found from aggregate traffic flow and provides a traffic characterization with its temporal correlation and available statistical multiplexing gain. A 'maximal rate envelope' is measured to characterize the behavior of aggregate flow. The rate envelope describes the traffic flow rate associated with the corresponding interval length. This provides the framework for the development of a new envelope-based MBAC.

The framework for the algorithm has what is called a schedulability confidence level. This reflects the variation and temporal correlation of past envelope measurements, and the uncertainty of the prediction of the future workload. It allows control of the QoS parameters that applications are ultimately concerned with, such as loss probability and delay-bound violation probability so they do not exceed the measured envelope. An extensive set of simulation experiments uses traces of compressed video as well as model-generated long-range dependent traffic. The scheme has been implemented on a test-bed of prototype routers.

The new Measurement-Based admission control approach utilizes the measured values of aggregate traffic envelopes. It consists of a measurement algorithm and an admission control algorithm. The measurement algorithm continually updates the recent empirical aggregate envelope and measures the envelope's temporal variation. The admission control algorithm has a check for aggregate schedulability with an associated predicted confidence level, and also an estimation of the loss probability. The new call is admitted if the predicted performance parameters satisfy the QoS requirements of the new flow as well as all existing flows.

First the aggregate rate envelope is found. An interval length associated with the flow rate is specified. By measuring the maximal rate envelope (defined next) of the aggregate flow, the short time-scale burstiness of the traffic is estimated. This allows for analysis of the dynamics of a buffered multiplexer with a new admission. Then the variation of the aggregate flow's rate envelope is measured, to characterize longer time-scale fluctuations in the traffic characteristics. The confidence values of the schedulability condition can be determined with the variation in the measured envelope. The expected fraction of bits dropped can be estimated should the schedulability condition fail to hold.

When a new flow arrives the aggregate schedulability test is performed. This test ensures that for a given confidence level the cell loss rate is within an acceptable level. This confidence level is necessary as there is no *a priori* assurance that the past envelope will prove adequate for the aggregate flow. Consider a new flow bounded by $r_k, k = 1,...,T$ that requests admission for traffic and having a service rate *C*, with minimum delay *d*, the minimum interval for the measured rate envelope τ, and buffer capacity of at least *C.d*. The aggregate flow is characterized by peak rate R_k and mean $\overline{R}_k$ and variance ${\sigma_k}^2$, $k = 1,...,T$. With a new admission no loss will occur with confidence level $\Phi(\alpha)$ if:

$$\max_{k=1,2,...T} \{k\tau(\overline{R}_k + r_k + \alpha\sigma_k - C)\} < Cd$$

The rate envelope of the aggregate process is significantly less than the sum of individual worst-case envelopes.

Having completed the schedulability test, the loss probability test is important. This is because if the traffic exceeds the aggregate envelope it will result in loss and delay. To satisfy loss requirements the MBAC has the following test. The aggregate flow has satisfied the schedulability test and has mean bounding rate $\overline{R}_k$ and variance ${\sigma_k}^2$ over intervals of length $k\tau$. For link capacity C, buffer size B and schedulability confidence level $\Phi(\alpha)$, the cell loss probability is:

$$P_{loss} \approx \max_{k=1,2,...,T} \frac{\sigma_k \Psi(\alpha)}{\overline{R}_T}$$

where: $$\Psi(\alpha) = \frac{1}{\sqrt{2\pi}} e^{-(\alpha^2/2)} - \alpha[1 - \Phi(\alpha)]$$

5.3 Comparison of Measurement-Based Algorithms with other Approaches

In the study *[13]* the two basic approaches to admission control are compared. The first is the parameter-based approach computing the amount of network resources from the current traffic flow levels. Then there is the Measurement-Based approach, which relies on the measurement of actual traffic load in making admission decisions. Three Measurement-Based algorithms are described based on ideas of measured bandwidth, acceptance region and equivalent bandwidth respectively. The simulation studies for several network scenarios evaluate the link utilisation and the adherence to service commitment achieved by these four algorithms. These three algorithms are investigated and provide an interesting spectrum of ideas.

The authors in *[13]* claim that service commitments made by Measurement-Based algorithms can never be absolute. Their Measurement-Based approaches are used in the context of service models that do not make guaranteed commitments and to provide this they have a controlled-load service model. The controlled-load service is designed for adaptive real-time applications that can tolerate variance in packet delays. The controlled-load service is suited to the decentralized and heterogeneous Internet. The same principles can be applied to an admission control service for the ATM protocol. The network switches and routers perform admission control at the call level to ensure that sufficient resources are available to serve the flows.

5.3.1 The Measured Sum Algorithm

The admission control algorithms *[13]* are compared with the Simple Sum algorithm, which simply calculates the sum of requested resources and checks that it does not exceed link capacity. Let v be the sum of reserved rates, μ the link bandwidth, α the name of a flow requesting admission, and r^{α} the rate requested by flow μ, so: $v + r^{\alpha} < \mu$

The Measured Sum algorithm uses measurement to estimate the load of existing traffic. This algorithm admits the new flow if the following succeeds:

$$\hat{v} + r^{\alpha} < v\mu$$

where v is a utilization target, and $\hat{v}$ the measured load of existing traffic. The point is made that with a simple *M/M/1* queue, variance in queue length diverges as the system approaches full utilization. This is an issue at very high utilization when a Measurement-Based approach may fail due to the large delay variations. It is thus necessary to keep link utilization below this level and so is set at $v = 0.9$.

5.3.2 The Acceptance Region Algorithm

This is the second Measurement-Based algorithm from *[13]* and it computes an acceptance region that maximizes the reward of utilization against the cell loss. The algorithm ensures that the measured instantaneous load plus the peak rate of a new flow is below the acceptance region. The link bandwidth, switch buffer space, a flow's token bucket filter parameters, the flow's burstiness and desired probability of actual load exceeding bound, are all used to compute an acceptance region for a set of flow types. It assumes Poisson call arrival process and independent, exponentially distributed call holding times.

5.3.3 Equivalent Bandwidth MBAC

The third Measurement-Based algorithm described *[13]* finds the equivalent bandwidth for a set of flows. The equivalent bandwidth of a set of flows is defined as the bandwidth $C(\varepsilon)$ such that the stationary bandwidth requirement of the set of flows exceeds this value with probability ε. The measured average arrival rate is approximated by measured average load and the peak rate is p. The admission control check when a new flow a requests admission is found to be:

$$\hat{C}_H + p^{\alpha} \leq \mu$$

The measurement mechanisms used in the study *[13]* are simplistic. However, they do provide an insight into how Measurement-Based algorithms can be examined. The first is a simple time-window measurement mechanism to measure network load with the "Measured Sum" algorithm. The average load every S sampling period is computed. At the end of a measurement window *T*, the highest average from the just ended *T* is used as the load estimate for the next *T* window.

When a new flow is admitted to the network, the estimate is increased by the parameters of the new request. If a newly computed average is above the estimate, the estimate is immediately raised to the new average. At the end of every *T*, the estimate is adjusted to the actual load measured in the previous *T*. Other measurement mechanisms used are point samples and exponential averaging. Point samples is a measurement mechanism used with the acceptance region algorithm, it takes an average load sample for a given period. Exponential averaging uses an estimate of the average arrival rate, instead of instantaneous bandwidth, to compute admission decisions with the equivalent bandwidth approach.

5.4 The Performance of Measurement-Based Algorithms

The previous section describes work in *[13]* which provides a basis for further consideration by the same authors in *[70]*. Their work is extended to provide a more comprehensive comparative study. Six Measurement-Based algorithms are compared and the performance of the algorithms is examined. In an effort to better support applications with real-time constraints, several new per-flow cell delivery services have been proposed, instead of those providing worst-case guarantees, and better than 'best-effort' services. That is they provide an enhanced Quality of Service without making hard guarantees. Specifications for these services might provide a delay *target,* rather than a bound. Parameter-based admission control algorithms that are based on worst case bounds are derived from parameters describing the flow, and so will result in low network utilization in the face of bursty network traffic. Measurement-Based admission control algorithms (MBACs) are more appropriate because they base admission control decisions on measurements of existing traffic rather than on worst-case bounds about traffic behavior. MBACs can achieve much higher network utilization than parameter-based algorithms while still providing acceptable service. Since traffic measurements are not always good predictors of future behavior, the Measurement-Based approach to admission control can lead to occasional cell losses or delays that exceed desired levels. These are acceptable with the relaxed nature of the service commitment provided.

The algorithms are evaluated according to two perspectives to satisfy the goals set by CLR and QoS constraints. First, the performance frontier or loss-load curve achieved by each algorithm is compared, where the loss-load curve depicts the rate of losses that occur at a given level of utilization. Second, the question is how close is the resulting performance is to the target. Next are descriptions of the six admission control algorithms. Each algorithm has two key components, a measurement process that produces an estimate of network load, and a decision algorithm that uses this load estimate to make admission control decisions.

5.4.1 The Hoeffding Bounds Algorithm

The admission control algorithm described in *[73]* computes the equivalent bandwidth for a set of flows using the Hoeffding bounds. A new flow is admitted if the sum of the peak rate of the new flow and the measured equivalent bandwidth is less than the link utilization.

5.4.2 Tangent at Peak and at Origin Algorithms

The first of four algorithms presented in *[13]* is based on the tangent at the peak of an equivalent bandwidth curve computed from the Chernoff Bound, and uses a point sample measurement process. A second algorithm uses a tangent to the equivalent bandwidth curve at the origin. This admission control algorithm also uses the point sample measurement process.

5.4.3 The Measure CAC

The Measure admission control algorithm *[69],* which is based on large deviation theory, admits a new flow if the sum of the peak rate of the flow and the estimated bandwidth of existing flows is less than the link bandwidth. The estimated bandwidth takes as input a target loss rate using the scaled cumulative generating function of the arrival process. An extension of this work is presented in the next section *[68]*.

5.5 The Shape-Function

This final algorithm is a new direction beyond the Effective Bandwidth and Measurement-Based approaches. In *[68]* a CAC algorithm is described using measurements made on existing connections and the declared parameters of the new connections. This is done using what is known as the Shape-Function. It was developed by Botvich and Duffield *[67]*, by the application of Large Deviation theory to queuing systems. The Shape-Function of the connections is estimated and so predictions can be made about their effect on the network. Using real traffic collected from a network, the performance of this CAC scheme is compared with that of the Mosquito *[69]* algorithm. In contrast, the Mosquito algorithm is found from an estimation of Effective Bandwidths *[11][21][33]-[44]*. The Mosquito algorithm's approach is based on the theory of Large Deviations, a probabilistic theory of rare events, which when applied to queuing systems, can be used to estimate bandwidth requirements (see Chapter 4 and the Appendix). The additive nature of the effective bandwidth approach fails to take the economies of scale into account. These arise from statistical multiplexing because it is based on large buffer asymptotics. An alternative approach involves estimation of the Shape-Function *[67]* from the multiplexed traffic. For the online estimation of bandwidth requirement there are comparisons for two estimators, the Shape-Function estimator *[68]* and the Mosquito estimator *[69]*.

The main issue to address is the loss of cells due to overflow at a buffer. A multiplex of N ATM streams is considered arriving at a buffer which has finite storage capacity B. Cells are then removed from the buffer at fixed rate S called the line-rate. The cell loss ratio for a multiplex of N lines with a buffer size B and a line-rate is denoted by $CLR(N, B, S)$. The logarithm of $CLR(N, bN, aN)$ is asymptotically linear in the number of sources *N* if the line-rate per source *a* and buffer size per source *b* are fixed. The multiplexing gain available in shared resource systems due to the statistical properties of the individual traffic streams is shown in *[68]*.

The arrival streams are modeled as stationary stochastic processes and are the total number of cells which have arrived up to time *t* from source *a*. For *N* sources feeding a buffer of size *Nb* which is being served at rate *Na,* the proportion of cells lost will satisfy the logarithmic asymptotic for Shape-Function *I(b):*

$$\log CLR(N,Nb,Na) \sim -NI(b), \qquad \text{as } N \to \infty$$

This holds for a wider class of traffic and should be valid for long-range dependent traffic *[76]-[78]*. The limit of the cumulative generating function for each source is found and the Shape-Function is derived from their Legendre transforms *[71]*.

The estimators measure the bandwidth requirements of the current traffic *[68]*. The effectiveness of the estimators is compared. The Shape-Function estimator is:

$$BWR(nB,C) := \min\{sN : e^{-NI(b,s)} \le c\}$$

where *c* is the target CLR.

The Mosquito Estimator *[69]* is based on large buffer asymptotics. In contrast, the Shape-Function estimator is found by assuming asymptotics for a large number of sources. For the Mosquito Estimator arrivals processes the loss ratio decays exponentially with buffer size. The decay rate is determined by the line-rate and by the CGF of the multiplexing of sources. The bandwidth requirement is found from simulations of a bufferless system and the observing cell loss.

5.6 Experiments and Conclusions

The algorithms presented are further examined with experiments and simulations. The sources are varied, from 'ON-OFF' to video and long-range dependent types. The findings provide insight into the strengths and weaknesses of the measurement approach, with particularly interesting results for LRD traffic.

The 'certainty equivalent' controller *[12]* is the first algorithm presented in this Chapter. It uses the star wars video trace as a traffic source, with calls arriving according to a Poisson process. The conclusions are that the scheme works well for large link capacities, with too many measurement errors for small link capacities. The simulation experiments in *[72]* continue the ideas of the 'certainty equivalent' by evaluating the aggregate traffic envelope algorithm's performance and comparing it with *[73][74]*. The algorithm uses the maximal rate envelope to capture multiplexing properties of the aggregate traffic flows. The autocorrelation structure *[5][75]* is also shown from the traffic envelope. The maximal envelope characterizes the extreme values of traffic flow and the variation of the maximum rate tends to be less than the variance of flow itself. The sources are MPEG compressed video and heavy-tailed 'ON-OFF' sources that form long-range dependent traffic in aggregate. The aggregate envelope MBAC achieves higher utilizations than both those in *[73][74]* while still satisfying the QoS requirements. The experiments indicate that aggregate flow rather than user-specified per-flow peak rates allow more control for exploiting statistical multiplexing gain. At higher link capacities higher utilization is again achieved with buffering gain. The aggregate envelope MBAC performs well over a wide range of link capacities and buffer sizes.

With long-range dependence *[5][75]-[78]* a new area of interest is highlighted. The 'ON-OFF' sources in experiments with the traffic envelopes *[72]* have heavy-tailed distributions exhibiting self-similarity when aggregated. These are Pareto 'ON-OFF' sources described *[13][74]*, as follows. The Pareto distribution is a heavy-tailed distribution that is described by two parameters, the location and shape. A Pareto shape parameter that is less than one gives data with infinite mean, a shape parameter less than two results in infinite variance. Each Pareto 'ON/OFF' source by itself does not generate a LRD series but its aggregation does. The aggregate envelope approach out-performs the other techniques as there is a utilization gain due to buffering. The temporal correlation of successive traffic envelopes is exploited to incorporate the effects of flow arrivals and departures. The traffic dynamics are captured at time-scales larger than that of the envelope and measurement window. This means an effective and versatile algorithm for a wide range of traffic types, buffer sizes and link capacities.

The Measured Sum, Acceptance Region and Equivalent Bandwidth MBACs are explored in *[13]*. There is a Simple-Sum algorithm (which simply adds PCR) and it has no cell loss, while the Measure-Sum achieves this if the utilization target is decreased to 80% with LRD. The Acceptance Region algorithm is thought to be overly optimistic for sources with heavy-tailed 'ON' and 'OFF' distributions. The Equivalent Bandwidth algorithm performs better with lower peak rates, but still lags behind the other Measurement-Based algorithms. Using a conservative approach the algorithm displays no cell loss. The simulations also explore a long-range dependence with two kinds of source model. Studies *[74][76]-[78]* have found that network traffic can exhibit long-range dependence, which implies that congested periods can be long and a slight increase in the number of active connections can result in large increase in cell loss rate. This may mean that long-range dependent traffic might have a damaging effect on Measurement-Based admission control algorithms. It is investigated with a simulation study with LRD source models. The model is again an 'ON/OFF' process with Pareto distributed 'ON' and 'OFF' times, as discussed earlier. The findings have very significant and grave implications for MBAC algorithms. Fortunately they are found to be refuted by the further LRD studies described next.

5.6.1 Experiments with the Performance of MBACs

The algorithms in *[70]* are presented to comment on the performance of MBACs and are evaluated in terms of the two goals of achieving high network utilization and low cell loss. They are the Heoffding Bounds, the Tangent at Peak and at Origin and the Measure Algorithms, the Measured Sum and Aggregate Traffic Envelopes are also examined. The simulations focus on three specific issues, that is the impact of heterogeneous traffic, the comparison between MBACs and an ideal parameter-based algorithm, and the implications of long-range dependent traffic on Measurement-Based admission control. Two kinds of source models are used in the experiments. The first is an 'ON/OFF' source and the second kind of source model uses a trace of video traffic to drive the simulation. The average utilization and packet loss rate are measured for each.

The results show that the cell loss rate as a function of link utilization show a 'performance frontier' in the display for the algorithms *[70]*. There is little difference between the performance frontiers. This indicates that all of the algorithms have a very similar performance in the tradeoff between loss rate and utilization. This result was found to hold across several different traffic models such as those with burstier traffic and long-range dependent traffic.

The experiments were elaborated with the examination of a heterogeneous traffic mix, and the MBACs displayed different performance frontiers. Then a comparison with an Ideal Algorithm was performed. There is a question that is concerned with the differences in the performance of the algorithms being so small. Also, how do they perform at optimum? A simple algorithm that accepts or rejects calls according to a quota is used as the 'ideal'. In contrast to the Measurement-Based approach, the quota algorithm admits a call based on an average behavior. However, the MBACs must assume worst case with the new flow. They respond to fluctuations in flow and so their performance is degraded relative to this ideal, but unrealistic algorithm.

The issue of long-range dependence is addressed with experiments in *[70]* using the video sources, with some interesting findings. The contrast with the quota algorithm shows that LRD has significant implications for the Measurement-Based algorithms. The Measurement-Based approach shows a *better* performance than that of the quota algorithm. This is explained by the fact that LRD traffic displays variations over long time frames, and the Measurement-Based algorithms can adjust the flow to respond. The quota algorithm has a fixed number of flows instead. The adaptation to these long-term fluctuations is a distinct advantage of the Measurement-Based algorithm. The quota algorithm is shown not to be the optimal after all. The results show that the measurement estimation and admission decision can be separated. The other interesting result is that the MBACs have a better performance with long-range dependent traffic than parameter-based algorithms. The second criterion for tunability proved to be disappointing. The authors suggest further research on this issue.

5.6.2 Experiments with the Shape-Function Algorithm

The CAC algorithms discussed in *[68]* are the simple Effective Bandwidth algorithm, the Shape-Function algorithm, and the Mosquito algorithm. The advantage gained by exploiting statistical multiplexing is shown, the CAC algorithm using any of the three estimating techniques, admits significantly more calls than the peak rate allocation scheme. As was found in *[69],* the Mosquito estimator is less conservative than the simple Effective Bandwidth estimator. The performance of the Shape-Function estimator lies between the other two estimators.

The pessimistic CAC algorithm allocates resources using the declared peak rate of each source, is optimal in the sense that the CAC algorithm is assumed to have-complete knowledge of the statistical properties of every connection requesting admission. To find the optimum admission scheme the number of calls was found empirically that could be multiplexed for a given BWR (the link-rate) and CLR.

The three algorithms, particularly the Mosquito algorithm, perform very close to the optimum. Loss occurred when using the Simple Effective Bandwidth or the Shape-Function algorithms, indicating that the Simple Effective Bandwidth algorithm is too conservative. Greater utilisation is achieved by using the Shape-Function algorithm. Since the Mosquito algorithm admits more connections than either the Simple Effective Bandwidth or the Shape-Function algorithms it may be expected that this algorithm exhibits a higher CLR, this was found to be true. It was noted that most connections experience no cell loss while others lose many cells. The Shape-Function estimator has been found to perform better than the simple effective bandwidth estimator but less well than the Mosquito estimator in terms of the number of connections admitted. However, the Shape-Function did not cause any violation of QoS requirements for CLR, unlike the Mosquito algorithm.

5.7 Summary

The allocation of bandwidth can be achieved in two ways. First, with a parametric model of the traffic as was discussed in previous chapters. This model is based on information declared by the connection at call setup time, and then a model that estimates the requirements. The second approach is to measure the bandwidth requirement directly. Measurement-Based admission control (MBAC) algorithms are shown in this chapter to provide better performance. This is despite the fact that the admission decision is made on the current state of the network only, that is without prior knowledge of the traffic statistics. The studies in *[13][70]* provide a comprehensive range of Measurement-Based algorithms, as well as useful analytical techniques. The simulations with LRD traffic prove to be very interesting, especially with new evidence from studies *[74]-[78]* which suggest the importance of self-similarity when modeling network traffic. The Measurement-Based approach is found to be particularly suitable for these sources.

Chapter 6

Numerical Evaluation

6.1 Introduction

This Chapter reports the results of studies that demonstrate four important CAC algorithms. They are the Convolution algorithm, the Chernoff Bound algorithm, the Gaussian Approximation algorithm and the Effective Bandwidth algorithm. There is also a simulation of cell-scale levels with which to compare and verify the algorithms. The numerical evaluation is a series of experiments designed and programmed by the author. The simulations were performed by computer programs written in 'C' language. The results of the simulation experiments are presented in a set of graphs. The experiments were repeated for mixtures of traffic to examine the effects of combinations of different traffic types.

6.2 The Traffic Scenario for Simulations

ATM networks provide performance guarantees to their connections, using traffic models to estimate resource requirements. Chapter 2 describes traffic models in greater detail. The traffic model is implemented as a simulation with 'C' programs. Bursty traffic sources such as video are characterized by the 'ON-OFF' sources in traffic modeling, as described in Section 2.3.1.1, *[1][3][5][6][19]*. The information 'burst' is transmitted at peak rate for the 'ON' period, and none is transmitted in the 'OFF' period. The 'ON-OFF' source is assumed to have exponentially distributed 'ON' and 'OFF' periods.

The admission control algorithms have a set of input parameters. Their arrivals process is simulated by aggregation of the traffic sources. The summation of bandwidth allocation for the sources is estimated for each algorithm as shown in *Figure 5.1*. There are three different traffic types. In the first group the Type I connection has characteristics PCR = 20Mbps and SCR = 10Mbps. The second connection is Type II with characteristics PCR = 10Mbps and SCR=5Mbps. Type

III is the third connection with a lower-speed with characteristics PCR = 5Mbps and SCR=3Mbps.

Type I:	20 Mbps PCR	10 Mbps SCR
Type II:	10 Mbps PCR	5 Mbps SCR
Type III:	5 Mbps PCR	3 Mbps SCR

Table 6.1 Traffic Source Types

The following experiments focused on the single buffer/trunk system. It was important to examine the effect of buffering and multiplexing on the combinations of different traffic sources in the network. The total traffic multiplexed into a node got smoothed out due to buffering, it is interesting to see how the various source types were affected. Some experienced an increase in admission while others lost admission to the network. Each algorithm was examined and the results were illustrated in what is known as an ‘admission region’. This is a graphical display of the admissions made for a combination of sources. The interactions with mixtures of sources can be observed and the impact of combining traffic can be seen.

The bufferless fluid flow model *[5][6][14]* was described in Section 2.4.1. It was used as a basis for the numerical evaluation, with a fluid flow representation of the aggregate traffic rate from all the sources. There was a *M/D/1* queuing structure for the ‘flow’ at the output of the switch or multiplexer. There was a buffer large enough to accommodate bursts of traffic *[14]-[18]*, but not so large as to violate cell delay criteria. The assessment of Quality of Service of the connections was evaluated in terms of Cell Loss Probabilities (CLP), and the simulations could verify whether the amount of cell loss predicted by the CAC algorithms actually occurred.

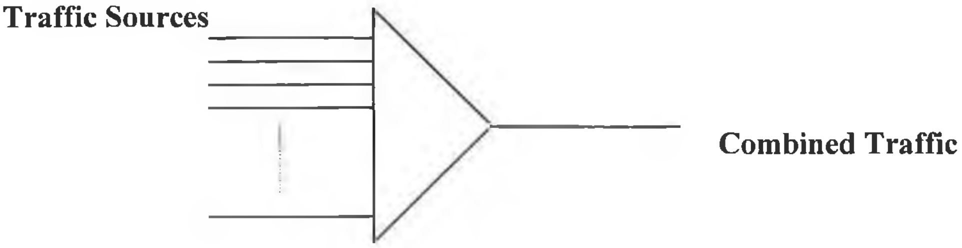

Figure 6.1 Traffic Scenario for Experiments

In the examples studied in this Chapter, we assumed link capacity C to be *155Mbps*, and to have N homogenous 'ON-OFF' type j sources with exponentially distributed transmission times. The connections were described by traffic descriptors comprising of the Sustainable Cell Rate (SCR) and Peak Cell Rate (PCR). The Quality of Service criterion is that the CLP, or Cell Loss Probability was restricted to 10^{-6}.

The experiments that were performed in the numerical evaluation (to be described in this Chapter) were based on a number of approximations with the use of the fluid flow model. The fluid flow model has been shown in *[21]* to admit 22%-27% more than analytical results for the Chernoff Bound Algorithm, for example. Work done and displayed in this Chapter studied the algorithms for multiple source classes. The need to buffer cells during the cell inter-arrival time was estimated, and was important when considering this model. If the number of cells involved was large and the buffering made correspondingly substantial then the fluid model predictions were close to those produced by a discrete model. It was found *[21]* to be optimistic if there was a large cell inter-arrival time for a small number of sources. The comments in *[33]* reflect on these findings, and also point out that the asymptotic approximation itself was likely to be inaccurate for large burst periods. When long burst periods are multiplexed however, a reasonably accurate estimation can be obtained from the stationary bit rate distribution.

6.3 The CAC Algorithms Investigated

There are four CAC algorithms explored by the numerical evaluation results and comparisons. They are the Convolution algorithm, the Chernoff Bound algorithm, the Gaussian Approximation algorithm and the Effective Bandwidth algorithm. A description of each that was used as a basis for the numerical experimentation is found in Chapters Three and Four. The algorithms were chosen to demonstrate a variety of algorithms with differing approaches. They were based on the measure of Cell Loss Probability (CLP). The bufferless fluid flow model calculated the numbers of cells that were discarded when the instantaneous total traffic load exceeds the link capacity *C*. The CLP for 'ON-OFF' sources under the bufferless fluid flow model was estimated and compared with that of a Quality of Service requirement of 10^{-6}. The Effective Bandwidth algorithm (in contrast to the other three) calculated the Equivalent Bandwidth or 'equivalent capacity' (see Chapter 4). The bandwidth was required to be within the network's capacity. The algorithm estimated this to be less than 155Mbps, in these experiments. The formulae formed the basis for the computations performed by computer programs in the coming sections, to examine the algorithms with the numerical evaluation experiments.

6.4 The Cell-Scale Simulation

The effectiveness of the CAC algorithms above is determined by their ability, without excessive computations, to estimate the Cell Loss Probability. The ideal algorithm will overstate the CLP compared to what happens in practice by a small margin, i.e., it will provide a tight upper bound for the CLP. The actual CLP will be determined by simulating the operation of the ATM multiplexer exactly, using cell-scale simulation. A simulation program was written to achieve this. It multiplexed *N* sources and records the cell loss observed in the buffer. It assumed a link rate for the outgoing link and all inputs of 155Mbps, giving a timeslot duration of approx. 2.7 μs. Each was bursty, with geometrically distributed on and off times.

During the 'ON' period, the source generated bits at a constant rate, and thus generated cells periodically (although the cell generation time occasionally slips by one time slot, if the source has not accumulated enough bits). This bit rate corresponded to the 'MAX' rate processed by the CAC algorithms.

The other parameter used by the CAC algorithms, the average bit rate, was related to the source model as follows:

$$\frac{AVG}{MAX} = \frac{r_{01}}{r_{01} + r_{10}}$$

where r_{01} (r_{10}) was the probability that the source state will change from 'OFF' to 'ON' ('ON' to 'OFF') between time slots, as shown in *Figure 6.2*. The value of r_{10} was chosen to give the required burst length in timeslots (equal to $1/r_{10}$), and the second transition probability was chosen to give the required average bit rate. The remaining parameter in the simulation model was the buffer capacity, which was chosen to be commensurate with the mean number of cells in a burst.

The simulation program was run for a sufficient number of time slots for at least ten lost cells to be recorded, giving a reasonable measure of confidence in the measured cell loss probability. However, in simulating a low-loss mode of operation, no cell loss was to be recorded in a simulation run time of practical duration. The maximum number of cells generated was limited to 10^7 and if no cell loss was observed, it may have been reasonably concluded that the CLP was below 10^{-6}, although formal calculation of the significance level of this outcome would have required a decorrelation technique such as batching to be applied.

Since the cell-scale simulation model had two additional parameters (the mean burst length, and the multiplexer buffer capacity) compared to the CAC algorithms, it allowed their robustness to be determined. It may have been expected that the CAC algorithms would produce conservative results for most choices of these parameters, since, in the worst case where the buffer capacity was low, and the mean burst length was long, high cell loss rates may have been expected even when the link utilisation is relatively low.

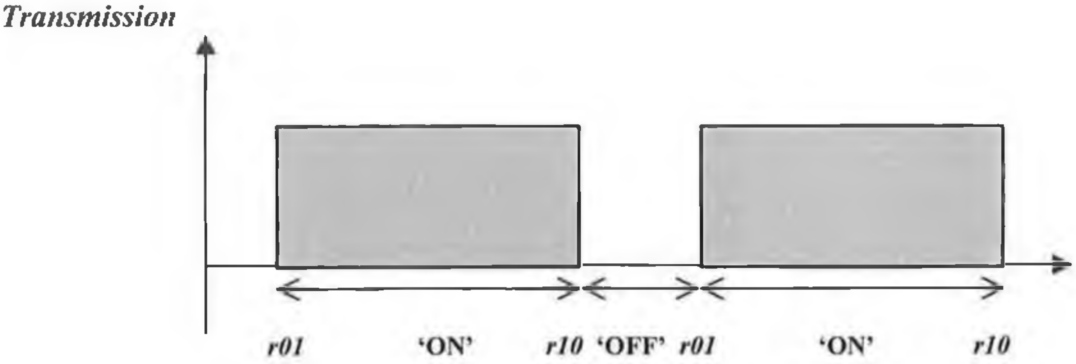

Figure 6.2 Transition Probabilities for 'ON' to 'OFF' and 'OFF' to 'ON'

6.5 Experimental Work

The performance of the algorithms for connection admission control was evaluated by comparing their predictions concerning CLP with the CLP measured by the cell-scale simulation. An effective algorithm should have admitted fewer connections than the cell-scale simulation and more than that of a peak rate algorithm. In this section the results are found from the numerical experiments which were performed with the four algorithms. First to be examined is the QoS criterion, which is the CLP, for the algorithm. The CLP was found for an increasing number of sources, and the experimental results plotted. These indicated the behavior of the algorithms with a single traffic source type. Next there were experiments with mixtures of heterogeneous sources. Three types of source characteristics were considered in these experiments, which are labeled Type I, Type II, and Type III. The properties of these three traffic types are shown in *Table 6.1*. The results are plotted in three-dimensions to display the admission regions for different mixtures of types of sources.

6.5.1 Homogeneous Sources

The Cell Loss Probability versus number of sources admitted by the algorithms was calculated for a single source type (Type II) and is shown in *Figure 6.3a, Figure 6.3b* and *Figure 6.3c*, for a link with capacity of 155Mbps. The target CLP can be less chosen to be 10^{-6} or better to meet Quality of Service requirements.

For the Convolution Algorithm displayed in *Figure 6.3a*, a sharp linear increase in readings indicated a reliability that was extended over a period. The graph then evened out. There is a mild increase around the subsequent almost horizontally linear reading. The Cell Loss Probability had a small rate of increment beyond that of the threshold value that can to be disregarded. The cell-scale simulation was also drawn, it clearly admitted almost twice as many sources as this algorithm, indicating how conservative it is. The cell-scale simulation was found from an experiment run over an extended period of time with bursty sources. The sources did not all run at the same time and so the admission could have been higher. The algorithm provided a quick estimate that was required to be within the available capacity. The results were compared with those of the other algorithms in *Figure 6.5*, where the Convolution Algorithm is shown to be a conservative approach compared with the other algorithms.

The Cell Loss Probability for the Chernoff Bound algorithm varied at the CLP limit in *Figure 6.3b*. The algorithm displays a fairly reliable gradual increase and the cut off point for CLP is clear. Experiments with smaller sources (1 Mbps PCR) in *[21]* found a slight fluctuation in CLP levels that may highlight a small margin of error. The cell-scale simulation again clearly admitted almost twice as many sources, so the algorithm is shown to have a conservative estimate. A comparison of CLP for the algorithms in *Figure 6.5* indicates that this is the most moderate approximation of the four algorithms. Chernoff Bound is used in conjunction with other less stringent approaches to provide a lower bound to their assessments *[20][21]*.

For the Gaussian Approximation Algorithm, the CLP is found from the graph in *Figure 6.3c* of the estimates for an incremental number of sources. There was a sharp decline corresponding to the aggregate PCR becoming equivalent to the link rate of 155Mbps. The graph then extends beyond this point, gradually increasing as shown in the diagram. The estimates are similar to that of previous algorithms but are more generous, allowing extra sources to be admitted. The results are compared to those of the cell-scale simulation and again found to be well within these findings. The Gaussian Approximation was used to develop the idea of 'equivalent capacity', *[33]*. The distribution of the stationary bit rate was approximated by a Gaussian distribution. The assumption was that the Gaussian distribution allowed the use of standard approximations to estimate the tail of the bit rate distribution. In particular it meant that the cumulative tail probability of exceeding a QoS value could be determined. These ideas lead us to those next of Effective Bandwidth.

The Effective Bandwidth approach calculated the equivalent capacity for sources rather than estimating cell loss. *Figure 6.3d* shows a steady, almost linear increase in required capacity as the number of Type II sources was incremented. The admission occurs with up to twenty six sources. The experiment calculated the equivalent capacity *[33]* for a given link rate and source type. The study in *[33]* checked the accuracy and investigates the limitations of this approach. They expressed concern with an over-estimation by the fluid flow approximation with

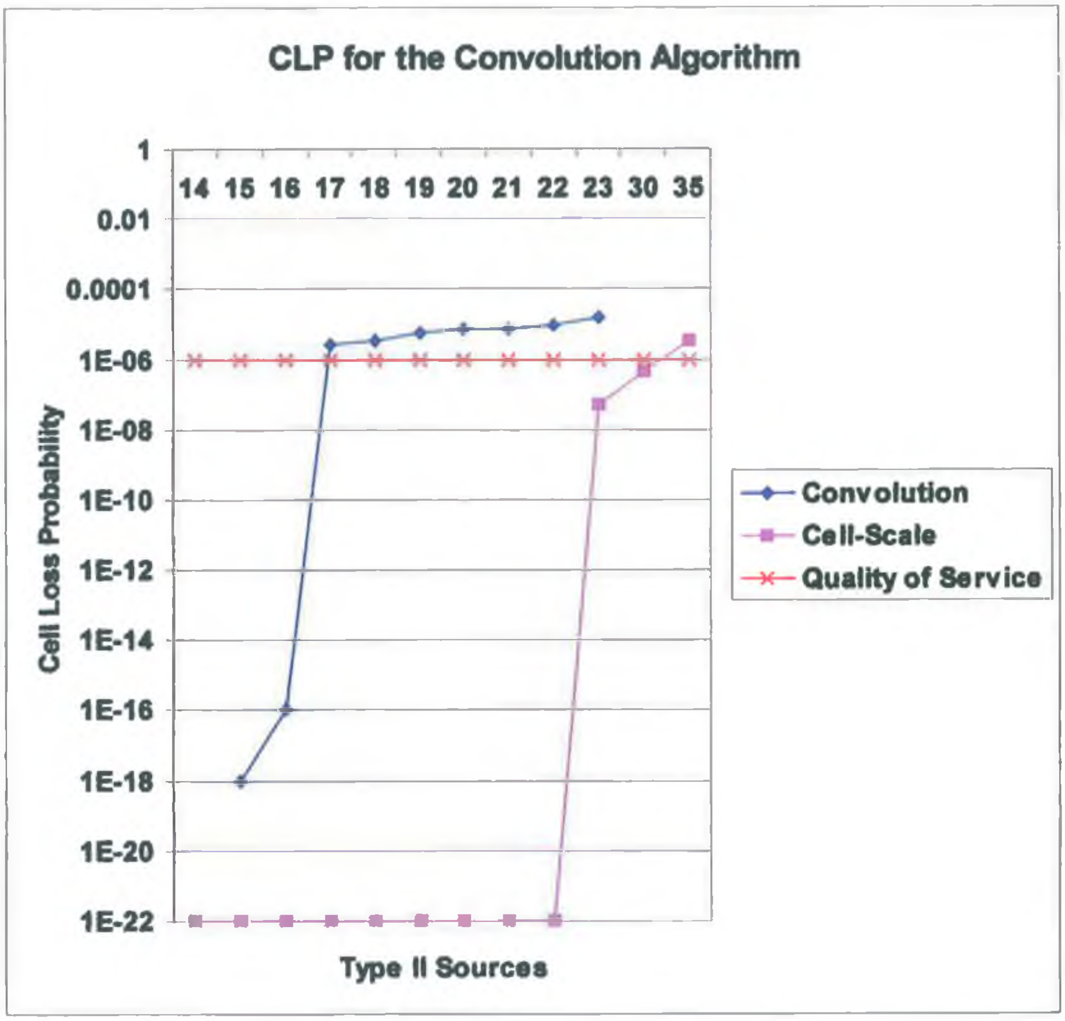

Figure 6.3a Cell Loss Probability in the Convolution Algorithm

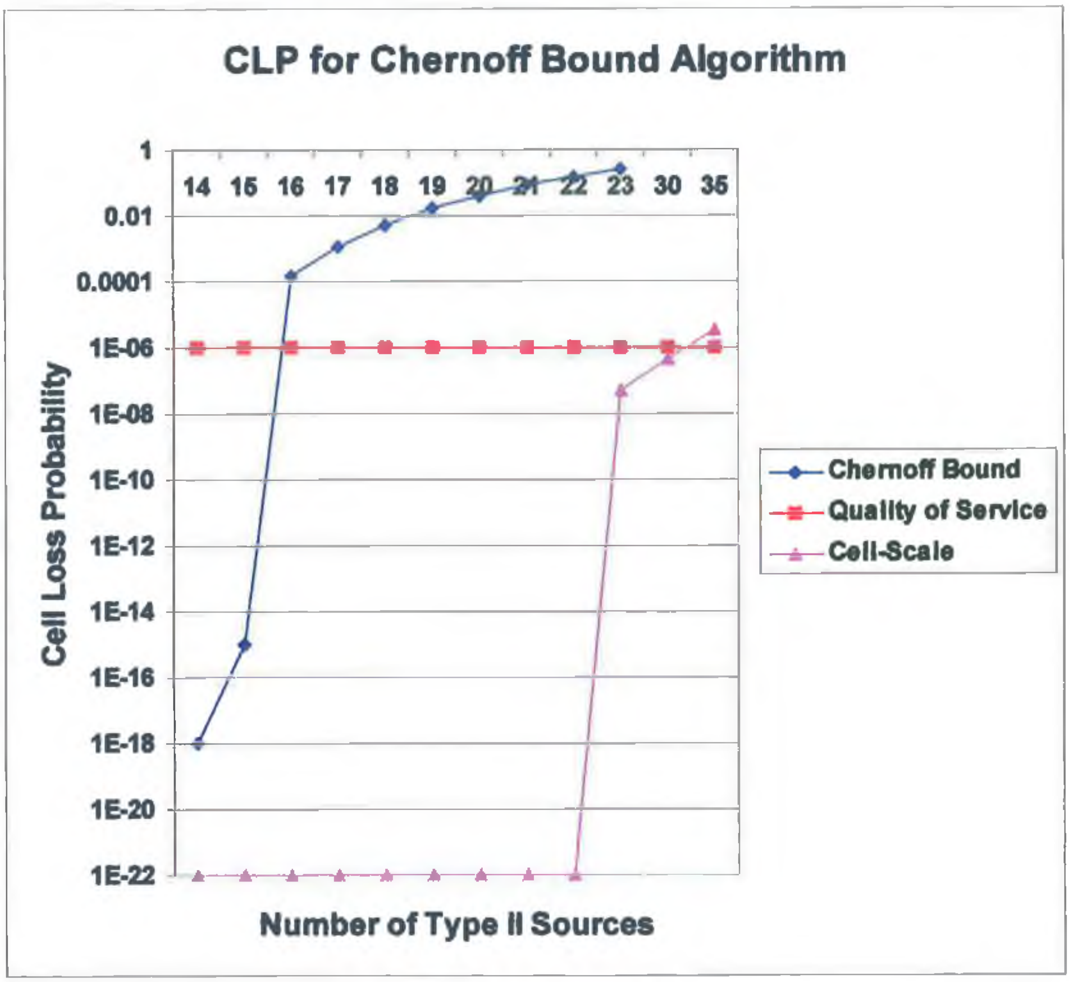

Figure 6.3b Cell Loss Probability for the Chernoff Bound Algorithm

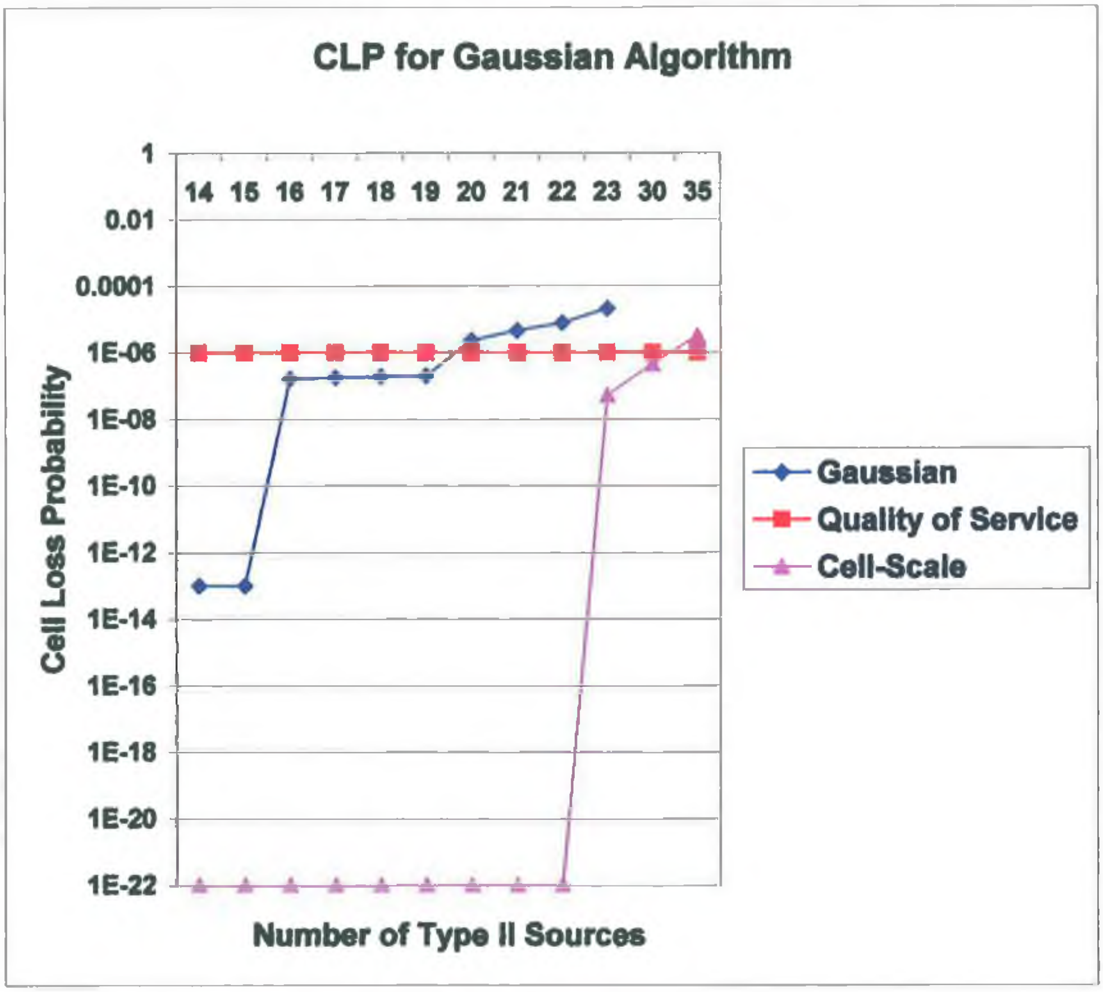

Figure 6.3c Cell Loss Probability for the Gaussian Approximation Algorithm

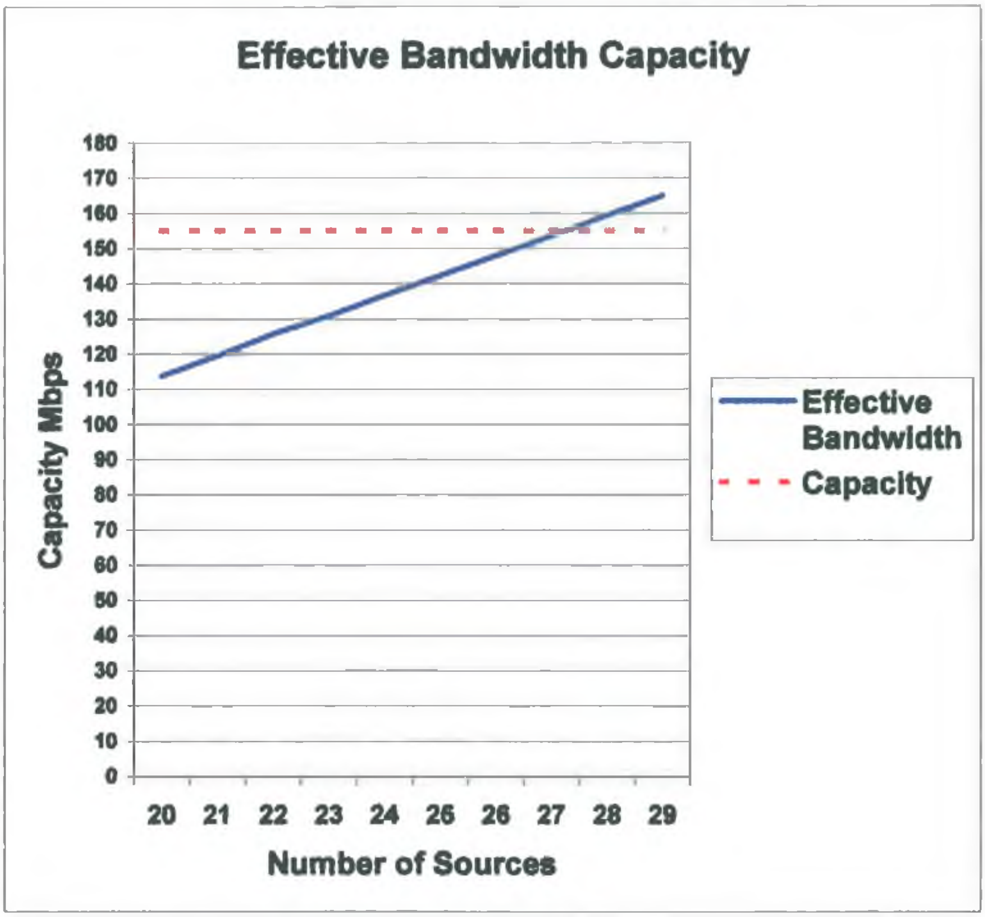

Figure 6.3d Capacity for Type II Sources with the Effective Bandwidth Algorithm with Link Capacity 155Mpbs

burstier traffic. The comparison with the cell-scale simulation results can be made to provide reassurance that the allocation of bandwidth is within limits. Also, these Effective Bandwidth results were very generous when compared to previous algorithms, showing how well the algorithm performed. The results are compared in *Figure 6.6* showing the admission regions, the Effective Bandwidth algorithm was far superior to the others.

6.5.2 Heterogeneous Traffic

The next set of examples investigated the case of non-identical sources. The experimental results with mixtures of traffic types revealed some variations in the responsiveness of the algorithm. This was a study of some of the aspects of the interactions between connections inside the network. There was an investigation into the potential impact of high-speed bursty connections on the bandwidth requirements of lower-speed ones. In the following scenario there were three groups with an incrementing number of connections in each. The experimental results were then used to create the admission regions displayed in *Figure 6.4a* to *Figure 6.4h*. We wished to study the potential changes in the bandwidth requirements of the connections of each type of traffic as a result of their interactions. In particular to investigate the significance of the 'gating' effect of high-speed bursts, which could modify the effective peak rate of a low-speed connection and therefore its bandwidth requirements. This 'gating' effect was caused by high-speed bursts, which when present forced the decrease in admission from the low-speed connection and requires a higher bandwidth allocation. There are significant dips in the graphs, the valleys and peaks reflect the 'gating' phenomenon.

6.5.2.1 The Convolution Algorithm

The first experiment involved a gradual increase of Type I and Type II with small increments of the two types of sources. The lower-speed source was Type III and was incremented for a corresponding fixed level of larger sources until the CLP constraint was reached. We see the effect on lower-speed traffic when the higher-speed traffic was incremented. The admission region in *Figure 6.4a* is a representation of the amount of sources that were admitted for the combinations of traffic types with the Convolution Algorithm. This is a very useful display when examining the algorithm for effectiveness with different traffic. *Figure 6.4a* shows admission regions for mixtures of three types of traffic sources for the described scenario. The efficiency of the Convolution algorithm was less with a larger input of Type I traffic, due to the 'gating' effect. The smaller traffic sources then plummeted significantly when mixed with the traffic types which required greater bandwidth. The graph shows the fluctuations when this occurred. It indicated that the algorithm was most efficient for an even amount of the three source types.

Next there was a similar trial that charted the admission of larger sources to the heterogeneous mix in *Figure 6.4b.* The Type II and III smaller sources were incremented for fixed amounts and then the admissible amounts of Type I sources were found. In *Figure 6.4b* the fluctuations were found to be far less predominant than that of the previous example. A series of minor dips and peaks followed the descent of the graph towards a more even mixture of types. The algorithm was most efficient with larger amounts of Type II sources, with a marked decline as the smaller sources were decreased. More of the larger sources (Type I) were admitted but the overall admission region was at a minimum. The findings suggest that a more even balance of source amounts proved to be the most effective approach for heterogeneous traffic.

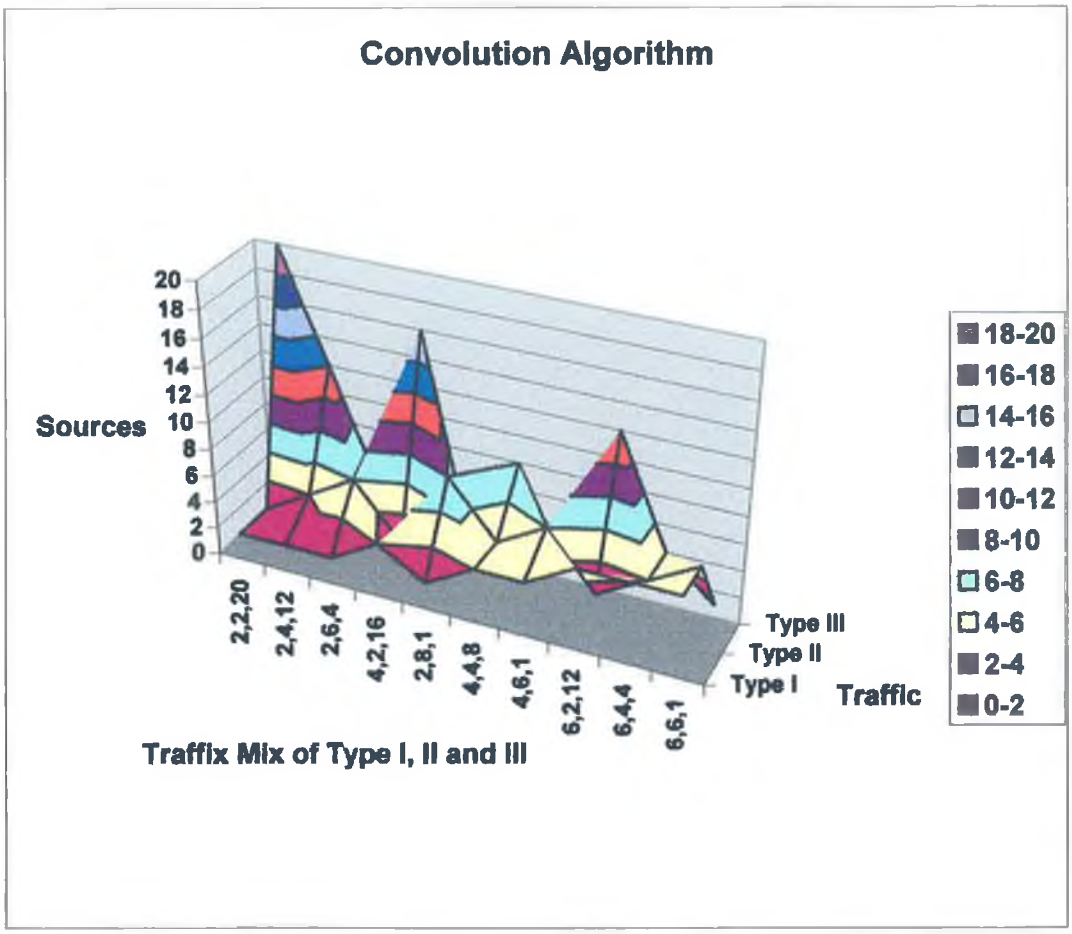

Figure 6.4a Admission Regions for Heterogeneous Traffic for the Convolution Algorithm Varying Type III Traffic

Note – The 'Admission Region' refers to the volume enclosed by the surface displayed above and is the number of sources admitted for the amounts of each traffic type. A vertical slice parallel to the z-axis will show the amounts of each traffic type admitted for the surface, as indicated by the x axis with the three different amounts shown. A fall in the surface means less sources are admitted, a peak in the graph indicates more sources are admitted for the criteria imposed, in this case the CLP.

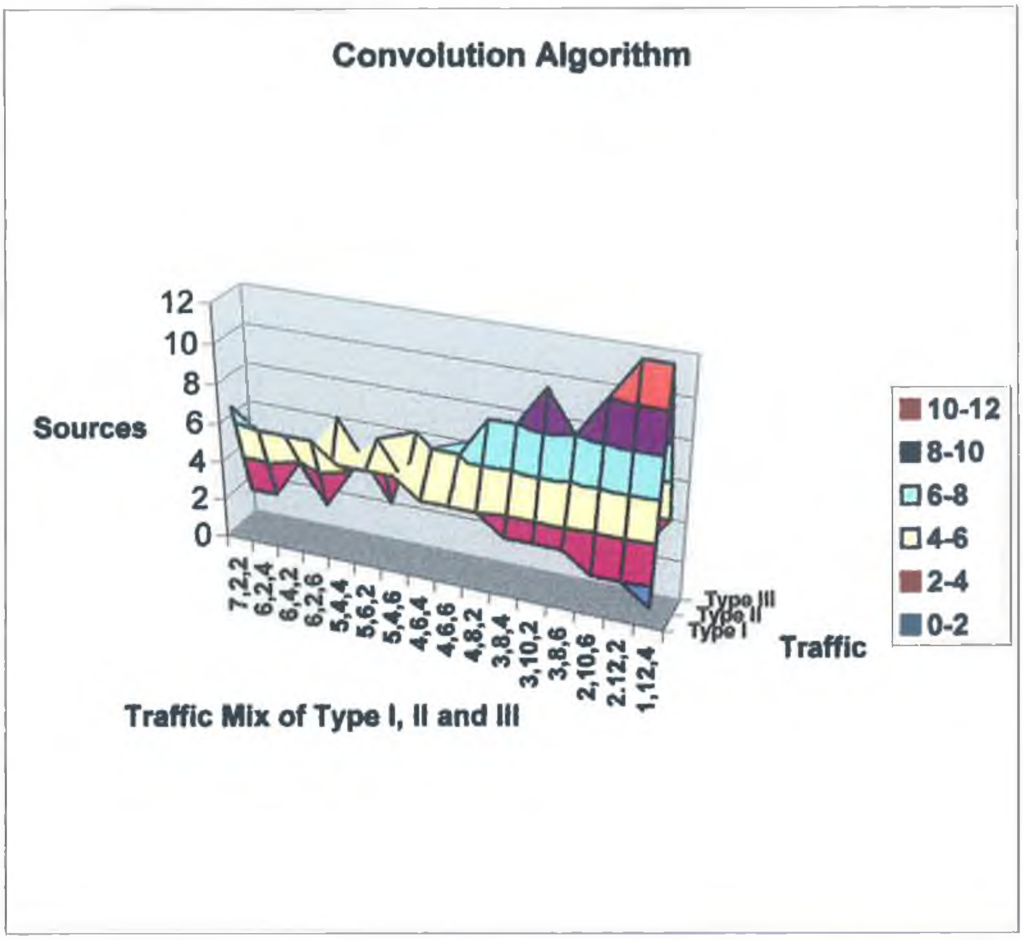

Figure 6.4b Admission Regions for Heterogeneous Traffic for the Convolution Algorithm Varying Type I Traffic

6.5.2.2 The Chernoff Bound Algorithm

The admission region for the Chernoff Bound Algorithm in *Figure 6.4c* shows the amount of sources that were admitted for a combination of three traffic types. The Type I and II (larger speed) sources were incremented and the number of Type III sources was found. The admission region was gradually increased with the number of sources of the traffic types. There was a steady increase with the bigger source types and the third traffic type falls off. The far peak in *Figure 6.4c* to the left shows that the admission of the smaller sources fell dramatically when the larger sources were predominant. This is again due to the 'gating' effect caused by high-speed bursts, which means a higher bandwidth allocation was required and the graph dips. These low points on the graph in *Figure 6.4c* indicate that the larger sources don't 'squeeze in' well with a predomination of smaller sources. There are two small peaks in the graph indicating that the more even amounts of the sources provide the most effective combination of admission. The CPL graph *Figure 6.3b* shows a fluctuation that is mirrored in the traffic mix in *Figure 6.4c*. The algorithm is slightly less efficient when a more refined estimation was required, such as almost all larger sources with a few of Type III. For more even mix of sources the admission region peaks. There was a similar result for a heavier mixture of Type I and Type II. The admission regions were compared with that of other algorithms in *Figure 6.6*.

The experiment was repeated (as shown in *Figure 6.4d*), this time for the admission of larger Type I sources. Type II and III (with PCR of 10Mbps and 5Mbps respectively) were incremented slowly. Readings for the corresponding number of Type I (PCR 20Mbps) sources were recorded. The admission region is a more even presentation than that of the previous experiment varying smaller sources, but with a marked decrease in size. The 'gating' effect is less obvious, with the graph sloping down to admit a greater number of larger sources. The algorithm is conservative and this is manifested in the display's mild peaks and troughs. The algorithm improves in admission when the source types were more evenly mixed. This reached a maximum with the increase in Type II, the middle range source. Overall, the results mirrored those demonstrated by the CLP displays, with a marginal degree of improvement overall shown with the heterogeneous traffic with increasing Type III sources. The algorithm was seen to be the most conservative estimate as suggested by the studies in *[9]*.

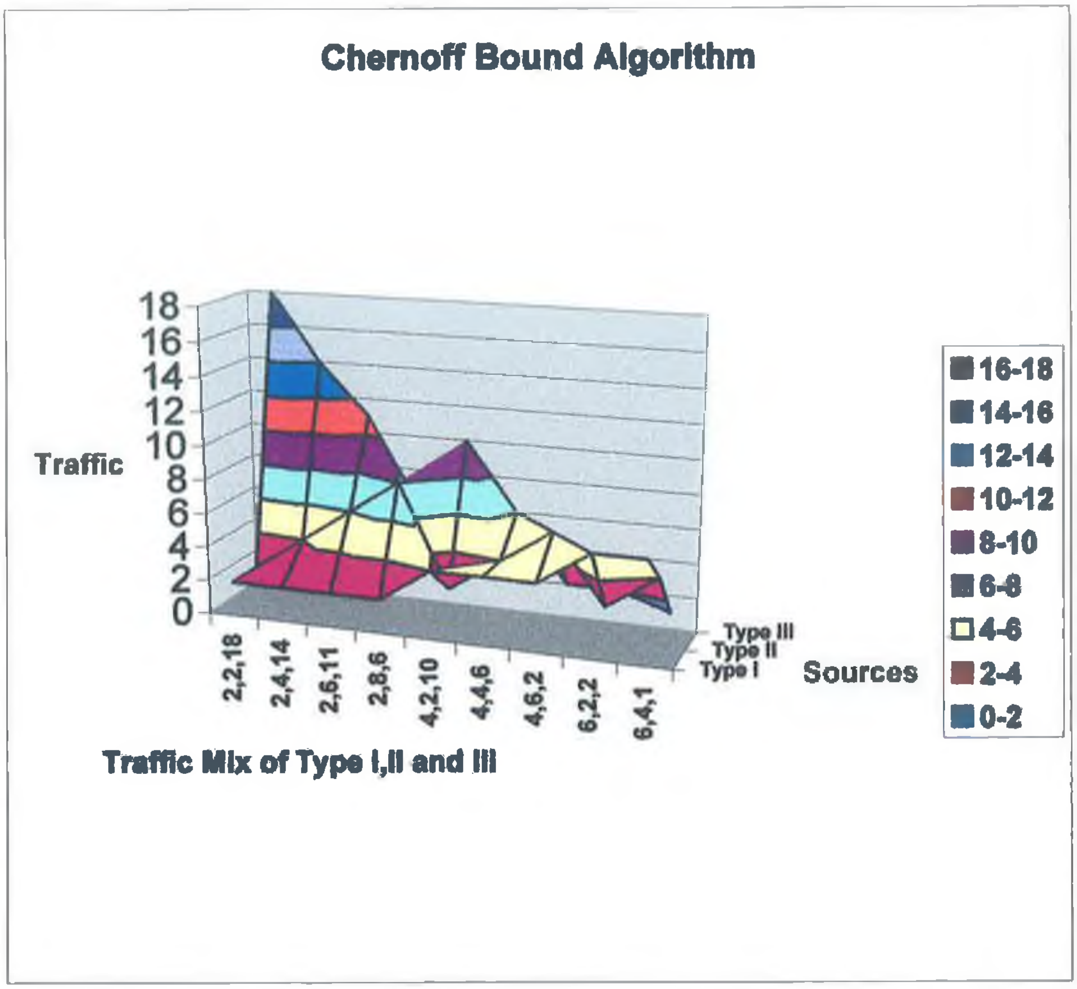

Figure 6.4c Admission Regions for Heterogeneous Traffic for the Chernoff Bound Algorithm Varying Type III Traffic

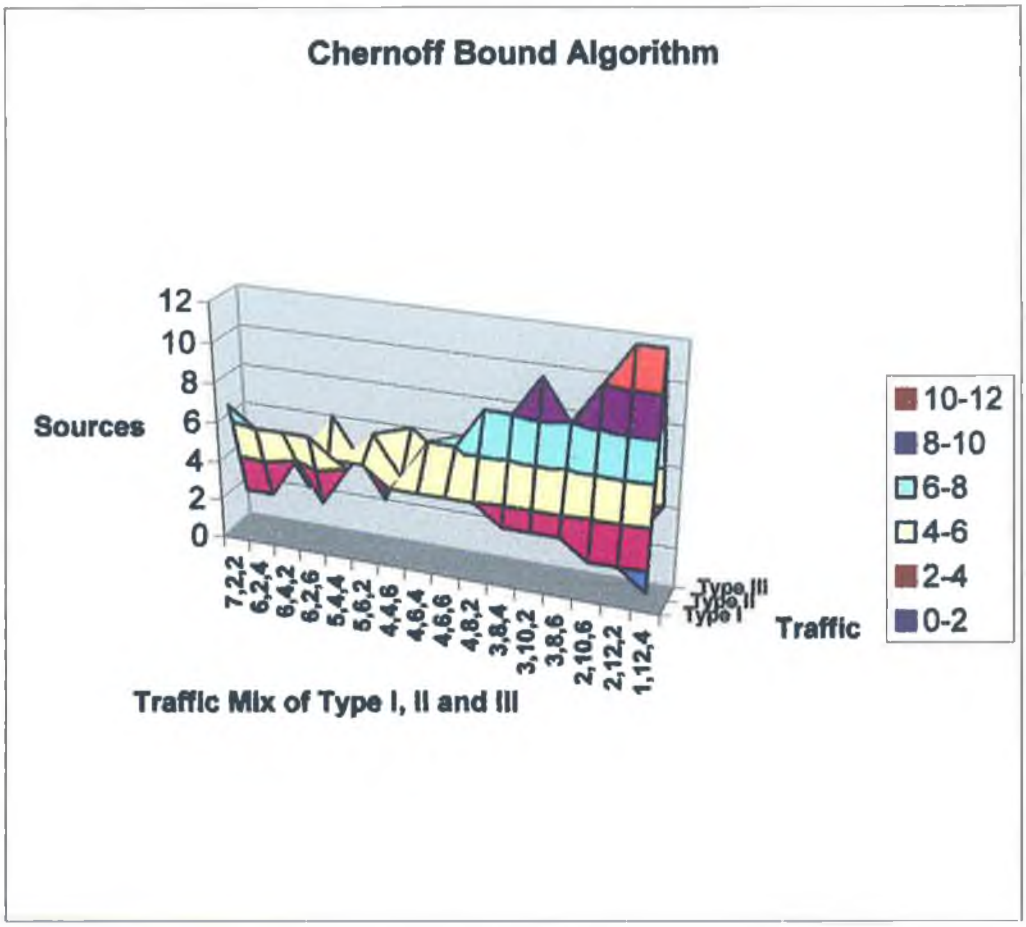

Figure 6.4d Admission Regions for Heterogeneous Traffic for the Chernoff Bound Algorithm Varying Type I Traffic

6.5.2.3 The Gaussian Approximation Algorithm

For the Gaussian Approximation algorithm, the outline of the inter-relationship between the admission regions of traffic types displayed variation in the amounts of different sources in *Figure 6.4e*. Type III source admission was found for increments of Type I and Type II. The gradual descent of the graph shows the admission for an increasing mix of source types. The top peak in *Figure 6.4e* is almost all Type III smaller sources. The admission region is very similar to that of the Convolution algorithm. This indicates the trend in the bandwidth requirements of the low-speed connections as a result of their interactions with high-speed. The 'gating effect' caused by Type I bursts resulted in a higher bandwidth allocation, as seen in previous experiments with other algorithms. This effect is seen in the steady decline of admission region as the higher-speed Type I and II sources gain predominance. In contrast to the previous algorithms however, the Gaussian Approximation algorithm shows a marked degree of efficiency with a greater number of larger sources. The 'gating' effect is less pronounced and there is a larger admission region. The algorithm is still somewhat less effective with a mixture of sources than with only one source type.

Then the admission of Type I traffic for increments of Type II and III was found. For the Gaussian Approximation Algorithm the admission region was overall higher than previous experiments with smaller traffic sources, with a myriad of peaks and troughs, *Figure 6.4f.* The higher points correspond to the more even allocations of source types. A larger admission of smaller sources seemed to cause a blockade of sorts and results in a trough. The Gaussian Approximation algorithm relies on the Central Limit Theorem that states that the distribution of aggregate traffic converges to a Gaussian distribution as the number of connections approaches infinity. What we see is more likely to be a cruder version of the 'gating' effect as mentioned earlier. The algorithm is shown to be more generous than previous algorithms in the allocation of bandwidth, particularly when the heterogeneous traffic mix contains larger sources.

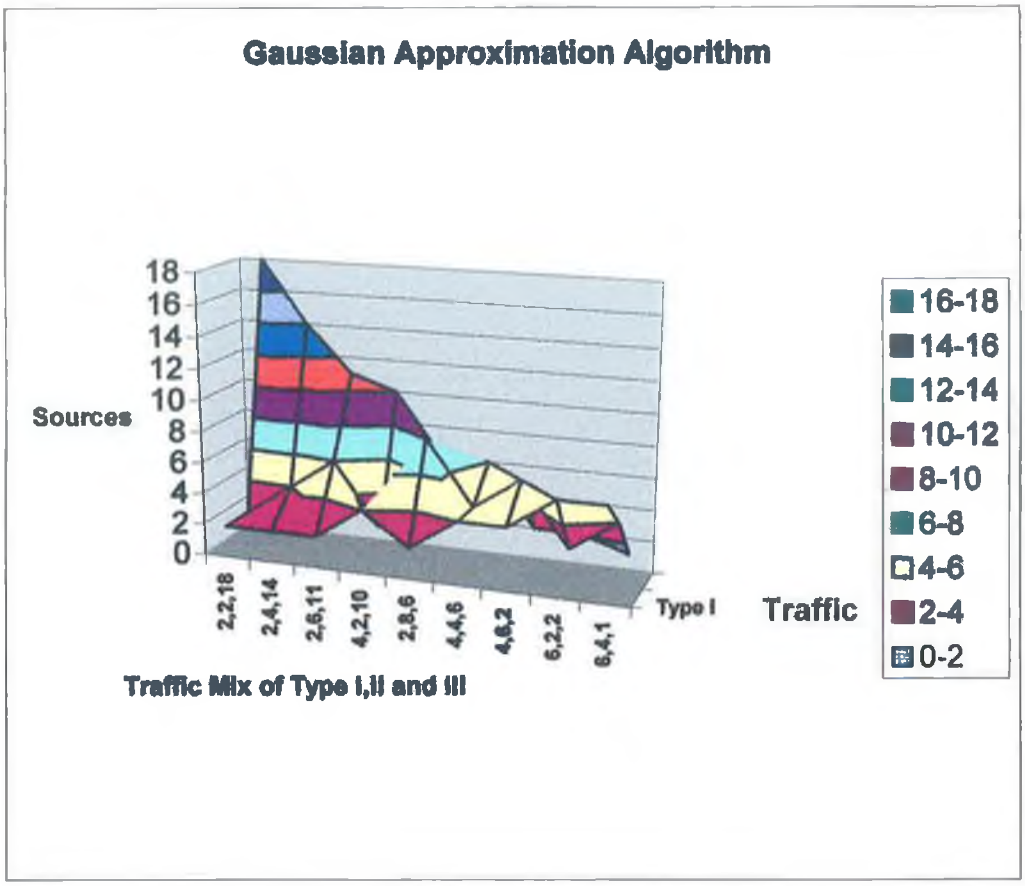

Figure 6.4e Admission Regions for Heterogeneous Traffic for the Gaussian Approximation Algorithm Varying Type III Traffic

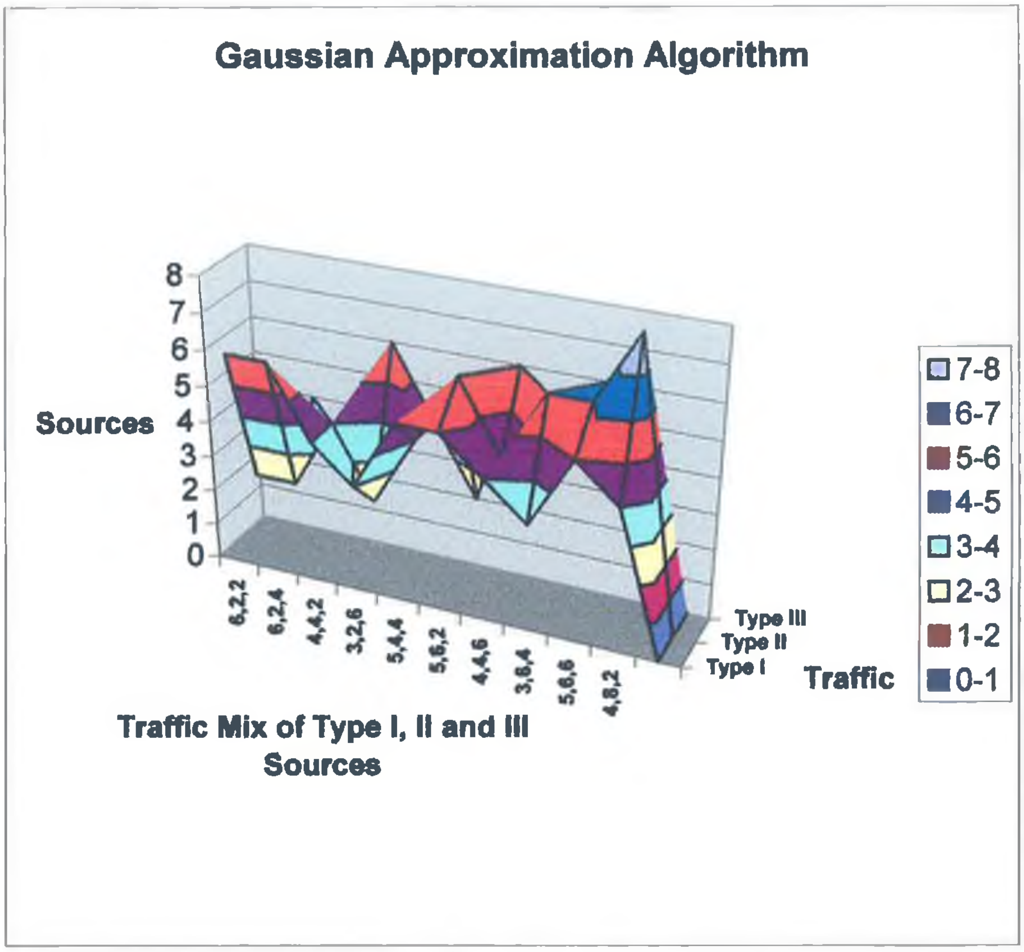

Figure 6.4f Admission Regions for Heterogeneous Traffic for the Gaussian Approximation Algorithm Varying Type I Traffic

6.5.2.4 The Effective Bandwidth Algorithm

The behavior of the Effective Bandwidth algorithm with a mixture of traffic source types is displayed, with the gradual increment of Type I and Type II and finding the corresponding Type III admissions. The experimental results were then used to create the admission regions in *Figure 6.4g*. We can see how the bandwidth requirements changed for the low-speed connections as a result of their interactions with high-speed traffic. The slope of the graph is gradual. The 'gating' effect is no longer apparent. This is because the Effective Bandwidth is independent of traffic submitted from other sources, hence there are no significant troughs as with previous algorithms. Notice how the admission region admitted more than the other algorithms, these are directly compared in *Figure 6.6*. The predominance of smaller Type III sources shows that the admission regions may be bigger at the start of the graph. The flexibility of sources with a lesser bit rate is striking. The admission region fluctuates slightly as the mixture of sources evens out. The combination of heavier Type I and II sources is displayed on the

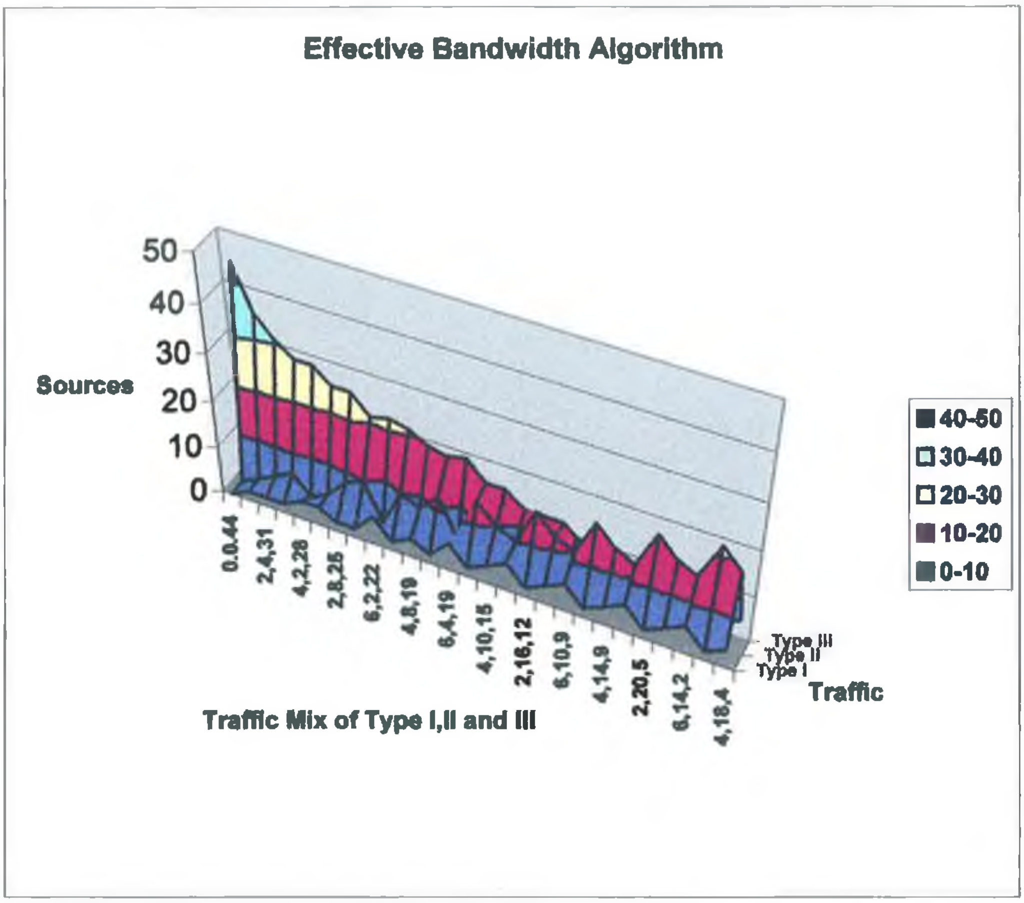

Figure 6.4g *Admission Regions for Heterogeneous Traffic for the Effective Bandwidth Algorithm Varying Type III Traffic*

right hand side of the graph. The volume enclosed is similar to that of mostly Type III sources on the extreme left. The best admission occurs when the mixtures were even. A majority of Type I traffic sources are shown with each smaller peak, with an even regularity. This indicates a consistency of admission with different levels of traffic mixes, which is an indication of superiority to the other algorithms presented.

For the Effective Bandwidth Algorithm, Type I was admitted for set amounts of smaller sources, as explained and is shown in *Figure 6.4h.* The results were quite different to those found with variations in the numbers of smaller sources admitted. The gradual increase in the smaller Type II and III traffic reduced the admission region. That is the performance dips for a more even mix of traffic. The peaks occur when the larger sources predominated the amounts. With a small amount of the slower sources there is a peak, then the admission region declines. It expands again as the Type II traffic increases. The admission policy is based on the estimation of required bandwidth capacity rather than cell loss. This has an impact on the efficiency of allocation for large sources. The graph is different to those of the previous examples. The direct comparison of admission regions is made in *Figure 6.6*, where the Effective Bandwidth algorithm displays a marked degree of superiority.

6.6 A Comparison of Algorithms

These studies describe experiments with simulations to demonstrate four of the CAC algorithms. The algorithms admission regions are assessed and the results demonstrate representations to indicate the significant findings of the trials. The cell-scale connection admission control simulations were run and produce idealised results. They were run without time constraints or consideration of implementation issues. The results are displayed in Figure 6.5 to compare the Cell Loss Probability of the algorithms. The ideal admission control simulation experiments can be compared with those of the CAC algorithms, this is discussed in the coming section. Then the admission regions of the four algorithms are mapped. They may be directly compared with the graphical representation in Figure 6.6.

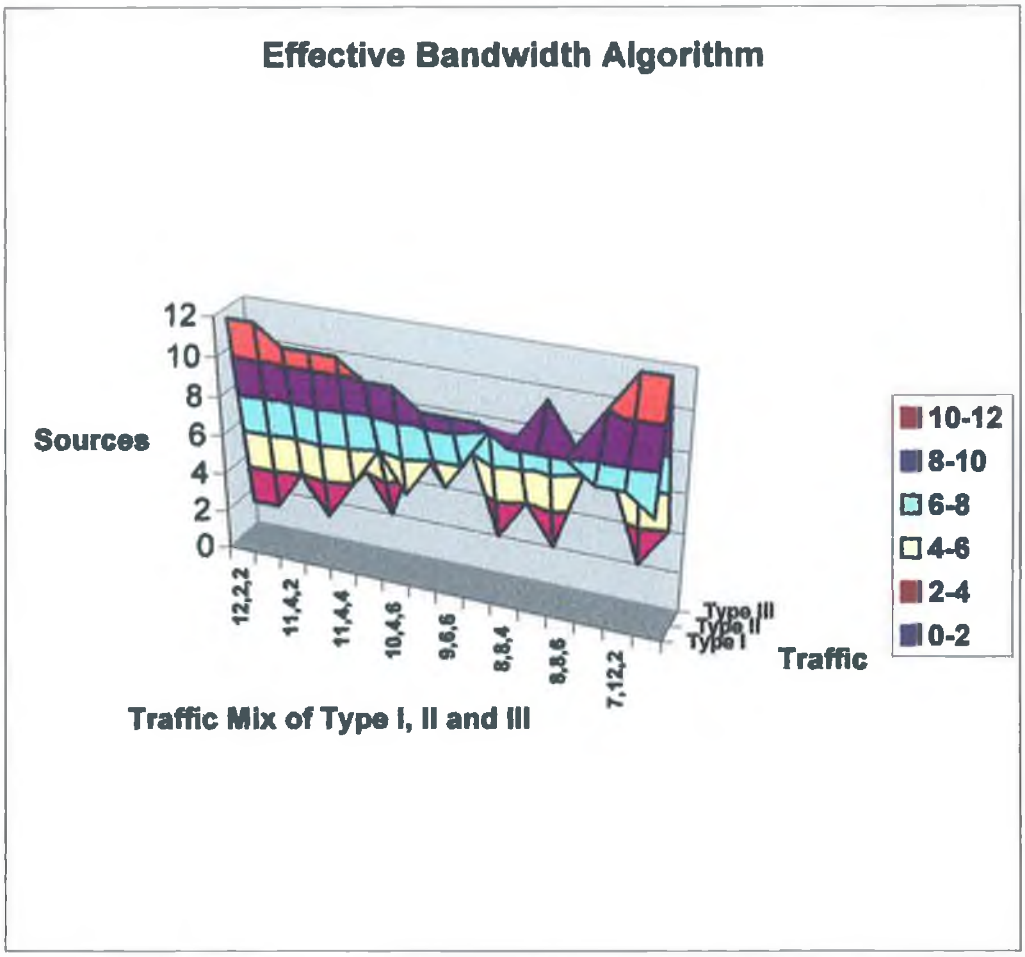

Figure 6.54h *Admission Regions for Heterogeneous Traffic for the Effective Bandwidth Algorithm Varying Type I Traffic*

6.6.1 A Comparison with Quality of Service Constraints

The results of numerical evaluation produced the graphs of CLP for the first three algorithms and also the network simulation admission control. The algorithms were compared and assessed relative to the ideal cell-scale simulation of CLP over time in *Figure 6.5*. As expected, the algorithms were found to be far more conservative than the simulation results. At a QoS level of CLP of 10^{-6} the algorithms admitted almost half that of the cell-scale approach. The algorithms must be conservative to stay with in allocation limits. Studies in *[33]* indicated that the bursty nature of traffic and the effects of statistical multiplexing provided the explanations. A word of caution was voiced in *[9]* that the assumptions concerning the nature of traffic itself may be flawed, and it may have been more self-similar in nature.

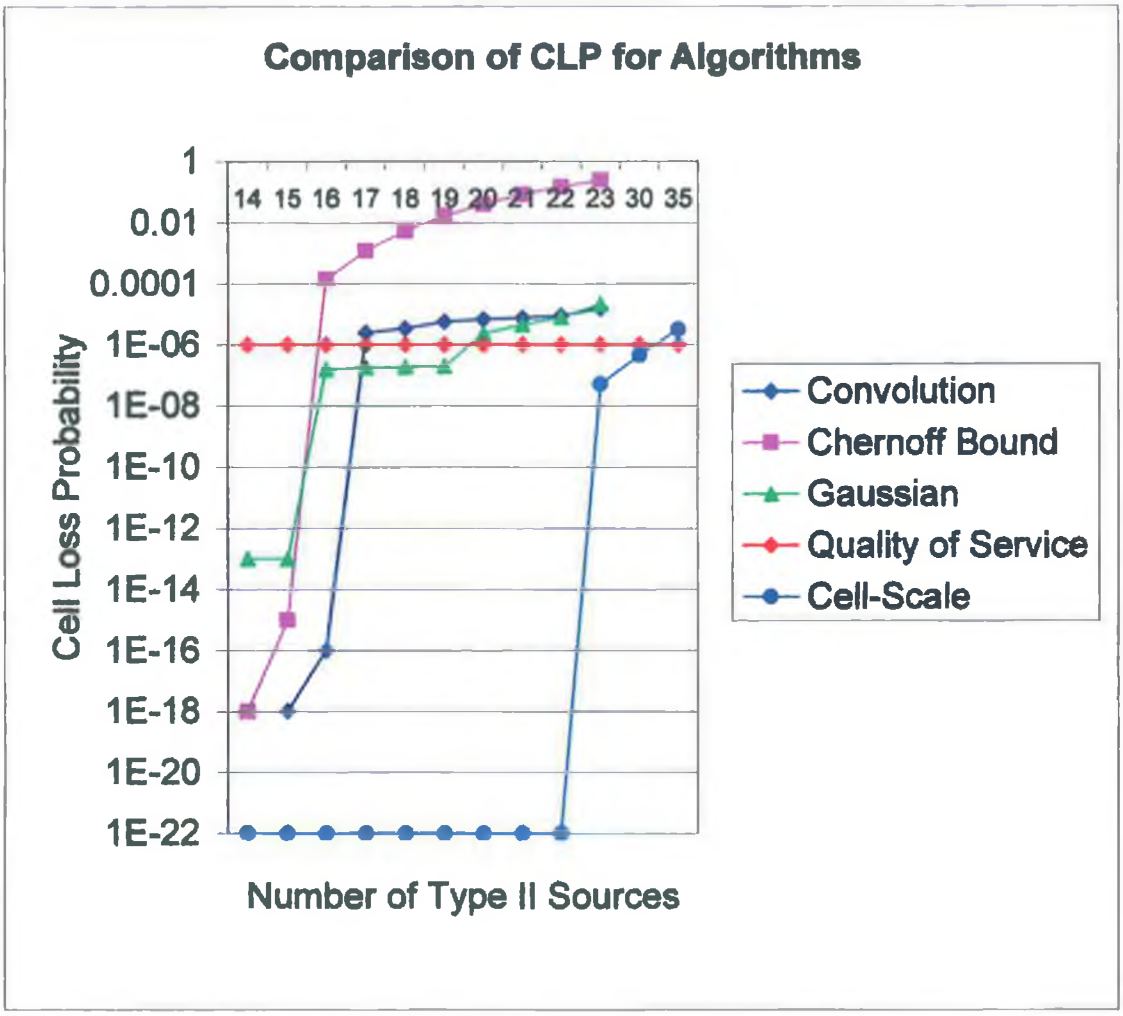

Figure 6.5 ***A Comparison of CLP for the Algorithms***

The Chernoff Bound is the most conservative. The Convolution algorithm displays a similar graph with more lenient admission. The Gaussian Approximation algorithm is shown to be the most generous in Figure 6.5. Compared to the cell-scale experiment however, all the algorithms are very stringent. In the next section we shall see how the Effective Bandwidth algorithm out-performed the others and was closest to the cell-scale simulation readings.

6.6.2 A Comparison of Admission Regions

The admission regions were plotted in *Figure 6.6* and comparisons were made from this graph. The Effective Bandwidth Algorithm was clearly the most effective algorithm. The graph peaked for Type II sources and the admission region for all traffic types was much larger than the three other algorithms. This validated recent work *[30]-[40]* which contained similar numerical evaluation and showed how the Effective Bandwidth algorithm was a better approach for resource allocation. The Chernoff Bound algorithm was the most conservative estimation with the smallest admission region for traffic. The Convolution and Gaussian Approximation Algorithms had similar admission regions. These results were consistent across a range of traffic types, as shown in *Figure 6.6*. They also validated the Cell Loss Probability study as shown in *Figure 6.5*. Similar studies in *[20]* combined the use of the Chernoff Bound and Gaussian Approximations. The admissible call region was found to be concave but becoming more linear with increasing values. These were contrasted with numerical results for equivalent capacity in *[33]*, where the observation was made that the stationary approximation results in a substantial overestimation for a small number of sources.

6.7 Conclusions

The basic objective of bandwidth management and traffic control strategy was to allow for a high utilisation of network resources, while sustaining an acceptable Quality of Service for all connections.

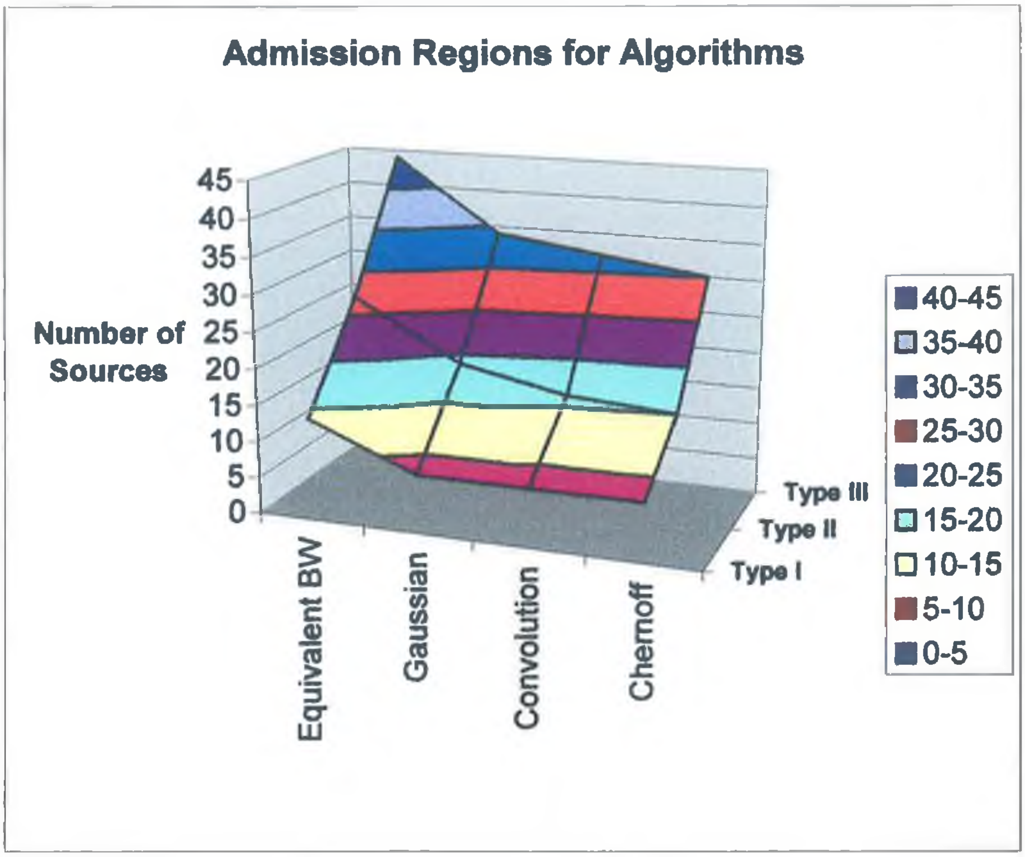

Figure 6.6 A Comparison of Admission Regions for the Four Algorithms

In this Chapter experiments with simulations were described, and the results were presented to demonstrate four of the CAC algorithms. There was also an experiment for cell-scale readings for admission to the network without time constraints or implementation issues. This provided a useful comparator to assess the admission for the CLP Quality of Service constraint. It was displayed with the algorithms for the parameter CLP and contrasted. There was also a display of the admission regions for a mixture of three different traffic source types for each algorithm. The algorithms admission regions were assessed and the results were demonstrated with graphical representations indicating the significant findings of the trials. Finally, the admission regions of the four algorithms were directly compared in Figure 6.6 to see which algorithm is the best.

The first experiments were to examine each algorithm's behavior with respect to the QoS constraint. This was a CLP chosen as 10^{-6}. The Convolution Algorithm showed a sharp drop in Cell Loss Probability for a smaller number of sources, with a leveling off around the admission region boundary. For the Chernoff Bound algorithm the CLP fluctuated a little at the boundary level. The Gaussian Approximation Algorithm had a broader base of potential readings of CLP. In contrast, the Effective Bandwidth algorithm provided a computationally simple approximation for the equivalent capacity or bandwidth requirement of a connection based on its statistical characteristics.

The next set of experiments investigated the case of a heterogeneous mixture of sources. It attempted to study aspects of the variations in the responsiveness of the algorithm and the interactions of different traffic types. It looks at the impact of high-speed bursty traffic on the bandwidth requirements of lower-speed traffic types. The experimental results were then used to create the admission region displayed in *Figures 6.4a* to *Figures 6.4g*. With the first three algorithms we investigated the significance of the 'gating' effect of high-speed bursts that forces a backlog. This higher 'effective' peak rate within the network can in turn require a higher bandwidth allocation, shown as the graphical representation forms a trough.

Interesting comparisons were then drawn between the studies of admission regions for each algorithm. The admission regions for each algorithm were presented in *Figure 6.6*. The effectiveness of each algorithm was mapped, it showed the admission regions for each. The Effective Bandwidth algorithm showed much wider admission region. The efficiency of the algorithm was striking. The graph dipped down to the Chernoff Bound algorithms readings. There was a steady similarity between the Convolution and the other algorithms. Numerical evaluations in *[30]-[45]* showed how the Effective Bandwidth algorithm was more effective for resource allocation.

The cell-scale comparator produced a set of results for the simulation of 'ON-OFF' sources for an extended length of time. Unhampered by time constraints or implementation considerations, an assessment of CLP was made for traffic and compared directly with the algorithms. The results showed that it admitted almost twice that of the algorithms, and so the four algorithms were very conservative. There are several explanations, the most important one was that the algorithms were designed to be within safe limitations, regardless of traffic load or type.

6.8 Summary

The Chapter presented four algorithms and the numerical evaluation of each. There was also a cell-scale simulation with a fluid flow model with exponentially distributed 'ON-OFF' sources. This acted as a comparator as the admission levels were found for an 'ideal' system without the time constraints or implementation issues of the algorithms. It admitted almost twice as many sources as the first three algorithms, indicating how conservative these algorithms are. It also validates the Effective Bandwidth approach that allowed for significantly more sources than the other algorithms. The experiments were extended to look at heterogeneous traffic. The 'gating' effect and the corresponding fluctuations are noted, and each algorithm displayed its own variations.

Chapter 7

Conclusions

7.1 A Review of CAC Algorithms

Asynchronous Transfer Mode (ATM) is a recommended transfer mode for the introduction of broadband services, and it is capable of integrating services as diverse as broadcast television and video. Integrating these services with a common platform brings a number of benefits, the most important being increased efficiency. One aspect of efficiency is statistical multiplexing, which offers potential gains yet prediction of future traffic may be difficult. The critical value of the Connection Admission Control is that it can allow for increased efficiency of link utilization and resource allocation, while still providing the required Quality of Service. Other evaluation priorities are network utilization and implementation and operational costs.

The exploration of algorithms for connection admission highlights interesting implications for resource allocation in ATM networks. The CAC plays a vital role in the management of these resources, allowing the most efficient use of bandwidth together with the most effective solutions. The review of algorithms in Chapters 3 demonstrates the wide range of approaches. Chapter 4 is devoted entirely to the Effective Bandwidth approach. The Measurement-Based approach is presented in Chapter 5. The series of numerical evaluations display the algorithms in Chapter 6, showing how different algorithms perform relative to each other. The overall question throughout is how the allocation of bandwidth for connections can be minimized while still meeting their QoS requirements.

7.2 Different Approaches for Connection Admission Control

There are two main approaches to admission control. First is the parameter-based approach that computes the amount of network resources required to support a set of calls from pre-defined traffic characteristics. The second is the Measurement-Based approach, which relies on the measurement of actual traffic in making admission decisions. A discussion by Duffield et al *[42]* contrasts these two processes, and presents an alternative to the modeling of the arrivals process with parameters. The large deviation rate-function of the arrivals process is used to estimate the QoS requirements directly, from the 'entropy' of the traffic streams. Measurement-Based algorithms *[10]-[13]* study the performance of a scheme that has no prior knowledge of the traffic statistics and makes the admission decision on the current network state only. In contrast to the other algorithms which look at the characteristics of source traffic and represent them as parameters, Measurement-Based algorithms make decisions on a monitored amount of traffic on the network. The main theoretical result is that for large-link capacities with separation of call and burst timescales, Measurement-Based algorithms can achieve the performance of an optimal scheme with knowledge of traffic statistics.

7.3 The Numerical Evaluation of the CAC Algorithms

The various CAC techniques all have objectives of achieving maximum link utilisation with QoS guarantees. These are balanced with estimations of computational complexity and real-time implementation issues. The numerical evaluation in Chapter 6 describe experiments with simulations to display and compare the CAC algorithms *[9][24]-[30][33]-[38][42]-[45]*. Exploration of the algorithms by simulation focuses on the Quality of Service parameter CLP and admission regions for different types of traffic.

A set of input parameters represents the arrivals process for the aggregate sum of sources, to demonstrate four of the admission control algorithms in different ways. These algorithms are displayed with the graphical representations that indicate the significant findings of the trials. The findings of the simulations studies for a number of algorithms give an interesting summary of some of the approaches to connection admission control. They give a graphical representation of the admission region for each algorithm to give an indication of the level to which call acceptance extends. The admission regions for each can be compared. The Effective Bandwidth algorithm was shown to be the most efficient. The comparisons drawn from each study validate recent work *[30]-[43]* which claims that the Effective Bandwidth algorithm is a better approach for resource allocation.

7.4 Directions and Further Work

A variety of algorithms used for connection admission control have been presented, along with a number of possible general approaches to the problem of bandwidth allocation in ATM networks. Early, groundbreaking work by Hui *[20]* and Guerin et al *[33]* presented the ideas of time-scaling and equivalent capacity. The use of large deviations theory was introduced, and the derivation of Effective Bandwidth was developed by Chang, Thomas *[35]*, Mitre et al *[32][34]*, Kelly, Gibbens *[36][37]* and other authors. It provides an interesting basis for the calculation of resource allocation, and the study in Chapter 4 and the Appendix provides some details of the application of this technique for Connection Admission Control.

The directions of further work are first the refinement of techniques from the large deviations theory for the Effective Bandwidth approach, and secondly the use of Measurement-Based algorithms. The use of Artificial Intelligence and neural networks is constantly developing, and finally there is the prioritisation of network traffic. Measurement-Based algorithms *[12][13]* study the performance of a scheme that has no prior knowledge of the traffic statistics in contrast to the other algorithms that look at the characteristics of source traffic and represent them as parameters. The third area of recent work looks at the potential of Artificial Intelligence and neural networks fuzzy logic approaches *[46]-[58]* to be applied to solve the many demands required of the CAC, particularly for multimedia services with bursty traffic. It is possible to integrate a variety of services with different Quality of Service requirements by classifying service types according to different priorities. This approach can be combined with others to develop a new sophisticated connection admission control.

Many of the techniques described for the CACs are also applicable to other networks such as Multiprotocol Label Switching (MPLS) and those supporting the Internet Resource Reservation Protocol IntServ/RSVP. The traffic behavior in such networks will differ from that in ATM networks because of differing packet formats and flow techniques, but the underlying principles of the admission control will be the same. The design and development of networks and new protocols continues and improves, to provide for an expanding array of broadband services.

Future work in the study of connection admission control has many important issues to consider, such as the nature of resource allocation and statistical multiplexing, and the integration of different types of services and the prioritization of calls. The ability to guarantee multiclass QoS for different types of services involves the mapping of user requirements to traffic parameters, and setting up the subsequent compliant connection. This is an area of intense interest and is leading to useful studies with a variety of queuing models. The input from other related areas of research such as traffic modeling should provide fruitful benefits. An example is the study of the self-similar nature of bursty traffic. So the continuation of improvements and new ideas in this area of research has wide-ranging implications for the critical issues of resource management and service provisioning in high-speed ATM networks.

Appendix

Effective Bandwidth

The following sections present the theory of Effective Bandwidth and equivalent capacity for statistical multiplexing *[11][28]-[45]*.

A.1 Defining Effective Bandwidth

The allocation of bandwidth for statistically multiplexed sources needs to meet Quality of Service requirements while targeting good link utilization. This leads to the idea of *'Effective Bandwidth'*. Since the likelihood that all sources will transmit at peak rate all the time is small, the allocation is less than the bandwidth required for the peak rate of the sources. Effective Bandwidth theory allows for the derivation of bandwidth allocation techniques from the behavior of individual and aggregate sources.

The concept of Effective Bandwidth is used to describe the utilisation of network resources in terms of the statistical characteristics of the sources and their Quality of Service requirements. It provides a measure associated with the source for performance guarantees expressed in terms of loss or delay. The CAC is simply a consideration of the sum of Effective Bandwidths to be less than a threshold.

Kelly and Gibbens *[36][37]* state the definition of Effective Bandwidth of a source as depending on two parameters, the space and time scaling. The choice of these time scales depends on the characteristics of the resource, capacity, buffer size, traffic model etc.

The Effective Bandwidth is given by the statistical descriptor:

$$\alpha(s,t) = \frac{1}{st} \log E[e^{sX[\tau,\tau+t]}] \qquad (A.1)$$

where s is the space scale (in bytes or cells) and t is the time scale (in seconds). $X[\tau, \tau+t]$ is the workload arriving at a resource in time period $[\tau, \tau+t]$ and the expectation is taken over the distribution of random periods. This means that $\alpha(s,t)$ lies between the mean and peak arrival rates of the source measured over an interval t, hence the improvement for link utilisation if the Effective Bandwidth can be allocated instead of the peak rate bandwidth requirement. The definition of Effective Bandwidth for $X[0,t]$ is the amount of work that arrives from a source in the interval $[0,t]$. Assume that $X[0,t]$ has stationary increments, the Effective Bandwidth of the source is defined as:

$$\alpha(s,t) = \frac{1}{st}\log E[e^{sX[0,t]}] \quad \textit{for} \;\; 0 < s,t < \infty \qquad (A.2)$$

with the following properties:

(i) If $X[0,t]$ has independent increments, then the Effective Bandwidth $\alpha(s,t)$ does not depend on t.

(ii) If the random variable X exists such that $X[0,t] = Xt$ for $t > 0$ then $\alpha(s,t) = \alpha(st,1)$ and so $\alpha(s,t)$ depends on s,t only through the product st. Otherwise $\alpha(s/t,t)$ is strictly decreasing in t.

(iii) If $X[0,t] = \sum_i X_i[0,t]\Sigma iXi$ where $(X_i[0,t])_i$ are independent, then

$$\alpha(s,t) = \sum_i \alpha_i(s,t) \qquad (A.3)$$

(iv) The Effective Bandwidth $\alpha(s,t)$ is increasing in s for any fixed value of t, and lies between the mean and peak arrival rate measured over the interval of length t, that is:

$$\frac{EX[0,t]}{t} \le \alpha(s,t) \le \frac{\overline{X}[0,t]}{t} \qquad (A.4)$$

where $\overline{X}\,[0,t]$ is $\sup\{x: P\{x[0,t] > x\} > 0\}$ the essential supremium.

The form of the Effective Bandwidth *α(s,t)* near $s = 0$ is determined by the mean, variance and higher moments of *X[0,t]*, while its form *α(s,t)* near $s = \infty$ is primarily influenced by the distribution of *X[0,t]* near the maximum. If the Effective Bandwidth *α(s,t)* is finite for some $s > 0$, then for a given *t:*

$$\alpha(s,t) \le \frac{1}{t}\mathrm{EX}[0,t] + \frac{s}{2t}\mathrm{VarX}[0,t] + o(s) \qquad as \quad s \to 0 \quad (A.5)$$

If the Effective Bandwidth *α(s,t)* is bounded above as $s \to \infty$ then for a given

$$\alpha(s,t) = \frac{\overline{X}[0,t]}{t} + \frac{1}{st}\log P\{X[0,t] = \overline{X}[0,t]\} + o\left(\frac{1}{s}\right) \qquad (A.6)$$

as $s \to \infty$.

A.2 Examples of Effective Bandwidth for Different Source Models

The scales of time and space are determined by the source and Quality of Service required, and by the capacity of buffer lengths. Kelly *[36]* derives the *α (s,t)* Effective Bandwidth descriptors for different source traffic models - Bernoulli bufferless models, periodic models, fluid models and fractal Brownian motion input models. They lead to admissible regions that give the time and space scales, *s* and *t*, for these sources.

A.2.1 Periodic Sources

The model is used to describe packets streams from constant rate information sources, for a source which produces *b* units of workload at times *{Ud +nd, n= 0,1,...}* where *U* is uniformly distributed on the interval *[0,1]:*

$$\alpha(s,t) = \frac{b}{t}\left[\frac{t}{d}\right] + \frac{1}{st}\log\left[1 + \left(\frac{t}{d} - \left[\frac{t}{d}\right]\right)\left(e^{bs} - 1\right)\right] \qquad (A.7)$$

and so:

$$\lim_{t \to \infty} \alpha(s,t) = \frac{\left(e^{bs} - 1\right)}{ds} \qquad (A.8)$$

For *b=d=1* shows that for *t* decreasing, the Effective Bandwidth increases, with a dramatic leap from one to zero. The model has been used for packet streams from constant rate information sources.

A.2.2 Fluid Sources

A two-state Markov chain describes the stationary fluid source. The transition rate from state *2* to state *1* is λ, from *1* to *2* is μ, and with the workload produced only when the Markov chain is in state *1* at constant rate *h*.

$$\alpha(s,t) = \frac{1}{st}\log\left\{\left(\frac{\lambda}{\lambda+\mu}, \frac{\mu}{\lambda+\mu}\right)\exp\left[\begin{pmatrix} -\mu+hs & \mu \\ \lambda & -\lambda \end{pmatrix} t\right]\begin{pmatrix} 1 \\ 1 \end{pmatrix}\right\} \qquad (A.9)$$

and:

$$\lim_{t \to \infty} \alpha(s,t) = \frac{1}{2s}\left(\left(hs - \mu - \lambda + \left((hs - \mu + \lambda)^2 + 4\lambda\mu\right)^{1/2}\right)\right) \qquad (A.10)$$

A stationary source described by a finite Markov chain with stationary distribution π and *q*-matrix *Q*, the workload is produced at rate h_i while the chain is in state *i*. From the backward equations for the Markov chain:

$$\alpha(s,t)=\frac{1}{st}\log\{\pi\exp[(Q+\mathbf{h}s)t]\mathbf{1}\} \qquad (A.11)$$

where $h = diag(h_i)$, and:

$$\lim_{t\to\infty}\alpha(s,t) = \frac{1}{st}\Phi(s) \qquad (A.12)$$

where Φ *(s)* is the largest real eigenvalue of the matrix $Q + hs$ as shown by Elwalid and Mitra *[34]*.

If $h_1 > h_i$, $i \neq 1$, then (A.6) becomes:

$$\alpha(s,t) = h_1 - \frac{1}{s}\left(\mu_1 - \frac{1}{t}\log\pi_1\right) + o\left(\frac{1}{s}\right) \qquad (A.13)$$

as $s\rightarrow\infty$, where μ_1 is the transition rate out of the state with peak rate. For $t = \infty$ for a fluid source with relevant limits for s and t is discussed in Chang, Thomas *[35]*.

A.2.3 Gaussian Sources

For a Gaussian sources

$$X[0,t] = \lambda t + Z(t) \qquad (A.14)$$

where *Z(t)* is normally distributed with zero mean, then the Effective Bandwidth is found from the first two terms of (A.5) then:

$$\alpha(s,t) = \lambda + \frac{s}{2t}VarZ\ (t) \qquad (A.15)$$

When the process is for heavy traffic models, *Var Z (t)* $= \sigma^2 t$ *[6][14].* When the process *Z* is fractional Brownian motion with Hurst parameter $H \in (0,1)$:

$$VarZ\ (t) = \sigma^2 t^{2H} \qquad (A.16)$$

$$\alpha(s,t) = \lambda + \frac{\sigma^2 s}{2} t^{2H-1} \qquad (A.17)$$

The behavior of $\alpha\ (s,t)$ as $t \rightarrow \infty$ depends on if *H<1/2, H=1/2 or H>1/2,* not on *s.* When *H<1/2*, lim $t \rightarrow \infty$ $\alpha\ (s,t)$ is finite, and does not depend on *s, H=1/2* then the limit depends on *s*, or $\alpha\ (s,t)$ grows as a fractional power of *t*. *H>1/2* means there is long-range order exhibited, and has been proposed as a model for Ethernet traffic.

A.2.4 General 'On-Off' Sources

The source alternates between long periods in an 'ON' state with an Effective Bandwidth $\alpha_1\ (s,t)$ and long periods in an 'OFF' state where it produces no workload. If *p* is the proportion of time spent in the 'ON' state, the values of *t* are small compared with the periods spent in the 'ON' or 'OFF' states.

$$\alpha(s,t) = \frac{1}{st} \log\left[1 + p\left(\exp\left(st\,\alpha_1(s,t)\right) - 1\right)\right] \qquad (A.18)$$

The 'ON' periods may at a finer time scales appear as a periodic source, with bursts having a structure on a finer timescale, so the definition of Effective Bandwidth depends on the range of *s* and *t*.

A.3 Multiplexing Models

The arrivals process is assumed to be the aggregation of the sources, with examples described in the previous sections.

$$X[0,t] = \sum_{j=1}^{J} \sum_{i=1}^{n_j} X_{ji}[0,t] \qquad (A.19)$$

The $(X_{ji}[0,t])_{ji}$ are independent processes with stationary increments whose distributions may depend on *j* but not on *i,* and the resource such as the switch has to cope with the aggregate arriving stream of work. The number of sources of type *j* is n_j, and for source of type *j* the Effective Bandwidth is $\alpha_j(s,t)$:

$$\alpha(s,t) = \sum_{j=1}^{J} n_j \alpha_j(s,t) \qquad (A.20)$$

The point of looking at multiplexing models is to figure out the constraints that exist, and to see if the sum of Effective Bandwidths for n_j number of sources is within the acceptance region for resource and Quality of Service requirements. The acceptance region is defined by a set of vectors *($n_1, n_2, \ldots\ n_j$),* for which a given performance in terms of queuing delay or buffer overflow is guaranteed.

The constraints are *(s*, t*, C*)* with the relationship:

$$\sum_{j=1}^{J} n_j \alpha_j(s^*,t^*) \leq C^* \qquad (A.21)$$

In the following sections, the choices of values for *(s*, t*, C*)* constraints and the acceptance region vectors are described for different types of multiplexing models. For bufferless models, the above equation is established based on the results of Hui *[20],* which establishes a conservative bound for a non-linear acceptance region for bufferless models. Then in Section A.3.2, a linear limiting form for the acceptance region is found for a buffered model with Levy input (*M/G/1* models). This includes fluid sources that are studied and a linear limiting form of the acceptance region found.

A.3.1 Bufferless Models

A simple model made up of the aggregation of sources X_{ji} of type *j:*

$$X = \sum_{j=1}^{J} \sum_{i=1}^{n_j} X_{ji} \qquad (A.22)$$

X_{ji} represents independent random variables with scaled logarithmic moment generating functions:

$$\alpha_j(s) = \frac{1}{s} \log E\left[e^{s X_{ji}}\right] \qquad (A.23)$$

If X_{ji} the instantaneous arrival rate of work from a source type j at a bufferless source of capacity C, and let X_{ji} *[0,t]* = $X_{ji} t$ so that $\alpha_j(s/t,t) = \alpha_j(s)$ for all values of t - from property (ii) of the definition of Effective Bandwidth, defined in the first section of this Appendix.

The constraint to be satisfied found from Chernoff's bound,

$$\log P\{X \geq C\} \leq \log E\left[e^{s(X-C)}\right] = s(\alpha(s) - C) \qquad (A.24)$$

so the constraint $\log P\{X \geq C\} \leq -\gamma$ will be satisfied by vector $n = (n_1, n_2, \ldots n_J)$ if it lies within the set A, where:

$$A = \left\{ n : \inf_s \left[s\left(\sum_{j=1}^{J} n_j \alpha_j(s) - C \right) \right] \leq -\gamma \right\} \qquad (A.25)$$

The region of set A is used to find the global bound on the acceptance region. Since A has a convex complement in R_+^J and this complement is defined at the intersection of R_+^J with a family of half spaces. By replacing *n* by *n** we get *s** for the infinium of A.25 above, and so the half-space touching at point *n** on the boundary of region *A* is:

$$\sum_{j=1}^{J} n_j \alpha_j (s^*) \leq C - \frac{\gamma}{s^*} \qquad (A.26)$$

This condition is a conservative global bound, of the form as defined in A.21 by the constraints are *(s*, t*, C*)* with the relationship as follows:

$$\sum_{j=1}^{J} n_j \alpha_j (s^*, t^*) \leq C^* \qquad (A.21)$$

This defines the bound on the acceptance region, so if *n* satisfies the condition A.21 then the performance guarantees of $log\ P\ \{\ X\ \geq C\ \} \leq\ -\gamma$ are met, representing the queuing delay or buffer overflow.

Let $A(\gamma,C)$ be the subset of R_+^J, such that n $\in A(\gamma,C)$ implies $log\ P\ \{\ X \geq\ C\ \}\ \leq\ -\gamma$, from Chernoff's theorem *[17]* :

$$\lim_{N \to \infty} \frac{1}{N} \log P \left\{ \sum_{j=1}^{J} \sum_{i=1}^{n_j N} X_{ji} \geq CN \right\} = \inf_s \left[s \left(\sum_{j=1}^{J} n_j \alpha_j (s) - C \right) \right] \qquad (A.27)$$

The infinium of the above equation is strictly increasing in each component of *n*, and so:

$$\lim_{N \to \infty} \frac{A(\gamma N, CN)}{N} = A \qquad (A.28)$$

This convergence statement means the approximation leading to region *A* becomes more accurate as the number of sources increases, and the tail probability decreases. This indicates the probability of resource overload, which we can convert to the proportion of work lost from an arriving stream by the next two steps: relating the expected size of overloads to tail probabilities, and then dividing by stream rates of the arriving streams. From Chernoff's bound, the expected rate of load loss is:

$$E\left(X-C\right)^{+} = \int_{0}^{\infty} P\{X \geq C + x\}dx$$

$$= \int_{0}^{\infty} \exp\left[s(\alpha(s) - (C + x))\right]dx$$

$$= \frac{1}{s}\exp\left[s(\alpha(s) - C)\right]$$

so the following is deduced:

$$E\left(X-C\right)^{+} \leq \frac{1}{s*}\exp\left[s*(\alpha(s*) - C)\right] \qquad (A.29)$$

where $s*$ in the infinium of A.27.

If the global bound condition A.26 is satisfied:

$$\sum_{j=1}^{J} n_j \alpha_j(s*) \leq C - \frac{\gamma}{s} \qquad (A.26)$$

then $P\{\ X > C\ \} \leq exp\ (-\ \gamma)$ and also $E\left(X-C\right)^{+} \leq e^{-\gamma/s^{*}}$ are assured. The proportion of load lost is $E(X-C)^{+}/E(X)$. If $n \in A\ st\ (\gamma, C)$ the subset of R_{+}^{J}, this means the proportion of work lost is not greater than $e^{-\gamma}$.

A.3.1.1 Improved Approximations

The inequalities A.24 below and A.29 (which is the constraint to be satisfied) are found from Chernoff's bound:

$$\log P\{X \geq C\} \leq \log E\left[e^{s(X-C)}\right] = s(\alpha(s) - C) \qquad \textit{(A.24)}$$

This provides bounds on the probability of resource overload or proportion of work lost. Estimates to get closely related 'tilted approximations' may be found *[31]* and are discussed by Hui *[20]:*

$$P\{X \geq C\} \sim \frac{1}{s*(2\pi\sigma^2(s*))^{1/2}} e^{s*(\alpha(s*)-C)} \qquad \textit{(A.30)}$$

and:

$$E[X - C]^+ \sim \frac{1}{s*(2\pi\sigma^2(s*))^{1/2}} e^{s*(\alpha(s*)-C)} \qquad \textit{(A.31)}$$

where $\sigma^2(s) = \frac{\partial^2}{\partial s^2}(s\alpha(s))$.

A.3.1.2 Approximate Linearity

To look at how well approximated is the region A.25 of the set A region defined by A.26:

$$A = \left\{ n : \inf_s \left[s\left(\sum_{j=1}^{J} n_j \alpha_j(s) - C \right) \right] \leq -\gamma \right\} \qquad \textit{(A.25)}$$

The linearly constrained region used to approximate set A is:

$$\sum_{j=1}^{J} n_j \alpha_j (s^*) \leq C - \frac{\gamma}{s^*} \qquad (A.26)$$

A Gaussian source with a normally distributed load is used to find this approximation between the above two equations, as it is the easiest to calculate.

Let the normal distributed load be:

$$\alpha_j (s^*) \leq \lambda_j + \frac{s \sigma_j^2}{2}$$

with load mean λ_i and variance σ_j.

The region of A.24 of set A becomes:

$$\sum_j n_j \lambda_j + \left(2\gamma \sum_j n_j \sigma_j^2 \right)^{1/2} \leq C \qquad (A.32)$$

The tangent plane at point n^* on the boundary of the region in the above equation A.33 is of the form A.26 with:

$$s^* = \frac{C - \sum_j n_j^* \lambda_j}{\sum_j n_j^* \sigma_j^2} \qquad (A.33)$$

and so the Effective Bandwidth of s^* (for the global bound of set *A)* is:

$$\alpha_j (s^*) = \lambda_j + \frac{\gamma \sigma_j^2}{C(1 - \delta^*)} \qquad (A.34)$$

where $\delta^* = \sum_j n_j \lambda_j / C$, the traffic intensity. The coefficients are relatively insensitive to the traffic mix n^*, provided $1/(1-\delta^*)$ does not vary too greatly with

n^*, to put this another way, provided that the traffic intensity is not too close to 1 on the boundary of the acceptance region.

Let $C^* = C - \gamma / s^*$, the effective capacity appearing to the right hand side of equation A.26, then:

$$C^* = C - \gamma \frac{\sum_j n_j^* \sigma_j^2}{C(1 - \delta^*)} \qquad (A.35)$$

$$= C - \gamma \frac{\text{variance of load}}{\text{mean free capacity}} \qquad (A.36)$$

A.3.2 M/G/1 Models

If the model has Q distributed as a stationary workload in a queue with server of capacity C and an infinite buffer, we will look at the proportion of time the buffer occupancy exceeds a level b. The arrival stream $X[0,t]$ is made up of processes $X_{ji}[0,t]$ with independent increments, so $\alpha_j(s) = \alpha_j(s,t)$ as before. To define the queue size at time τ:

$$Q(\tau) = (X[0,\tau] - C\tau) - \inf_{0<t<\tau} \{X[0,t] - Ct\} \qquad (A.37)$$

letting $\tau \to \infty$.

The Pollaczek-Khinchin formula *[14][15]* is:

$$E[e^{sQ}] = \frac{C - \alpha(0)}{C - \alpha(s)} \qquad (A.38)$$

Cramer's estimate *[31]* describes the tail behavior of the distribution for Q (the workload in the queue). There is a finite constant κ such that the interior of the interval on which α(s) is finite, so that $\alpha'(\kappa)$ is finite. Then Cramer's estimate is:

$$P\{Q \geq b\} \sim \frac{C - \alpha(0)}{\kappa\alpha'(s)} e^{-\kappa b} \qquad (A.39)$$

as $b \to \infty$.

Let $A(\gamma,b)$ be the subset of R_+^J, such that $n \in A(\gamma,b)$ implies $\log P\{Q \geq b\} \leq -\gamma$.

Then as a result of Cramer's estimate:

$$\lim_{N \to \infty} A(\gamma N, bN) = A \qquad (A.40)$$

where:

$$A = \{n : \sum n_j \alpha_j\left(\frac{\gamma}{b}\right) \leq C\} \qquad (A.41)$$

A is a region defined by a constraint of the form A.21, which is:

$$\sum_{j=1}^{J} n_j \alpha_j(s^*, t^*) \leq C^* \qquad (A.21)$$

with $A \subset A(\gamma, b)$ so the linearly constrained region A is a conservative global bound as well as an asymptotic limit.

A.3.2.1 Finite Buffers

If there is a finite buffer size *b*, the proportion of time the buffer occupancy exceeds this level indicates the excess workload lost. So we can use the *M/G/1* queue equations by removing the time intervals when the workload is above *b*, from Cramer estimate, A.39 the proportion of workload lost with a buffer size *b*, *L(b)* is:

$$L(b) \sim \frac{C(C-\alpha(0))^2}{\kappa\alpha^{\prime}(s)\alpha(0)} e^{-\kappa b} \qquad (A.42)$$

as $b \to \infty$

If *A prop(γ, b)* is a subset of R_+^J, such that $n \in A\ prop(\gamma, b)$ implies *logL(b)* ≤ - γ then:

$$\lim_{N\to\infty} A_{prop}(\gamma N, bN) = A \qquad (A.43)$$

A.3.2.2 Brownian Input

Let *Z(t)* be standard Brownian motion, then $X_{ji}[0,t] = \lambda_j t + \sigma_j Z(t)$ and Z^1 is a Brownian motion for the superpositions:

$$X[0,t] = \left(\sum_j n_j \lambda_j)t + \left(\sum_j n_j \sigma_j^2\right)^{1/2} Z^1(t)\right) \qquad (A.44)$$

From the basic formulae of Brownian Motion the constraint is:

$$P\{Q \geq b\} = \exp\left\{\frac{-2b(C - \sum_j n_j \lambda_j}{\sum_j n_j \sigma_j^2}\right\} \qquad (A.45)$$

thus the constraint *log P { Q≥b }* ≤ - γ becomes the following formula:

$$\sum_j n_j\left(\lambda_j + \sigma_j^2 \frac{\gamma}{2b}\right) \leq C \qquad (A.46)$$

which is the same as $\sum_j n_j \alpha_j(s^*, t^*) \leq C^*$ with $s^* = \gamma/2b$.

The Pollaczek-Khinchin formula A.38 it follows that $EQ = \alpha'(0)/(C - \alpha(0))$, and so the constraint $EQ \leq L$ provides a linear acceptance region is satisfied if:

$$\sum_{j=1}^{J} n_j \left[\alpha_j(0) + \frac{\alpha_j^{/}(0)}{L} \right] \leq C \qquad (A.47)$$

A.3.3 Buffer Asymptotic Models

Tail probabilities decay exponentially for models more general than the M/G/1 queue. Next we will see how the formulae A.39 and A.40, which were found in the previous section, will still hold. To get these to hold for asymptotic models, the increments for the queue are assumed to be ergodic rather than stationary, for *Q* the workload as before in the queue of server capacity *C* with an infinite buffer, with arrival stream *X[0,t]*. For asymptotic behavior, the limit of convergence and the rate of convergence of the Effective Bandwidth is of interest. If there is a limit for convergence:

$$\lim_{t \to \infty} \alpha(s,t) = \alpha(s) \qquad (A.48)$$

and there is a constant κ such that $\alpha(\kappa) = C$, and $\alpha'(\kappa)$ is finite:

$$\lim_{b \to \infty} \frac{1}{b} \log\{ Q \geq b\} = -\kappa \qquad (A.49)$$

The usefulness of the limit depends on whether the rate of convergence occurred within the timescale of interest, the convergence $\alpha(\kappa,t)$ to $\alpha(\kappa)$ should occur in this timescale so that the Cramer's estimate can be used. An example for an M/G/1 model should occur in the time frame of the time taken to fill the buffer t_1, and the time taken to empty it t_2, as b increases and $t_1 = b/(C-\alpha(0))$, $t_2 = b/(\kappa\alpha'(\kappa))$.

The ideas of buffer asymptotics can be extended to examples where the limit A.48 does not exist, but a large deviations principle can be applied. An example is Fractional Brownian input, with $\alpha(s,t)$ given by A.17:

$$\alpha(s,t) = \lambda + \frac{\sigma^2 s}{2} t^{2H-1} \qquad (A.17)$$

and so for the:

$$\alpha_{ji}(s,t) = \lambda_i + \frac{\sigma_j^2}{2} s t^{2H-1} \qquad (A.50)$$

so to find the Effective Bandwidth, we use equation A.20 to show that:

$$\lim_{b\to\infty} \frac{\log P\{Q \ge b\}}{b^{2(1-H)}} = -\frac{1}{2\sigma^2}\left(\frac{C-\lambda}{H}\right)^{2H}\left(1-H\right)^{-2(1-H)} \qquad (A.51)$$

It is possible to deduce from this equation that the condition that $P\{Q \ge b\} \le exp(-\gamma)$ becomes the next equation, as γ, $b \to\infty$:

$$2\gamma\sum_{j=1}^{J} n_j\sigma_j^2 \le b^{2(1-H)}\left(\frac{C-\sum_{j=1}^{J} n_j\lambda_j}{H}\right)^{2H}\left(1-H\right)^{-2(1-H)} \qquad (A.52)$$

with γ / b held constant.

For *H* = *1/2*, the above equation becomes A.46:

$$\sum_j n_j \left(\lambda_j + \sigma_j \frac{\gamma}{2b} \right) \leq C \qquad (A.46)$$

For *H* →*1* it becomes A.32:

$$\sum_j n_j \lambda_j + \left(2\gamma \sum_j n_j \sigma_j^2 \right)^{1/2} \leq C \qquad (A.32)$$

The long range order is mostly effected by the scaling relationship between *γ*, *b* and *C*, as opposed to the form of the acceptance region *A*.

References

[1] De Prycker M, "Asynchronous Transfer Mode Solution for Broadband ISDN", *Prentice Hall,* 1995.

[2] Goralski W, "Introduction to ATM Networking", *Mc Graw Hill,* 1995.

[3] Mc Dysan D, Spohn S, "ATM Theory and Applications", *Prentice Hall,* 1995.

[4] Stallings W, "High-Speed Networks TCP/IP and ATM Design Principles" Prentice Hall, *February 1997.*

[5] Michiel H, Laevens K, "Teletraffic Theory in the Broad-Band Era ", *Proceedings of the IEEE, December 1997, pp.2007-2033.*

[6] Adas A, "Traffic Models in Broadband Networks", *IEEE Communications Magazine, July 1997, pp.82-89.*

[7] ATM Forum, "Traffic Management Specification, Version 4.0", *The ATM Forum, af-tm-0056.000, 1996.*

[8] ATM Forum, "ATM User-Network Interface Specification, version 4.0", *The ATM Forum, af-sig-0061.00 July 1996.*

[9] Fontaine M and Smith DG, "Bandwidth Allocation and Connection Admission Control in ATM Networks", *Electronics & Communication Engineering Journal, August 1996, pp.156-164.*

[10] Grossglauser M, Keshev S, Tse D, "RCBR: A Simple and Efficient Service for Multiple Time-Scale Traffic ", *IEEE/ACM Transactions on Networking,, Vol. 5, No.6, December 1997, pp.741-755.*

[11] Dziong Z, Juda M, Mason L, "A Framework for Bandwidth Management in ATM Networks - Aggregate Equivalent Bandwidth Estimation Approach ", *IEEE/ACM Transactions on Networking, Vol. 5, No.1, February 1997, pp.134-147.*

[12] Tse D, Grossglauser M, "Measurement-Based Call Admission Control: Analysis and Simulation", *Proceedings of the IEEE INFOCOM 1997, pp. 981-989.*

[13] Jamin S, Shenker S, Danzig P, "Comparison of Measurement-Based Admission Control Algorithms for Controlled-Load Service", *Proceedings of the IEEE INFOCOM 1997, pp. 973-980.*

[14] Nelson R, "Probability, Stochastic Processes, and Queuing Theory", *Springer-Verlag 1995.*

[15] Kleinrock L, "Queueing Systems Volume 1:Theory", *Wiley 1996.*

[16] Bonaventure O, Nelissen J, "Guaranteed Frame Rate:A Better Service for TCP/IP in ATM Networks*", IEEE Networks, January/February 2001, pp. 46-54.*

[17] Billingsley P, "Probability and Measure ", *Wiley 1995.*

[18] Allen A, "Probability, Statistics and Queueing Theory", *Computer Science and Applied Mathematics, Academic Press Inc, 1978.*

[19] Ross K, "Multiservice Loss Models for Broadband Telecommunication Networks", *Springer, 1996.*

[20] Hui JH, "Resource Allocation for Broadband Networks", *IEEE Journal on Selected Areas in Communications, Vol. 6, No. 9, December 1988, pp. 1598-1608.*

[21] Elwalid A, Mitre D, Wentworth R, "A New Approach for Allocating Buffers and Bandwidth to Heterogeneous, Regulated Traffic in an ATM Node", *IEEE. Journal on Selected Areas in Communications, Vol. 13, No. 6, August 1995, pp. 1115-1127.*

[22] Chen F, Mellor J, Mars P, "A Hybrid Approach for Generating Fractional Brownian Motion", *Proceedings of the IEEE SIGCOM 1996.*

[23] Tsybakov B, Georganas N, "ON Self-Similar Traffic in ATM Queues: Definitions, Overflow Probability Bound, and Cell Delay Distribution", *IEEE/ACM Transactions on Networking,, Vol. 5, No. 3, June 1997, pp. 397-409.*

[24] Murase T, Sukuki H, Sato S and Takeuchi T, "A Call Admission Control Scheme for ATM Networks using Simple Quality Estimation", *IEEE Journal on Selected Areas in Communications, Vol. 9, Dec 1991, pp. 1461-1470.*

[25] Saito H, Shiomoto K, "Dynamic Call Admission Control in ATM Networks", *IEEE Journal on Selected Areas in Communications, Vol. 9, No. 7, September 1991, pp. 982 -989.*

[26] Lee T, Lai K, Dann S, "Design of a Real-Time Call Admission Controller for ATM Networks", *IEEE Journal on Selected Areas in Communications, Vol. 4, No. 5, October 1996, pp. 758 -765.*

[27] Fabregat-Gesa R, Sole-Pareta J, Marzo-Lazaro J, Domingo-Pascual J, "Bandwidth Allocation Based on Real Time Calculations Using the Convolution Approach", *Proceeding of the IEEE GLOBECOM 1994.*

[28] Miyao Y, "A Call Admission Control Scheme for ATM Networks", *Proceeding of the IEEE GLOBECOM 1991.*

[29] Murase T, Sukuki H, Sato S and Takeuchi T, "A Call Admission Control for ATM Networks Based on Individual Multiplexed Traffic Characteristcs ", *Proceeding of the IEEE GLOBECOM 1991.*

[30] Gibbens R, Kelly F, Key P, "A Decision-Theoretic Approach to Call Admission Control in ATM Networks", *IEEE Journal on Selected Areas in Communications, Vol. 13, No. 6, August 1995, pp. 1101 - 1114.*

[31] Bucklew J, "Large Deviation Techniques in Decision, Simulation and Estimation", *Wiley Series in Probability and Mathematical Statistics, 1990.*

[32] Mitra D, Reiman MI, Wang J,"Robust Dynamic Admission Control for Unified Cell and Call QoS in Statistical Multiplexers", *IEEE Journal on Selected Areas in Communications, Vol. 16, No.5, June 1998, pp. 692 - 707.*

[33] Guerin R, Ahmadi H, Naghshineh M, "Equivalent Capacity and Its Application to Bandwidth Allocation in High-Speed Networks", *IEEE*

Journal on Selected Areas in Communications, Vol. 9, No. 7, September 1991, pp. 968 - 981.

[34] Elwalid A,Mitra D, "Effective Bandwidth of General Markovian Traffic Sources and Admission Control in High-Speed Networks", *IEEE Transactions on Networking, Vol. 1, No. 3, June 1993, pp. 329 -343.*

[35] Chang C, Thomas J, "Effective Bandwidth in High-Speed Networks", *IEEE Journal on Selected Areas in Communications, Vol. 13, No. 6, August 1995, pp. 1091 - 1100.*

[36] Kelly F, "Notes on Effective Bandwidths ", *Stochastic Networks Theory and Applications, Edited by FP Kelly, S Zachary, I Ziedins, Royal Statistical Society Lecture Notes Series 4, Oxford Science Publications, 1996.*

[37] Gibbens R J, "Traffic Characterisation and Effective Bandwidths for Broadband Network Traces", *Stochastic Networks Theory and Applications, Edited by FP Kelly, S Zachary, I Ziedins, Royal Statistical Society Lecture Notes Series 4, Oxford Science Publications, 1996.*

[38] Kesidis G, Walrand J, Chang C, "Effective Bandwidths for Multiclass Markov Fluids and Other ATM Sources", *IEEE/ACM Transactions on Networking, Vol 1, No.4, August 1993, pp. 424-428.*

[39] de Veciana G, Kesidis G, Walrand J, "Resource Management in Wide-Area ATM Networks Using Effective Bandwidths", *IEEE/ACM Transactions on Networking, Vol1, No.4, August 1993.*

[40] Leonard C, "On Large Deviations for Particle Systems Associated with Spacially Homogeneous Boltzmann Type Equations", *Probability Theory and Related Fields, 101, 1-44, 1995.*

[41] Einmahl U, Kuelbs J, "Dominating Points and Large Deviations for Random Vectors", *Probability Theory and Related Fields, 105, 329-345, 1996.*

[42] Duffield NG, Lewis JT, O Connell N, Russell R, Toomey F, "Entropy of ATM Traffic Streams: A Tool for Estimating QoS Parameters", *IEEE Journal on Selected Areas in Communications, Vol. 13, No.6, August 1995, pp. 981 -990.*

[43] Ganesh A, "Estimating Effective Bandwidths from Traffic Data", *IEEE SIGCOM 1996.*

[44] Weiss A, "An Introduction to Large Deviations for Communication Networks", *IEEE Journal on Selected Areas in Communications, Vol. 13, No.5, August 1995, pp. 938 - 952.*

[45] Tse D, Gallagher R, Tsitsiklis, "Statistical Multiplexing of Multiple Time-Scale Markov Streams", *IEEE Journal on Selected Areas in Communications, Vol. 13, No.5, August 1995, pp. 1028 - 1038.*

[46] Haykin S, "Neural Networks A Comprehensive Foundation", *Prentice Hall 1994.*

[47] Zurada J M, "Introduction to Artificial Neural Systems", *West 1992.*

[48] Onyiagha G, Krasniqi X, Clarkson T, "Adaptive Access Control of ATM Traffic Using Neural Networks", *Proceedings of the IEEE SIGCOM 1996.*

[49] Hiramatsu A, "Integration of ATM Call Admission Control and Link Capacity Control by Distributed Neural Networks" , *IEEE Journal on Selected Areas in Communications, Vol. 9, No.7, September 1991.*

[50] Morris R, Samadi B, "Neural Networks in Communications: Admission Control and Switch Control", *Proceedings of IEEE ICC 1991.*

[51] Davoli F, Maryni P, "A Two-level Stochastic Approximation for Admission Control and Bandwidth Allocation", *IEEE Journal on Selected Areas in Communications, Vol. 18, No.2, February 2000, pp. 222 - 233.*

[52] Chang C, Lin S, Cheng R, Shiue Y, "PSD-based Neural-net Admission Control", *Proceedings of the IEEE INFOCOM 1997.*

[53] Uhehara K, Hirota K, "Fuzzy Connection Admission Control for ATM Networks Based on Possibility Distribution of Cell Loss Ratio", *IEEE Journal on Selected Areas in Communications, Vol. 15, No.2, February 1997, pp. 179 - 190.*

[54] Youssef S, Habib I, Saadawi T, "A Neurocomputing Controller for Bandwidth Allocation in ATM Networks", *IEEE Journal on Selected Areas in Communications, Vol. 15, No.2, February 1997, pp. 191 - 199.*

[55] Bensaou B, Lam S, Chu H, Tsang D, "Estimation of the Cell Loss Ratio in ATM Networks with a Fuzzy System and Application to Measurement-Based Call Admission Control ", *IEEE/ACM Transactions on Networking,, Vol. 5, No.4, August 1997, pp.572-584.*

[56] Gersht A, Lee KJ, "A Congestion Control Framework in ATM Networks", *IEEE Journal on Selected Areas in Communications, Vol. 9, No. 7, September 1991, pp. 1119 - 1130.*

[57] Cheng RG, Chang CJ, Lin J, "A QoS-Provisioning Neural Fuzzy Connection Admission Controller for Multimedia High-Speed Networks", *IEEE/ACM Transactions on Networking, Volume 7, Number 1, February 1999, pp.111-121.*

[58] Ren Q, Ramamurthy MG, "A Real-Time Dynamic Admission Controller Based on Traffic Modeling, Measurement, and Fuzzy Logic Control", *IEEE Journal on Selected Areas in Communications, Vol. 18, No.2, February 2000, pp. 184 - 196.*

[59] Schormans J, Pitts J, "New Models for Admission Control of Priority Traffic in ATM Networks", *Proceedings of the IEEE SIGCOM 1996.*

[60] Shiomoto K, Yamanaka N, "A Simple Multi-QoS ATM Buffer Management Scheme Based on Adaptive Admission Control", *IEEE SIGCOM 1996.*

[61] Lee H, Park CG, Kim Y, "Providing Multiclass QoS in Shared ATM Output Multiplexer", *Proceedings of the IEEE SIGCOM 1996.*

[62] Wang Q, Frost V, "Efficient Estimation of Cell Blocking Probability for ATM Systems", *IEEE/ACM Transactions on Networking, Vol. 1, No. 2, April 1993.*

[63] Marbach P, Mihatsch O, Tsitsiklis J, "Call Admission Control and Routing in Integrated Services Networks Using Neuro-Dynamic Programming", *IEEE Journal on Selected Areas in Communications, Vol. 18, No.2, February 2000, pp. 197 - 208.*

[64] Tong H, Brown T X, "Adaptive Call Admission Control Under Quality of Service Constraints: A Reinforcement Learning Solution", *IEEE Journal on Selected Areas in Communications, Vol. 18, No.2, February 2000, pp. 209 - 221.*

[65] Jamoussi B, Adoul-Magd O, "Performance Evaluation of Connection Admission Control Techniques in ATM Networks", *Proceedings of the IEEE SIGCOM 1996.*

[66] Lee H, Mark J, "Capacity Allocation in Statistical Multiplexing of ATM Sources", *IEEE/ACM Transactions on Networking, Vol. 3, No. 2, April 1995.*

[67] Botvich D, Duffield N, "Large deviations, economies of scale, and the shape of the loss curve in large multiplexers", *Queuing Systems, Vol 20, p. 293-320, 1995.*

[68] McGurk B, Walsh C, "Investigations of the performance of Measurement-Based Connection Admission Control algorithm", *Technical Report, Dublin Institute of Advanced Studies Applied Probability group, 1999.*

[69] Crosby S, Leslie L, McGurk B, Lewis J, Russell R, Toomey F, "Statistical properties of a near-optimal Measurement-Based CAC algorithm", *Proceedings of the IEEE ATM '97 Workshop, p. 103-112, Lisbon, Portugal, May 1997.*

[70] Breslau L, Jamin S, Shenker S, "Comments on the Performance of Measurement-Based Admission Control Algorithms" *Technical Report, International Computer Science Institute, 1998.*

[71] Kreyszig E, "Advanced Engineering Mathematics" *Wiley, 8th Ed., 1999.*

[72] Qiu J, Knightly E, "Measurement-Based Admission Control with Aggregate Traffic Envelopes", *IEEE/ACM Transactions on Networking, Vol. 9, No. 2 April 2001, pp 199-210.*

[73] Floyd S, "Comments on Measurement-Based Admission Control Algorithms for a Controlled-Load Service" *Technical Report, Lawrence Berkeley Laboratory, July 1996.*

[74] Breslau L, Jamin S, Shenker S, "A Measurement-Based Admission Control Algorithm for Integrated Services Packet Networks" *IEEE/ACM Transactions on Networking, Vol. 5, No. 2, pp. 56-70, 1997.*

[75] Hogg R, Tanis E, "Probability and Statistical Inference", *Prentice Hall, 5th Ed., 1997.*

[76] Sahinoglu Z, Tekinay S, "On Multimedia Networks: Self-Similar Traffic and Network Performance", *IEEE Communications Magazine, January 1999, pp. 48-52.*

[77] Tsybakov B, Georganas N, "On Self-Similar Traffic in ATM Queues: Definitions, Overflow Probability Bound, and Cell Delay Distribution" *IEEE/ACM Transactions on Networking, Vol. 5, No. 3 June 1997, pp. 397-409.*

[78] Willinger W, Taqqu M, Sherman R, Wilson D, "Self-Similarity Through High-Variability: Statistical Analysis of Ethernet LAN traffic at the Source Level" *IEEE/ACM Transactions on Networking, Vol. 5, No. 1 February 1997, pp. 71-87.*