Growing and pruning (GAP) RBF networks for call admission control in ATM traffic management

Mohit Aiyar

2005

Mohit, A. (2005). Growing and pruning (GAP) RBF networks for call admission control in ATM traffic management. Master's thesis, Nanyang Technological University, Singapore.

<https://hdl.handle.net/10356/4905>

<https://doi.org/10.32657/10356/4905>

Nanyang Technological University

Downloaded on 26 Apr 2023 12:42:46 SGT

**Growing and Pruning (GAP) RBF Networks
For Call Admission Control
In ATM Traffic Management**

Mohit Aiyar

School of Electrical & Electronic Engineering

A thesis submitted to the Nanyang Technological University
in fulfillment of the requirement for the degree of
Master of Engineering

TK
5105.35
M697
2005

Acknowledgments

Prof. N. Sundararajan has been a pillar of support during the course of this project. He has been instrumental in ensuring the successful conclusion of the research initiative. On a personal note, Prof. Sundararajan has been a tremendous source of inspiration and guidance. For his unwavering support and help, I am very grateful and express my heartfelt appreciation.

I would like to thank A/P Saratchandran for unstinting encouragement right from the inception of the project. His wealth of knowledge and experience in the field of research and as an illustrious member of the academic community provided invaluable insights which contributed to the project maintaining a steady course of progress.

Last, but definitely not in the least, I would like to convey a sincere vote of thanks and appreciation to my supervisors at Citibank N.A – Mr. Ajit Kanagasundram, Ms. Susan Kwek and Mr. James Berti for their kind understanding and support. Balancing the demands of a challenging full-time job and a part-time research project is by no means an easy undertaking – Ajit, Susan and James, by means of their staunch support and kind consideration, tremendously contributed towards the accomplishment of this task.

Contents

Acknowledgmentsii

Contents iii

Summaryvi

List of Figures..... viii

List of Tables x

1 **Introduction**..... 1

1.1 Motivation..... 1

1.2 Objectives 4

1.3 Major Contributions of the Thesis 5

1.4 Organization of the Thesis 7

2 **Asynchronous Transfer Mode (ATM) Networks**.....9

2.1 An Overview of B-ISDN & ATM 10

2.2 A Review of the ATM Architecture 12

2.3 A Study of Data Flows in ATM Networks 18

2.4 An Introduction to Traffic Management in ATM Networks 20

2.4.1	Quality of Service Attributes	20
2.4.2	ATM Traffic Service Categories	21
2.4.3	Traffic Contract	24
2.4.4	Traffic Management Schemes	26
2.4.5	Congestion Control	30
3	ATM Traffic Management – Call Admission Control	35
3.1	A Review of ATM Call Admission Control	36
3.2	Limitations of Conventional Call Admission Control	40
4	Neural Call Admission Controllers	44
4.1	Radial Basis Function Neural Networks	44
4.2	Minimal Resource Allocation Network (MRAN)	48
4.3	MRAN Applications	53
4.4	Growing and Pruning (GAP) Neural Network	55
4.4.1	Growing Criterion	55
4.4.2	Pruning Criterion	57
4.4.3	Nearest Neuron for Parameter Adjustment and Pruning	58
4.5	Neural Networks for Traffic Management of ATM Networks	61
5	OPNET Simulation Setup for ATM Networks and Call Admission Control Schemes	67
5.1	Relevant features of OPNET Radio / Modeler v8.0	68

5.2	A Review of the Hierarchical Model Architecture in OPNET	72
5.3	OPNET Simulation Setup	79
5.3.1	Application Configuration Panel	81
5.3.2	Profile Configuration Panel	83
5.3.3	Scenarios, Profiles and Applications	84
5.3.4	ATM Call Admission Control in OPNET	87
6	Quantitative Performance Evaluation of Neural CAC Schemes.....	91
6.1	Traffic Scenario 1 – Application: FTP.....	92
6.2	Traffic Scenario 2 – Applications: FTP and Email.....	100
6.3	Traffic Scenario 3 – Applications: FTP, Email and Voice	109
6.4	Summary of Performance Results	113
6.5	Cumulative Distribution Function Analysis of Call Blocking Ratio and ATM Cell Delay	117
7	Conclusions and Recommendations.....	124
7.1	Conclusions.....	124
7.2	Recommendations for Further Research.....	127
	Author's Publications	130
	References	131
	Appendix A - Traffic Scenario Parameters Setup in OPNET	141

Summary

This thesis presents a study on the use of recently developed neural networks MRAN (Minimal Resource Allocation Network) and GAP (Growing and Pruning neural network) for the performance enhancement of Call Admission Control in Asynchronous Transfer Mode (ATM) networks. GAP and MRAN generate a minimal radial basis function neural network by adding and pruning hidden neurons based on input data and are ideal for online adaptive control of fast time-varying non-linear systems. The use of GAP and MRAN in the study of call admission control schemes is new. The fast learning and accurate predictions obtained with the neural networks are shown to make better call admission control decisions under heavy traffic situations compared to conventional schemes.

The study is carried out using the well-known simulation package OPNET – Optimized Network Engineering Tools (OPNET) – Modeler v8.0. The simulation environment in OPNET consisted of the ATM network setup, traffic modelling, investigation of the admission control schemes and integration of the new neural network algorithms in OPNET.

The performance of the CAC schemes were evaluated under three main scenarios involving heavy traffic situations of varying multiplexed traffic consisting of bursty, variable bit rate and multiplexed traffic (combining both speech and video signals) in the form of applications such as FTP, Email and Voice. For each scenario, the performance metrics of Call blocking ratio, Cell delay and its variations and also the

Throughput were compared for three different traffic contracts of varying Quality of Service requirements – labelled High, Medium and Low respectively.

Simulation results indicate that the GAP and MRAN controllers perform better than the conventional call admission controller in making more effective call blocking decisions while maintaining a high quality of service. Results from a typical simulation across all traffic scenarios and traffic contract requirements indicate that the neural CAC schemes consistently provide a better Call Blocking Ratio and Throughput. Cell Delays experienced by the neural controllers were found to be approximately the same as or worse than the conventional schemes. The cell delay variation experienced by the neural controllers was found to be greater than the conventional schemes for all the scenarios studied. The additional delay experienced while using the neural controllers might be attributed to the overhead introduced by the complex computations needed by the neural Call Admission Control schemes. The Cumulative Distribution Function determined for Call Blocking Ratio and ATM Cell Delay based on an execution set of 20 simulation runs across two scenarios corroborated the above observations.

The study results indicate that the GAP and MRAN driven call admission control methods result in enhanced call admission control in ATM networks subjected to different and varying traffic situations and can be evaluated further in a real implementation in a real network.

List of Figures

Figure 2.1 ATM Protocol Structure	13
Figure 2.2 - ATM Cell Structure – UNI and NNI.....	16
Figure 2.3 – ATM Connection identifiers.....	17
Figure 2.4 – Multilevel traffic controls in ATM networks	27
Figure 4.1 – Radial Basis Function Network Model	46
Figure 4.2 –Neural Networks in a traffic control system.....	63
Figure 5.1 – OPNET Modeler / Radio Version 8.0 Common Desktop Environment	69
Figure 5.2 - Simulation Project Cycle.....	72
Figure 5.3– ATM Model Suite – ATM_Advanced Component Suite.....	77
Figure 5.5– Sample ATM Process Model.....	79
Figure 5.6– ATM Network Model.....	80
Figure 5.7 - Application Model Hierarchy	81
Figure 6.1.1 – Call Blocking Ratio – Medium QoS	92
Figure 6.1.2 – Call Blocking Ratio – High QoS	93
Figure 6.1.3 – ATM Cell Delay – Medium QoS	94
Figure 6.1.4 – ATM Cell Delay – High QoS	95

Figure 6.1.5 – ATM Cell Delay Variation – Medium QoS	96
Figure 6.1.6 – ATM Cell Delay Variation – High QoS	97
Figure 6.1.7 – Traffic Throughput on the backbone link – Medium QoS	98
Figure 6.1.8 – Traffic Throughput on the backbone link – High QoS.....	99
Figure 6.2.1 – ATM Call Blocking Ratio – Medium QoS.....	100
Figure 6.2.2 – ATM Call Blocking Ratio – High QoS	101
Figure 6.2.3 – ATM Cell Delay – Medium QoS	103
Figure 6.2.4 – ATM Cell Delay – High QoS.....	104
Figure 6.2.5 – ATM Cell Delay Variation – Medium QoS	105
Figure 6.2.6 – ATM Cell Delay Variation – High QoS	106
Figure 6.2.7 – ATM Traffic Throughput on Link 1 – Medium QoS	107
Figure 6.2.8 – ATM Traffic Throughput on Link 1 – High QoS.....	108
Figure 6.3.1 – ATM Call Blocking Ratio – Medium QoS.....	110
Figure 6.3.2 – ATM Cell Delay – Medium QoS	111
Figure 6.3.4 – ATM Cell Delay Variation – Medium QoS	112
Figure 6.4.1 – Cumulative Distribution of Call Blocking Ratio – Traffic Scenario 1 (High QoS).....	118
Figure 6.4.2 – Cumulative Distribution of Call Blocking Ratio – Traffic Scenario 2 – (Medium QoS).....	119
Figure 6.4.3 - Cumulative Distribution of ATM Cell Delay – Traffic Scenario 1 - (High QoS)	121
Figure 6.4.4 - Cumulative Distribution of ATM Cell Delay Traffic Scenario 2 - (Medium QoS).....	122

List of Tables

Table 2.1 - ATM Layer Service Categories.....	23
Table 2.2 - ATM Service Categories	24
Table 5.1 – Summary Parameter setup for Quality of Service levels.....	84
Table 5.2 – Summary Parameter setup for Traffic Scenario 1	85
Table 5.3 – Summary Parameter Setup for Traffic Scenario 2.....	86
Table 5.4 – Summary Parameter Setup for Traffic Scenario 3	87
Table 6.4.1 – Overall Performance Comparison of the Call Admission Controllers – Traffic Scenario 1	114
Table 6.4.2 – Overall Performance Comparison of the Call Admission Controllers – Traffic Scenario 1	114
Table 6.4.3 – Overall Performance Comparison of the Call Admission Controllers – Traffic Scenario 2.....	114
Table 6.4.4 – Overall Performance Comparison of the Call Admission Controllers – Traffic Scenario 2.....	115
Table 6.4.5 – Overall Performance Comparison of the Call Admission Controllers – Traffic Scenario 3.....	115
Table Appendix A – Parameter setup for Traffic Scenario 1	142
Table Appendix B– Parameter setup for Traffic Scenario 2.....	144
Table Appendix C – Parameter Setup for Traffic Scenario 3	146

Chapter 1

Introduction

1.1 Motivation

The convergence of the telecommunication and computer industries has led to new paradigms being adopted in the world of networking. One such scheme that has been widely accepted and implemented to great benefit is the Asynchronous Transfer Mode (ATM). ATM was selected by the telecommunication (carrier) industry as the technology to deliver the Broadband Integrated Services Digital Network (B-ISDN) carrier service [1]-[4]. The ATM design makes it imminently suited to handling different kinds of communication traffic (voice, audio, video and data) in an integrated manner – providing seamless interworking between the LAN and WAN network environments.

ATM technology supports a wide variety of services and applications. ATM, also known as cell relay, is in some ways similar to packet switching networks and frame relay methodologies. Like packet switching and frame relay, ATM allows multiple logical connections to be multiplexed over a single physical interface wherein user information is transmitted between communicating entities using fixed size 53 byte packets, referred to as ATM cells. [5] ATM is a streamlined protocol with minimal error and flow control capabilities. The ATM technique provides an attractive solution

to the contentious issue of integrating different types of services, with widely different bit rates, through common interface and switching fabrics. It is a compromise between packet switching and circuit switching techniques. It can support both connection and connection-less traffic, with provisions for guaranteed Quality of Service (QoS) via pre-allocation of resources. Statistical multiplexing is used to exploit the bursty nature of the varied traffic types, hence providing efficient usage of bandwidth

However, the bursty nature of the traffic and the task of multiplexing vastly different traffic types result in some serious traffic management issues necessitating the deployment of effective traffic management techniques [6]. Congestion Control, Traffic Policing, Network Resource Management and Call Admission Control are some of the traffic management approaches that are used in conjunction with each other to achieve a non-congested network offering high and reliable Quality of Service [7] – [8]. Call Admission Control (CAC) is the first line of defence for the network in protecting itself from excessive loads. CAC determines whether the QoS needs of a new call request can be supported in the light of the existing connections supported by the network. The objective of CAC is to restrict the access of new calls so that the cell and call level QoS is maintained within the path network. An additional objective is to maximise network revenue over time. Numerous conventional CAC schemes have been proposed and implemented [9] – [10]. Conventional CAC schemes make an admission decision primarily on the basis of the traffic characteristics specified at the start of the call. By definition, therefore, conventional schemes are most effective during steady state. However, the unpredictable nature of the B-ISDN traffic incident on the network poses a challenge which conventional schemes find difficult to grapple with. The rationale and principles underlying the nature and choice of thresholds or equivalent capacity under dynamic conditions is incoherent. Networks are forced to make decisions based on incomplete information so that the decision process is full of uncertainty. Thus, because of the unpredictable statistical fluctuations of the system, these control schemes will always be subject to decision errors, which degrade performance.

Traditional analytical methods require complicated calculations and exhaustive simulations to derive relationships between input traffic and service quality and cannot adapt to unexpected traffic.

Neural networks, with their property to efficiently adapt and predict, stand out as a very viable and suitable alternative. Recently, use of artificial neural networks (ANNs) in traffic management of ATM networks is gaining momentum [11] – [12]. The self-learning capability of the neural network is used to characterize the relationship between input traffic and system performance. Neural network adaptability is used to learn the relation between offered traffic and service quality from observed data. The neural network learns how to adapt to unexpected traffic as a non-linear function from noisy data. When incorporated into a network, the neural network creates a decision function by learning the behaviour of the operating multiplexer. When the characteristics of the multiplexed calls change, the neural network correctly modifies its nonlinear function by interpolating or extrapolating from the past experienced results. Initially all the weights of the neural network are set to near optimal values. Then these initial weights are adjusted to the optimal values through training with real traffic patterns. This method is suitable to a situation where the traffic characteristics rapidly change because the neural networks can adaptively control by learning the relationship between the offered traffic characteristics and the service qualities from the results of actual network operation.

Optimized Network Engineering Tools (OPNET) [13] – [14] provides simulation tools to adequately portray realistic ATM networks and analyse the veracity of proposed hypotheses [15]. OPNET Modeler/Radio Version 8.0 is used by expert end-users from the academic field or from professional institutions and corporations. As a pre-requisite to using OPNET effectively, knowledge of how to model realistic traffic sources and devise true-to-life network configurations is needed [16]. Integration of customizations such as neural networks into the default models provided by OPNET requires an in-depth knowledge of the engineering behind the OPNET suite of products.

A minimal radial basis function (RBF) neural network called Minimal Resource Allocation Network has been developed in [6] – [7]. Recently, the authors have proposed an enhanced version of the MRAN neural network called GAP in 2002 [59]. This research initiative has been motivated by the prospect of using MRAN and GAP as neural networks to enhance the performance of conventional CAC schemes in ATM networks. OPNET v8.0 has been used as the simulation tool to map out and validate this scheme.

1.2 Objectives

The primary objective of this research is to study artificial neural network based traffic management of ATM networks, in particular RBFN neural networks and their application in Call Admission Control of ATM networks. More specifically, the objectives of the work in this thesis can be summarised as follows:

1. Performance Evaluation of neural Call Admission Control schemes when compared with conventional approaches to traffic management.
2. Development of an adaptive neural call admission control algorithm using
 - a. The Radial Basis Function neural network MRAN (Minimal Resource Allocation Network) and
 - b. An improved version of MRAN called GAP (Growing and Pruning) neural network.
3. Performance comparison of the GAP CAC scheme against the MRAN based and Conventional CAC schemes. OPNET is used to conduct a comprehensive and quantitative performance evaluation under well defined mild, medium and heavily congested multiplexed traffic scenarios while studying the following figures of merit – a. Call Blocking Ratio b. ATM Cell Delay c. ATM Cell Delay Variation d. Throughput.

1.3 Major Contributions of the Thesis

Major contributions made in this thesis are:

- Performance Evaluation of Neural Call Admission Control schemes when compared with conventional approaches to traffic management. The thesis looks into the conventional Call Admission Control schemes and explores the drawbacks faced by these approaches. Section 3.2 delves further into the limitations of conventional schemes. The thesis looks at how neural networks can be used to alleviate the drawbacks faced by the conventional schemes. Section 4.5 specifically proposes neural Call Admission Control schemes and the significant advantages they provide over conventional approaches.
- Development of an adaptive neural call admission control algorithm using the Radial Basis Function neural networks MRAN (Minimal Resource Allocation Network) and GAP (Growing and Pruning) neural network. The neural CAC schemes proposed using MRAN and GAP react to unpredictable and varying incident traffic conditions by making effective and dynamic Call Admission control decisions. Section 5.3.4 elaborates on the execution of the algorithm(s) into the ATM network hierarchy. Both algorithms were seamlessly integrated into the network with minimal integration issues.
- Performance comparison of the GAP CAC scheme against the MRAN based and Conventional CAC schemes. The neural Call Admission Controllers were found to make more effective decisions when compared to the conventional schemes. The four performance metrics recorded were a. ATM Call Blocking Ratio b. ATM Cell Delay c. ATM Cell Delay Variation and d. ATM Throughput. The most relevant figure of merit from the list is the ATM Call Blocking Ratio and is the most appropriate indicator of the effectiveness of the call admission control decisions. The performance of the CAC schemes were observed under three main scenarios of varying multiplexed traffic consisting of FTP only, FTP and Email and FTP, Email and Voice respectively. For each

scenario, the metrics were collected for three different traffic contracts of varying Quality of Service – labelled High, Medium and Low respectively.

For traffic scenario 1 (FTP only) under a high QoS traffic contract, GAP-CAC resulted in about 50% of the total number of calls being blocked, compared to approximately 56% for MRAN-CAC and 100% for Conventional-CAC. For traffic scenario 2 (FTP and Email only) under a high QoS traffic contract, GAP-CAC resulted in about 51% of the total number of calls being blocked, compared to approximately 60% for MRAN-CAC and 100% for Conventional-CAC. For traffic scenario 3 (FTP, Email and Voice) under an medium QoS traffic contract, GAP-CAC resulted in about 51% of the total number of calls being blocked, compared to approximately 49% for MRAN-CAC and 59% for Conventional-CAC. The metrics were also collected for medium and low QoS traffic contracts for each traffic scenario. The summary of results is documented in Table 7.1.

- The thesis also elaborates on the steps to be taken to setup a realistic ATM network configuration using the OPNET simulation environment. Traffic sources are setup to adequately reflect realistic scenarios in a B-ISDN environment. The means of capturing and investigating meaningful results is shown. In addition, the approach taken to manually tap and calculate data rate statistics and to install the neural network in an existing OPNET model is detailed during the course of the thesis.

This thesis clearly illustrates that the use of GAP and MRAN based neural Call Admission Control schemes result in significant performance improvement and it is worthwhile to investigate their actual implementations in an experimental setup.

1.4 Organization of the Thesis

The research work undertaken has been presented through eight chapters in this thesis.

Briefly, chapters 2-5 provide insights into the research done in areas that constitute the underlying themes of this project – namely ATM and related issues, neural networks and their application in traffic management of ATM networks.

Specifically, chapter 2 provides insights into the theory behind Asynchronous Transfer Mode (ATM). The ATM specific traits relating to Virtual Path / Virtual Channel, QoS categories, traffic contract and others is detailed in this chapter. The traffic management issues are tabled and expounded upon.

Chapter 3 focuses on the ATM traffic management technique of Call Admission Control. Honouring the traffic contract of an incident connection, evaluating relevant statistics such as Call Blocking Ratio and the inherent limitations of conventional CAC schemes is further examined.

Chapter 4 provides a brief introduction to neural networks. In particular, a short review of the RBF neural networks is provided, along with the development of MRAN and GAP, and its applications in various fields, like function approximation and non-linear time series prediction.

Chapter 5 details the application of neural networks in overcoming the shortcomings of conventional CAC schemes.

The definition of the ATM network configuration and the modelling of the traffic sources to depict realistic traffic scenarios are elaborated upon in Chapter 6. The investigation of the conventional CAC schemes within the context of OPNET is shown and the location and implementation of the neural network within OPNET node and process models is highlighted.

Chapter 7 provides the main performance results based on the OPNET simulation. The metrics used to track the performance of conventional schemes as compared to the MRAN/GAP based implementations are presented.

Chapter 8 summarizes the implications of the performance results and draws conclusions on the outcome of the research. It puts the current status of the research initiative into proper perspective and makes suggestions for future work.

Chapter 2

Asynchronous Transfer Mode (ATM) Networks

Asynchronous Transfer Mode (ATM) was adopted by the International Telecommunications Union – Telecommunications Standardization Sector (ITU-T) as the transport method for the broadband integrated services digital networks (B-ISDN). It uses a packet-like switching and multiplexing technique in which the packets are fixed-size cells of 53 bytes each. Cells are then multiplexed on the links to obtain the flexibility necessary to support all-purpose services.

The ATM technique provides an attractive solution to the problem of integrating different types of services, with widely different bit rates, through a common interface and switching fabrics. It is a compromise between packet switching and circuit switching techniques. Packet switching uses statistical multiplexing and is therefore considered to be a very flexible solution. However, it does not provide effective control over the network resources or the delivered quality to the end-users. These Quality of Service (QoS) parameters are often defined in terms of performance objectives such as packet loss rate tolerance, packet delay and delay variability tolerance. On the other hand, circuit switching, which uses synchronous time division multiplexing, is a more rigid structure. Although it provides means for controlling the network resources, it does so at the expense of efficient use of the same resources. ATM is a good compromise. It can support both connection and connection-less traffic, with provisions

for guaranteed QoS via pre-allocation of resources. Statistical multiplexing is used to exploit the bursty nature of Variable Bit Rate (VBR), Unspecified Bit Rate (UBR) and Available Bit Rate (ABR) traffic classes, hence providing efficient usage of bandwidth. In contrast to circuit switching, it is flexible enough to support different types of calls with different bit rates over the same multiplexing interface. These advantages come at a cost – Elaborate flow, congestion control and call admission control techniques need to be deployed concurrently as a coherent traffic management strategy.

The following chapter elaborates the various aspects of ATM technology. The hierarchy of the ATM topology and the data flows are reviewed. The definitions of a Virtual Path and a Virtual Channel connection are explained. The recent exponential expansion of the scope of traffic types and profiles has led to the contentious issue of traffic management on ATM networks – these issues are tabled and their implications highlighted.

2.1 An Overview of B-ISDN & ATM

Rapid advances in computer and communication technologies have resulted in the increasing convergence of the two fields. The same digital techniques are used for the storage and transmission of data, voice, image and video. Merging and evolving technologies, coupled with increasing demands for efficient and timely collection, processing and dissemination of information, have led to the development and are encouraging the enhancement of integrated systems that transmit and process all kinds of data. This evolution has found its form in the Broadband Integrated Digital Networks.

Broadband Integrated Services Digital Networks (B-ISDN) is conceived as an all-purpose digital network facilitating world-wide information exchange between any two subscribers without limitations that can be imposed by the communication medium or media. By definition, B-ISDN networks support a host of applications ranging from data to voice and streaming image. Considering the set of applications that use ISDN, these network systems will have to support interactive and distributive services, bursty

and continuous traffic, connection oriented and connectionless services, point to point and complex communications – all in the same network. The B-ISDN also contains intelligence for the purpose of providing service features, maintenance and network management functions. This intelligence is usually distributed between the network and its terminals. The following attributes broadly capture the type of services that B-ISDN provides:

- High Bandwidth
- Bandwidth on demand
- Varying Quality of Service (QoS) parameters
- Guaranteed Service Levels
- Continuous or variable bit rate services

The B-ISDN has assumed greater relevance with the onset and the exponential growth in the application on the Internet and local area networks.

ATM was designated by the ITU-T as the transfer mode of choice for B-ISDN. With the objective of accelerating the convergence of standards and industry cooperation, an international consortium called the ATM Forum was founded to ensure interoperability between public and private ATM implementations and to promote the use of ATM products and services. Although it is not a standards body, the ATM Forum works closely with the standards organizations such as the International Telecommunications Union (ITU) and the Internet Engineering Task Force (IETF) in developing the definitions for ATM standards. This international consortium has grown from fewer than 10 members in 1991 to over 700 members currently, consisting of public and private network equipment vendors and service providers, software companies, as well as government organizations, national research laboratories and universities.

There has been much research done on Quality of Service (QoS) provisioning with traffic management in wired networks. Traffic management over ATM networks has matured, and for QoS provisioning aspects, ATM surpasses other protocols such as

Internet Protocol (IP), Frame Relay, etc. However, with the great success and popularity of the Internet and the World Wide Web (WWW), IP is and will continue to be the major protocol supporting them, even though ATM is a much better protocol for QoS than IP since IP is connectionless in nature. In recent years, many researches have focused on QoS over IP. Attempts have been made to get similar performance results as in ATM, but the task is difficult due to IP being connectionless in nature. In other words, attempts are being made to put more connection-oriented features into IP to improve QoS provisioning. There are mainly three protocols for QoS over IP: Integrated Services, Differentiated Services, and Multi-Protocol Label Switching (MPLS). MPLS is the most promising of the three since it has more connection-oriented features compared to the other suggested approaches. Nevertheless, the QoS obtaining in IP or its offshoots fall short when compared to that prevailing in ATM.

Future applications are expected to require increasingly higher bandwidth and generated a heterogeneous mix of network traffic. Existing networks cannot provide the transport facilities to efficiently support a diversity of traffic with various service requirements. ATM is potentially capable of supporting all classes of traffic in one transmission and switching fabric technology. It promises to provide greater integration of capabilities and services, increased and more flexible access to the network, and more efficient and economical service.

2.2 A Review of the ATM Architecture

ATM carries all traffic on a stream of fixed-size packets (cells), each comprising 5 bytes of header information and a 48-byte information field (payload). The reason for choosing a fixed-size packet is to ensure that the switching and multiplexing function could be carried out quickly and easily. ATM is a connection-oriented technology in the sense that before two systems on the network can communicate, they should inform all intermediate switches about their service requirements and traffic parameters. This is similar to the telephone networks where a fixed path is set up from the calling party to the receiving party. In ATM networks, each connection is called a virtual circuit or

virtual channel (VC), because it also allows the capacity of each link to be shared by connections using that link on a demand basis rather than by fixed allocations. The connections allow the network to guarantee the quality of service (QoS) by limiting the number of VCs. Typically, a user declares key service requirements at the time of connection setup, declares the traffic parameters and may agree to control these parameters dynamically as demanded by the network.

Figure 2.1 ATM Protocol Structure

The ATM protocol reference model is based on standards developed by the ITU. Communication from higher layers is adapted to the lower ATM defined layers, which in turn pass the information onto the physical layer for transmission over a selected physical medium. The protocol reference model is divided into three layers: the ATM adaptation layer (AAL), the ATM layer, and the physical layer (Figure 2.1).

The ATM Adaptation Layer (AAL) interfaces the higher layer protocols to the ATM Layer. It relays ATM cells both from the upper layers to the ATM Layer and vice versa. When relaying information received from the higher layers to the ATM Layer, the AAL segments the data into ATM cells. When relaying information received from the ATM Layer to the higher layers, the AAL must take the cells and reassemble the payloads into a format the higher layers can understand. This is called Segmentation and Reassembly (SAR). The traffic types incident on an ATM network have been broadly categorized into the following four categories or classes.

Class A - Constant Bit Rate (CBR) service. This class relates to a particular type of connection-oriented service where the bit rate is constant. Examples of this service include 64 Kbit/sec voice, fixed-rate uncompressed video and leased lines for private data networks.

Class B - Variable Bit Rate (VBR) service. This class also relates to a connection-oriented service in which the bit rate is variable but requires a bounded delay for delivery. Examples of this service include compressed packetized voice or video. The requirement on bounded delay for delivery is necessary for the receiver to reconstruct the original uncompressed voice or video.

Class C - Connection-oriented data service. Examples of this service include connection-oriented file transfer and in general, data network applications where a connection is set up before data is transferred. This service has variable bit rate and does not require bounded delay for delivery.

Class D - Connectionless data service. Examples of this service include datagram traffic and in general, data network applications where no connection is set up before data is transferred.

Although each AAL is optimized for a specific type of traffic, there is no stipulation in the standards that AALs designed for one class of traffic cannot be used for another. In fact, many vendors of ATM equipments currently manufacture products that use AAL5 to support all the above classes of traffic, and most activities at the ATM Forum have focused on AAL5. The AAL is also important in the internetworking of different networks and services. [17] elaborates on discussion points and the issues surrounding AAL 5 design.

In this special issue, the article by Henderson [18] describes a signaling AAL protocol standard recommended by ITU, the Service Specific Connection Oriented Protocol (SSCOP), which incorporates many design principles for a high-speed protocol with lightweight (i.e., reduced processing) operation. The article also contains a list of over thirty references on the topic of AAL, including documents of ITU Recommendation.

The ATM layer provides an interface between the AAL and the physical layer. This layer is responsible for relaying cells from the AAL to the physical layer for transmission and from the physical layer to the AAL for use at the end systems. When it is inside an end system, the ATM layer receives a stream of cells from the physical layer and transmits either cells with new data or empty cells if there is no data to send. When it is inside a switch, the ATM layer determines where the incoming cells should be forwarded to, resets the corresponding connection identifiers and forwards the cells to the next link. Moreover, it buffers incoming and outgoing cells, and handles various traffic management functions such as cell loss priority marking, congestion indication, and generic flow control access. It also monitors the transmission rate and conformance to the service contract (traffic policing). Traffic management is a hotly debated topic in

the ATM Forum, which will be covered, in greater detail in the later sections of this thesis.

Figure 2.2 - ATM Cell Structure – UNI and NNI

The fields in the ATM header define the functionality of the ATM layer. The format of the header for ATM cells has two different forms, one for use at the user-to-network interface (UNI) and the other for use internal to the network, the network-to-node interface (NNI) (Figure 2.2). At the UNI, the header dedicates four bits to a function called generic flow control (GFC), which was originally designed to control the amount of traffic entering the network. This allows the UNI to limit the amount of data entering the network during periods of congestion. At the NNI, these four bits are allocated to the virtual path identifier (VPI).

Figure 2.3 – ATM Connection identifiers

The VPI and the virtual channel identifier (VCI) together form the routing field (Figure 2.3), which associates each cell with a particular channel or circuit. The VCI is a single-channel identifier whereas the VPI allows grouping of VCs with different VCIs and allows the group to be switched together as an entity. The VPIs and VCIs have significance only on the local link. The contents of the routing field will generally change as the cell traverses from link to link. For the UNI, the routing field contains 24 bits and thus the interface can support over 16 million sessions. At the NNI, the field contains 28 bits, allowing for over 268 million sessions to share a link within a subnet. [19] discusses important issues in Private Network-to-Node Interface (P-NNI) routing. The payload type indicator (PTI) field is used to distinguish between cells carrying user data and cells containing control information. This allows control and signaling data to be transmitted on a different sub channel from user data and hence separation of user and control data. The cell loss priority (CLP) bit provides the network with a selective discard capability. This bit could be set by a user to indicate lower-priority cells that could be discarded by the network during periods of congestion. For example, whereas data applications generally cannot suffer any cell loss without the need for retransmission, voice and video traffic usually can tolerate minor cell loss. One would therefore assign a higher cell loss priority to the CLP bit for voice or video traffic than

data traffic. Furthermore, the network could also use the CLP bit to highlight those cells that have exceeded the negotiated rate limit of a user.

The header error check (HEC) field is used to reduce errors in the header that cause a misrouting of the cell for one user into another user's data stream. This field contains the result of an 8-bit CRC checking on the ATM header (but not on the data). When a switch or an end system terminates the header, multiple-bit errors will be detected with a high probability. Moreover, a single-bit error can be corrected. This is desirable since ATM is intended for use on fiber optics link, where current modulation techniques have reduced the bit error rate tremendously.

The Physical Layer defines the bit timing and other characteristics for encoding and decoding the data into suitable electrical/optical waveforms for transmission and reception on the specific physical media used. In addition, it also provides cell delineation function, header error check (HEC) generation and processing, performance monitoring, and payload rate matching of the different transport formats used at this layers.

2.3 A Study of Data Flows in ATM Networks

Logical connections in ATM are referred to as virtual channel connections (VCC). A VCC is the basic unit of switching on an ATM network. A VCC is set up between two end users through the network and a variable rate, full duplex flow of fixed size cells is exchanged over the connection. VCCs are also used for user-network exchange (control signalling) and network-network exchange (network management and routing). For ATM, a second sub layer of processing has been introduced that deals with the concept of virtual path. A virtual path connection (VPC) is a bundle of VCCs that have the same endpoints. Thus, all of the cells flowing overall the VCCs in a single VPC are switched together. The virtual path concept was developed in response to a trend in high-speed networking in which the control cost of the network is becoming an increasingly higher proportion of the overall network cost. The virtual path technique

helps contain the control cost by grouping connections that share common paths through the network into a single unit. Network management actions can then be applied to a smaller number of groups of connections instead of to a large number of individual connections. The knowledge of virtual channels and virtual paths is essential for a qualitative understanding on Call Admission Control (explained later).

The process of setting up a virtual path connection is de-coupled from the process of setting up an individual virtual channel connection. The virtual path control mechanisms include calculating routes, allocating capacity and storing connection state information. To set up a virtual channel, there must first be a virtual path connection to the required destination node with the sufficient available capacity to support the virtual channel, with the appropriate quality of service. The required state information related to VC / VP mapping is established during virtual channel setup.

The end points of a VCC may be end users, network entities or an end user and a network entity. In all cases, cell sequence integrity is preserved within a VCC – cells are delivered in the same order in which they are sent. A user of a VCC is provided with a Quality of Service specification by parameters such as cell loss ratio (ratio of cells lost to cells transmitted) and cell delay variation. The types of traffic parameters that can be negotiated include average rate, peak cell rate, and burstiness and peak duration. The network may need a number of strategies to handle congestion and to manage existing and requested VCCs. At the crudest level, the network may simply deny new requests for VCCs to prevent congestion. Additionally, cells may be discarded if negotiated parameters are violated or if congestion becomes severe. In an extreme situation, existing connections may be terminated. Bandwidth management is performed for each virtual path (VP) between adjacent nodes. Multimedia traffic, such as voice, video and computer data, are integrated into ATM networks, which leads to the cell stream flowing over each VC to be very bursty [20] – [22].

2.4 An Introduction to Traffic Management in ATM Networks

In order for ATM networks to deliver guaranteed quality of service on demand while maximizing the utilization of available network resources, effective traffic management mechanisms are needed. Almost every aspect of ATM network operation, from signaling requests and routing to network resource allocation and policing, contains some traffic management mechanisms. A better understanding of the traffic service categories, QoS parameters and the aspects of the traffic contract contribute towards attaining a comprehensive grasp of the traffic management mechanisms

2.4.1 Quality of Service Attributes

The performance metrics listed below are used as parameters to gauge the Quality of Service (QoS) provided by the ATM network:

- *Peak Cell Rate (PCR)*: The maximum instantaneous rate at which the user will transmit. For bursty traffic, the inter-cell interval and the cell rate varies considerably. The PCR is the inverse of the minimum inter-cell interval.
- *Sustained Cell Rate (SCR)*: This is the average rate as measured over a long time interval.
- *Cell Loss Ratio (CLR)*: The percentage of cells that is lost in the network because of error or congestion and is not delivered to the destination. Recall that each ATM cell has a cell loss priority (CLP) bit in the header. During periods of congestion, the network will first discard cells with $CLP = 1$. Since the loss of cells with $CLP = 0$ is more harmful to the operation of the application, CLR can be specified separately for cells with $CLP = 1$ and for those with $CLP = 0$.

- **Cell Transfer Delay (CTD):** The delay experienced by a cell between network entry and exit points is called the cell transfer delay. It includes propagation delays, queuing delays at various intermediate switches, and service times at queuing points.
- **Cell Delay Variation (CDV):** This is a measure of variance of CTD. High variation implies larger buffering for delay sensitive traffic such as voice and video.
- **Burst Tolerance (BT):** This determines the maximum burst size that can be sent at the peak rate. This is the bucket size parameter for the leaky bucket algorithm that is used to control the traffic entering the network. The algorithm consists of putting all arriving cells in a buffer (bucket) which is drained at the sustained cell rate (SCR). The maximum number of back-to-back cells that can be sent at the peak cell rate is called maximum burst size (MBS).
- **Minimum Cell Rate (MCR):** This is the minimum rate desired by a user.

The QoS parameters – CDV, CTD and CLR – are negotiated during connection setup. The network specifies other QoS parameters, such as Cell Error Ratio (CER), Severely Errored Cell Block Ratio (SECBR) and Cell Misinsertion Ratio (CMR). The ITU-T has defined QoS classes by assigning pre-defined values for the QoS parameters, namely Class 1 (stringent class), Class 2 (tolerant class), Class 3 (bi-level class) and Class U (unspecified class). ATM switches are free to support additional QoS levels (or classes) for providing a finer granularity QoS support.

2.4.2 ATM Traffic Service Categories

At a very broad level, incident traffic on an ATM network can be classified into the following service categories:

Constant Bit Rate (CBR): This class is used for emulating circuit switching, where the bit rate is constant. Cell loss ratio is specified for cells with CLP=0 and may or may not be specified for cells with CLP=1.

Variable Bit Rate (VBR): This class allows users to send at a variable rate. Statistical multiplexing is used and so there may be small nonzero random loss. Depending upon whether or not the application is sensitive to cell delay variation, this class is subdivided into two categories: real-time VBR (VBR-RT) and nonreal-time VBR (VBR-NRT). While cell transfer delay is specified for both categories. CDV is specified only for real-time VBR. An example of real-time VBR is interactive compressed video while that of nonreal-time VBR is multimedia email.

Available Bit Rate (ABR): This class is designed for normal data traffic such as file transfer and email. Although the standard does not require the cell transfer delay and cell loss ratio to be guaranteed, it is desirable for switches to minimize the delay and loss as much as possible. Depending upon the congestion state of the network, the source is required to control its rate. The users are allowed to declare a minimum cell rate (MCR), which is guaranteed to the VC by the network. Most VCs will ask for an MCR of zero. Those with higher MCR may be denied connection if sufficient bandwidth is not available.

Unspecified Bit Rate (UBR): This class is designed for those data applications that want to use any leftover capacity and is not sensitive to cell loss or delay. Such connections are not rejected on the basis of bandwidth shortage (i.e., no connection admission control) and not policed for their usage behavior. During congestion, the cells are lost but the sources are not expected to reduce their cell rate. Instead, these applications may have their own higher-level cell loss recovery and retransmission mechanisms. Examples of applications that use this service are email and file transfer. Of course, these same applications can use the ABR service, if desired.

The QoS parameters for the above classes of traffic are summarized in Table 2.1.

Attribute	CBR	VBR-RT	VBR-NRT	ABR	UBR
CLR for CLP = 0	Specified			Specified	Unspecified
CLR for CLP = 1	Optional			Specified	Unspecified
CTD	Specified			Unspecified	Unspecified
CDV	Specified		Unspecified	Unspecified	Unspecified
SCR and BT	NA	Specified		Not Applicable	
PCR and CDVT	Specified			Specified	
MCR	Not Applicable			Specified	NA

Table 2.1 - ATM Layer Service Categories

The service categories defined above have a correlation with the QoS parameters listed previously. The CBR service category is defined for traffic that requires a constant amount of bandwidth, specified by a PCR, to be continuously available. So long as the source ensures that the cells it transmits fall within the PCR mandated, the network will guarantee the transmission of the cells across to the specified destinations with minimal cell loss and within the fixed bounds of maxCTD and peak-to-peak CDV.

The *real time Variable Bit Rate (VBR-rt)* is characterised by PCR (and its associated tolerance), SCR and an MBS – this parameter is instrumental in controlling the bursty nature of VBR traffic. The network attempts to deliver cells of conforming connections within fixed bounds of maxCTD and peak-to-peak CDV. *Non-real-time VBR* sources are also specified by PCR, SCR and MBS, but are less sensitive to delay and delay variation than the real time sources. The network does not specify any delay and delay variation parameters for the VBR-nrt service.

The *Available Bit Rate (ABR)* service category is specified by a PCR and an MCR that is vouched for by the network. While an ABR connection’s bandwidth should not fall below the MCR benchmark, the bandwidth allocated by the network to an ABR connection may vary during the life of a connection.

Best effort applications are serviced by the *Unspecified Bit Rate (UBR)* service category. This category does not support any service guarantees.

Table 2.2 highlights the relationship between the ATM service categories and the QoS parameters.

Attribute	ATM Layer Service Category				
	CBR	RT – VBR	NRT– VBR	UBR	ABR
Traffic Parameters:					
PCR and CDVT	Specified				
SCR, MBS, CDVT	NA	Specified		NA	
MCR	NA				Specified
QoS Parameters:					
Peak to peak CDV	Specified		Unspecified		
Max CTD	Specified		Unspecified		
CLR	Specified				*

* CLR is kept low for sources that adjust cell flow according to control feedback. Network implementation determines whether or not a quantitative value for CLR is specified.

Table 2.2 - ATM Service Categories

A new service called Guaranteed Frame Rate (GFR) is being introduced at the ATM Forum and the ITU-T. GFR is based on UBR, but guarantees a minimum rate to connections. The service also recognizes AAL5 frames and performs frame level dropping as opposed to cell level dropping.

2.4.3 Traffic Contract

To provide a guaranteed QoS, a traffic contract is established during connection setup, which contains a connection traffic descriptor and a conformance definition. However, it is not necessary for every ATM virtual connection to have a specified QoS. The

reason for this is that, if ATM supported only specified QoS connections, then a large percentage of the network resources will be wasted. This can happen when one or more connections are not utilizing the full capacity of their QoS contracts. Unspecified QoS contracts can be supported by an ATM network on a "best-effort" basis. Such best-effort services are sufficient for supporting most of the existing data applications. ABR or UBR are usually specified in the traffic contract when the ATM network is providing a best-effort service. Thus, these two classes of traffic are sometimes referred to as best-effort traffic.

The ATM traffic categories, QoS parameters and traffic contract hold a relationship, which is very critical to effective traffic management in ATM networks.

The ATM Forum Traffic Management Specification [23] defines five service categories for ATM networks as detailed in Section 2.4.2. Each service category is defined using a traffic contract and a set of QoS parameters. The traffic contract is a set of parameters that specify the characteristics of the source traffic. The QoS parameters are negotiated by the source with the network, and are used to define the expected standard of service provided by the network. The network guarantees to abide by the parameters agreed to maintain a certain QoS – as long as the source conforms to the established parameters in the traffic contract. The network need not support the agreed QoS restrictions if the source violates the clauses in the traffic contract.

PCR, SCR, MBS and MCR, as detailed in Section 2.4.1, are traffic parameters; Cell Delay Variation Tolerance (CDVT) defines the allowable variation from PCR for a connection's traffic to be conforming. The Generic Cell Rate Algorithm (GCRA) is used to confirm that the network conforms to the PCR and SCR. The GCRA is a version of the token bucket algorithm that determines if an incoming cell is conforming to the traffic specifications. Non-conforming cells might have their CLP bit set by the Usage Parameter Control (UPC) function of the network.

2.4.4 Traffic Management Schemes

In recent years there has been tremendous growth in the use of computer networks. It is expected that this growth will continue well in the coming years, fuelled by emerging Internet applications such as telephony, electronic commerce, video distribution, and video conferencing. As the bandwidth demand and the number of different applications increase, network service providers will need to support a wider range of services with various qualitative requirements, while maintaining reasonably high levels of network resource utilization. To provide this service, effective traffic management assumes a very potent role. Traffic management in communication networks deals with the controlled use of network resources to prevent the network from becoming a bottleneck. In particular, when network resources are allocated to more connections / traffic than they can effectively support, network performance for users degrades. Therefore, it is necessary to allocate and control the traffic so that the network can operate at acceptable levels even at times when the offered load to the network exceeds its capacity.

Nordstrom, Carlstrom, Gallmo and Asplund predicted in [24] that traffic control in ATM networks is likely to be implemented in a multilevel architecture (Figure 2.4), which separates the control at the network, call and cell levels. The network levels allocate resources to virtual paths, given the offered call traffic and tolerated call-blocking probabilities. The virtual path concept allows several calls to be switched and handled together, which simplifies Call Admission Control, but decreases the network utilization, in particular if the path capacities are reserved deterministically. The call level performs admission control within the path network and allocates bandwidth and switch buffer capacity to individual calls. The cell level allocates resources during the cell transfer phase, and is responsible for cell traffic enforcement at access switches and arbitration of cells during switch overload. This segregation of task levels was later followed in project development.

Figure 2.4 – Multilevel traffic controls in ATM networks

The nature of ATM makes traffic control a challenging task. In ATM, user cells are asynchronously multiplexed onto high capacity links according to actual communication needs. The bursty natures of the traffic for some applications on ISDN and the range of bit rates experienced make traffic control application on ATM employ a multi level architecture. Various frameworks have been proposed for the ATM traffic control (e.g. [25]), but it is not easy to build an efficient traffic control system because of the diversity in multimedia traffic characteristics.

The ATM solution [26] is flexible enough to transport different services with different bandwidth requirements in an integrated manner. ATM also provides the potential to obtain improved bandwidth efficiency by statistical multiplexing of variable bit rate (VBR) or bursty traffic streams. Performance requirements during the information transfer phase of a connection can range from stringent cell delay variation for delay sensitive services such as voice or video to stringent cell loss probability for loss sensitive services such as data. In [27], the author presents a distributed flow control scheme where the scheme is distributed across the user system and the switches at the

edge of the ATM network. In the Deterministic Quantized Rate Based (DQRB) scheme, it was suggested that the Distributed Flow Control system be implemented at the periphery of the network to monitor and regulate flow control of traffic entering the network. Unlike other schemes, which indiscriminately monitor and penalise the services, the intention of the control entities proposed by the author is to see that the traffic enters the network in a more regulated fashion – a simple congestion control scheme. Statistical multiplexing of different types of services with different correlation's and burstiness properties, and different QoS requirements, require elaborate congestion control mechanisms in order to avoid excessive loss of information. The problem is complicated by the fact that most of these multimedia services have poorly understood traffic characteristics. Also some new services such as VBR video exhibit sudden changes in their bit rate arrival processes, which might lead to significant and abrupt changes in the network's workload, thus adding another degree of complexity to the problem. This has led many researchers [26] to believe that new congestion and call admission control algorithms with some form of adaptive and learning capabilities are required to meet these challenges. This will be discussed in later sections.

There are varied and diverse traffic management schemes in ATM networks. Broadly speaking, the schemes can be classified into the following categories: Call Admission Control (CAC), Congestion Control, Usage Parameter Control (UPC), Network Resource Management (NRM), feedback controls and priority controls. Call Admission Control is dealt with in detail in the following chapter.

Network Resource Management (NRM) provisioning may be used to allocate network resources to separate traffic flows according to service characteristics.

Feedback controls are defined as the set of actions taken by the network and by the users to regulate the traffic submitted on ATM connections according to the state of the network.

Usage parameter control / network parameter control (UPC/NPC) is defined as the set of actions taken by the network to monitor and control traffic in terms of traffic offered and validity of the ATM connection, at the user access level and the network access points, respectively. Their main purpose is to protect the network resources from malicious as well as unintentional behaviour that can affect the QoS of other already established connections by detecting violations of negotiated parameters.

Priority Control allows the user to generate different priority traffic flows by using the CLP bit. A congested network element may selectively discard cells with low priority to protect as far as possible the NP for cells with high priority.

Layered coding technique allows adjusting the rate depending on the available resources.

Traffic shaping is a mechanism that alters the traffic characteristics of a stream on a VCC or a VPC to achieve a desired modification of those traffic characteristics.

Fast resource management operates on the time scale of the round-trip propagation delay of the ATM connection.

Feedback control and UPC/NPC schemes may be performed through a leaky bucket mechanism [28], or a spacer mechanism like the virtual scheduling algorithm (VSA). These mechanisms enforce the average bandwidth and the burst factor of a source. A possible implementation of a leaky bucket is to control the traffic flow by means of tokens – an arriving cell first enters a queue. If the queue is full, cells are simply discarded. To enter the network, a cell must first obtain a token from the token pool. If there is no token, the cell must wait in the queue until a new token is generated. Tokens are generated at a fixed rate corresponding to the average rate of the connection. If the number of tokens in the token pool exceeds some predefined threshold value, the process of token generation stops. This threshold value corresponds to the burstiness of the transmission.

The remainder of this section briefly dwells on the aspects of Congestion Control.

2.4.5 Congestion Control

Congestion Control refers to the set of actions to relieve congestion by limiting its spread and duration. Congestion control is concerned with allocating the resources in a network such that the network can operate at an acceptable performance level when the demand exceeds or is near the capacity of the network resources. These resources include bandwidth of the links, buffer space and processing capacity of the intermediate nodes. Although resource allocation is necessary even at low loads, the problem becomes more important as the load increases because the issues of fairness and low overhead become increasingly important. Without proper congestion control mechanisms, the throughput of the network may be reduced considerably under heavy load.

Congestion control lies at the heart of the general problem of traffic management for ATM networks. In general, congestion arises when the incoming traffic to a specific link is more than the outgoing link capacity. The primary function of congestion control is to ensure good throughput and delay performance while maintaining a fair allocation of network resources to the users. For unspecified QoS traffic such as ABR service, whose traffic patterns are often highly bursty and unpredictable, congestion control poses more challenges than for other services. As described in [29], one way to classify congestion control schemes is based on the layer of ISO/OSI reference model at which the scheme operates. For example, there are data link, routing, and transport layer congestion control schemes. Typically, a combination of such schemes is used. The selection depends upon the severity and duration of congestion.

One method to avoid network congestion is to accept a new ATM connection during connection setup phase only when sufficient network resources are available to provide the acceptable QoS. This is called connection admission control (CAC), which is

needed for connections where the QoS must be guaranteed. The "busy" tone on telephone networks is an example of CAC. CAC will be discussed in greater detail in the following chapter.

ATM networks present difficulties in effectively controlling congestion not found in other types of networks, including frame relay networks. The complexity of the problem is compounded by the limited number of overhead bits available for exerting control over the flow of user cells. This area is currently the subject of intense research and no consensus has emerged for a full-blown traffic and congestion control strategy. Both the types of traffic patterns imposed on the ATM network and transmission characteristics of those networks differ markedly from those of other switching networks. Most packets switched and frame relay networks carry non real time data traffic. Typically, the traffic on individual virtual circuits or frame relay connections is bursty in nature and the receiving system expects to receive incoming traffic on each connection in such a fashion. However, this is in contrast to the unpredictability of traffic patterns in ATM networks. The types of congestion control mechanisms for these networks are therefore inadequate for ATM networks. ATM layer traffic and congestion control should support a set of ATM layer Quality of Service classes sufficient for all foreseeable network services. The specification of these QoS should be consistent with network performance parameters currently under study. The design of an optimum set of ATM layer traffic controls and congestion controls should minimize network and end system complexity while maximizing network utilization.

There are a number of approaches to congestion control such as reducing the implicit flow rate controls and modifying network equipment design. In [30] two types of flow control mechanisms are described: preventive and reactive as a means to enforcing traffic management. A preventive control, which uses a static resources allocation, seems inadequate to handle rough traffic in high-speed networks. On the other hand, the reactive control schemes, which employ a closed loop feedback system, seem more attractive than the preventive control schemes, especially on ATM networks. In [31] Pitsillides and Lambert propose a congestion control strategy that integrates Call

Admission Control and flow control and uses predictive adaptive (feedback and feedforward) control techniques. It provides robust, effective and efficient congestion control with guaranteed QoS and high utilization of link capacity. These schemes aid in reducing congestion. In simple terms if, for any time interval, the total sum of demands on a resource is more than its available capacity, the resource is said to be congested for that interval. The simple definition of congestion above also enables the classification of all congestion schemes into two classes: those that dynamically increase the available resource and those that dynamically decrease the demand. All congestion control schemes require the network to measure the total load on the network and then to take some remedial action. This feedback and control strategy helps in maintaining optimum network performance. Various traffic management schemes in high-speed networks are explored in [32]-[34]. In [33], Jagannathan and Talluri propose an autoregressive moving average (ARMAX)-based adaptive control methodology to prevent congestion in high-speed ATM networks. The adaptive controller uses feedback from the network to control the source transmission rates. The network switch buffer dynamics are modelled as a non-linear discrete-time system and the ARMAX controller is designed to predict the transmission rates of the traffic source to ease congestion in the network. The ARMAX parameters are tuned to take the fluctuating nature of the network traffic conditions into consideration. The paper explores the stability of the closed loop system so as to guarantee the requested QoS. In [8], a new congestion control scheme is proposed which uses the cell loss priority (CLP) bit, the explicit forward congestion indicator and the explicit backward congestion indicator. To deal with performance monitoring management, a framework for defining a generic intelligent and integrated model for network management is described. The intelligent management uses different parameters and builds a complex but efficient control scheme.

A standard congestion controller would have, in summary, the following three basic control capabilities:

- ❑ A capability for selectively shedding load under congestion conditions through the CLP indicator belonging to the ATM cell header. If the CLP indicator is set to 1, it indicates to the network that it can be shed under congestion conditions at any network element.
- ❑ A capability where for congestion conditions it can notify network elements in the forward path. The EFCI indicator in the ATM header helps facilitate this.
- ❑ A capability for backward notification of encountered congestion conditions along a path. This refers to the EBCI possibility.

The main difficulty found in monitoring performance and, specifically, when controlling traffic and congestion, is to tune the parameters. There is a need for a platform that makes intelligent decisions based on feedback received from the network.

Any effective congestion control scheme would aim at reducing the duration of congestion. In [35], the average length of a telephone is estimated to be between 100 to several hundred milliseconds. For the case of video, the peak rate may last several seconds, but these measurements were performed without considering congestion/flow control schemes. [35] suggests that reactive flow control is ineffective when the congestion period is less than the round-trip delay. For effective congestion control, it is imperative that congestion at intermediate switching elements is quickly detected and the rate of transmission at the source nodes is accordingly moderated.

However, congestion control schemes are not always effective. In many cases, congestion in the network results in loss of cells resulting in a degradation of the QoS. Congestion control schemes have to be fair and responsive, must have a low overhead and should possess the ability to work in bad environments [36]. These desirable properties are not always present. Call Admission Control helps in overcoming these traffic management woes.

In summary, this chapter helped in highlighting the relevance of ATM in today's world of B-ISDN and the increasing omnipresence of the Internet and associated applications.

Concepts related to the hierarchical architecture of ATM were highlighted and connection establishment and data flows within ATM were traced. Contentious traffic management concerns of the day were outlined and the various approaches taken to address the issues were explained and surveyed. Special focus was given to congestion control techniques, leading up to the use of Call Admission Control for traffic management. The theory underlying CAC and the issues surrounding it will be dealt with in the next chapter.

Chapter 3

ATM Traffic Management – Call Admission Control

As highlighted in the previous chapter, ATM adopts the statistical multiplexing technique and hence, traffic control is necessary to avoid possible congestion at each network node and to achieve the Quality of Service (QoS) requested by each connection. Traffic control methods are often divided into two categories – reactive control and preventive control. Reactive control methods are restricted by the stringent real time response timelines imposed by today's high-speed networks. Preventive control is, therefore, considered to be more suitable than reactive control for high-speed networks. Call Admission Control (CAC) and bandwidth policing are two effective forms of preventive traffic control.

This chapter provides an understanding of the concepts behind CAC. For purposes of this thesis, conventional CAC schemes are those that do not employ neural networks in making the connection admission decisions. The salient features of conventional CAC schemes are highlighted and the inherent limitations they are faced with are outlined. Possible solutions to address the shortcomings of conventional schemes are briefly explored, leading to the use of neural networks as a predictive tool in making the connection admission decision.

3.1 A Review of ATM Call Admission Control

Call or Connection Admission Control (CAC) is the first line of defence for the network in protecting itself from excessive loads. CAC determines whether a call request for a Virtual Path (VP) connection is accepted or not. The objective of CAC is to restrict the access of new calls so that the cell and call level Quality of Service (QoS) is maintained within the path network. An additional objective is to maximize network revenue over time. In essence, when a user requests a new VPC or VCC, the user must specify (implicitly or explicitly) the traffic characteristics in both directions for that connection. The user selects traffic characteristics by selecting a QoS from among the QoS classes that the network provides. Before opening a communication channel, a user declares to an ATM node the cell generation characteristics, such as the minimum cell generation rate and the average cell generation rate. The decision to restrict a call is based on the availability of network resources sufficient to deliver the call level and cell level QoS. This decision requires information about the connection's traffic characteristics. To provide QoS guarantees while efficiently utilizing network resources, ATM networks need sufficient information about incoming traffic. The information on each connection's traffic is contained in its traffic parameters provided during connection setup. The CAC along each node along the end-to-end path of the connection will use these values to determine the traffic model and decide whether to accept or reject the request for setup. If the CAC at every node along the path decides that there are enough resources to provide QoS guarantees, the request is accepted; otherwise it is rejected. The ATM node accepts the new call only if the QoS of the already connected calls is expected to satisfy their required values even after the new call is connected. By accepting the connection, the network forms a *traffic contract* with the user. Once the connection is accepted, the network continues to provide the agreed upon QoS as long as the user complies with the traffic contract. To ensure that this mechanism works effectively, the bandwidth requirements of the incoming call is compared to the availability of bandwidth in the network. The available bandwidth takes into consideration the maintenance of a good service level to existing connections. Furthermore, the node always keep track of the number of cells sent by

each user and it may discard some cells if the number of cells sent by a user exceeds the declared value depending on the policing function.

[32]-[43] list out several CAC schemes proposed in recent years. Cell Delay and long term time-averaged cell loss probability are often adopted as QoS parameters. In [42] and [43] a bufferless fluid flow model is studied. A virtual bandwidth technique was suggested in [43]. An example of effective bandwidth management used for CAC was proposed in [44] and [45]. The effective bandwidth of a set of connections multiplexed on a transmission link is defined as the minimum amount of bandwidth required to achieve a desired QoS. A computationally simple approximate expression for the effective bandwidth of individual and multiplexed connections was derived in [44]. Based on the approximate expression, CAC can be performed in real-time. [46] presents a fast implementation for the CAC scheme proposed in [42] and [43]. It is shown that if information about the current traffic on the link is stored in an appropriate manner, a connection request can be accepted or declined with only two multiplications and one division.

The main objective of connection admission control (CAC) is to guarantee, for each service, an appropriate QoS while maximizing network utilization. In CAC, the decision as to whether a new VC is admitted or not is based on its impact on network performance, which depends on the traffic characteristics of the VC and on the QoS levels of the existing VCs. Several CAC methods use descriptors declared by the user at VC setup time [43], [47]. They assume that the VCs source traffic characteristics, such as the peak rate, average rate and mean burst length are given. Provided that these characteristics are accurate, the methods achieve optimum bandwidth efficiency. If the traffic over the VC could be characterised completely, we could admit the maximum number of VCs while maintaining the QoS. The average rate and mean burst length are, however, hard to estimate at VC setup time, especially for data communication. They are intrinsically *a posteriori* parameters. These, and other issues with conventional CAC schemes, are elaborated on later in this chapter.

However, in the practical design of CAC facilitating systems, there are a number of events to be considered. One of the major considerations is deriving a proper model for the traffic source as described in [48]-[50]. Scheduling mechanisms and admission policies play an important role in optimizing resource allocation in networks offering integrated services. The scheduler in a network mediates the low-level contention for service between cells of different classes. Scheduling and admission control policies are the subject of a case study in [51]. Various other events such as switch fabric design, buffering strategies, service scheduling disciplines, usage parameter control, buffer management schemes and feedback flow controls have a direct bearing on the design of traffic management and CAC systems and have to be factored as depicted in [52].

A practical approach to Call Admission Control using generalized peakedness has been proposed in [53]. A peakedness approach is used for determining the effective bandwidth at each node for CAC where the peakedness is defined as the variance to mean ratio of the number of busy servers in an infinite trunk.

Another aspect of quality of service that may be requested or assigned is cell loss priority. A user may request two levels of cell loss priority for an ATM connection. The user through the CLP bit in the cell header indicates the priority of an individual cell. When two priority levels are used, the traffic parameters for both cell flows must be specified. Typically, this is done by specifying a set of traffic parameters for high priority traffic ($CLP = 0$) and a set of traffic parameters for all traffic ($CLP = 0$ or 1). Based on this breakdown, the network may be able to allocate resources more efficiently.

Once the connection has been accepted by the Connection Admission Control function, the Usage Parameter Control (UPC) function of the network monitors the connection to determine whether the traffic conforms to the traffic contract. The main purpose of the UPC is to protect the network resources from an overload on one connection that would

adversely affect the QoS on other connections by detecting violations of the assigned parameters and taking appropriate actions. The conformance test is performed by the Usage Parameter Control (UPC) in the ATM access node at the beginning of the connections end-to-end path and it can be done using the Generic Cell Rate Algorithm (GCRA). Conforming traffic will be allowed to enter the ATM network and will receive QoS guarantees while non-conforming traffic will be either dropped or tagged as low priority. By doing so enough resources to guarantee the cell level QoS is made available to ongoing and accepted calls.

But this guarantee is not satisfactory if it comes at the cost of unreasonably high rates of call denials for a particular class. There is a need to simultaneously guarantee a certain QoS at the call level as well. In order to improve the systems performance at the call level (fairness in call blocking), it may be necessary to block additional calls as a function of the System State i.e. the current portfolio of calls utilizing the output port of the switch. The decision to accept an additional call involves the calculation or estimation of the consequences of that call acceptance on cell loss and cell delay of itself and other ongoing calls as well as global statistics such as the Call Blocking Ratio.

The Call Admission Controller must try to optimize the utilization of the switch while obeying call level QoS requirements. The three main functions of CAC are the configuration of traffic regulation, buffer management and scheduling units. Admission control is based on the knowledge of statistical characteristics of ongoing and arriving connections. The way in which these statistics are specified gives rise to two different CAC approaches:

The parametric approach: The CAC assumes that the cell arrival processes for the contending connections are completely known, via a parametric description of the arrival distribution.

The non-parametric approach: The CAC does not know the complete cell arrival process, although it knows some of its characteristics.

3.2 Limitations of Conventional Call Admission Control

Any CAC decision is based on the requested QoS by any new connections incident on the network. Under ideal circumstances, this would involve accurately predicting the resulting QoS once the additional traffic has been admitted. Lack of knowledge of traffic characteristics is one of the major obstacles in the path of making an accurate prediction. Without accurate predictions of the resulting QoS, a network may adopt an overly conservative admission policy that, while assuring certain agreed-upon QoS, lowers network utilization. More accurate predictions of QoS would enable the network to make better-informed connection admission decisions, thus resulting in better QoS and improved network utilization.

There are various levels of traffic knowledge that a network might be aware of. This will, in turn, shape the kind of CAC policy that the network implements. The various degrees of traffic knowledge are listed below:

For systems in which the access *link speed* is known, with no knowledge of the traffic that the link will carry, the *(physical) peak-rate allocation* admission policy is the only viable option. Naturally, this scheme often achieves very little efficiency.

If a user's *peak traffic rate* on the half-buffer time scale is known, the *(traffic) peak-rate allocation* policy is justifiable. This approach yields an improved efficiency, even for links serving only one connection (i.e., links whose capacity is less than the user's physical peak rate can now serve the connection). Nevertheless, with such a policy no further gains from statistical multiplexing are possible.

Knowing a user's *mean traffic rate* in addition to the *peak*, would lead to some multiplexing gains and would even allow link utilization to approach unity at very large capacities.

Additional knowledge of the *variance* of a user's traffic on the half-buffer time scale would allow further gains. One quantitative indication of these gains is that, asymptotically, the idle time approaches zero at the same rate as in the next case.

Knowing the *complete marginal distribution* of a user's traffic on the single time scale of a half-buffer work time would allow further multiplexing gains, especially at the more practical link capacities below the asymptotic regime.

Knowing the *traffic distribution on more than one time scale* should lead to further increases in utilization.

Complete traffic information would allow the use of an ideal policy. The "only" restriction would be the computational complexity in implementing the policy.

The improvements along the progression (Cases I-VII) are to be expected from the simple observation that greater knowledge provides the possibility for a greater efficiency.

Conventional CAC and congestion control schemes based on mathematical analyses provide robust solutions for different steady state traffic environments. However, their design and implementation, which directly utilize capacity estimation and buffer thresholds, suffer from some fundamental limitations.

Limitations of Conventional CAC techniques:

It is not easy to determine accurately the effective thresholds or the equivalent channel capacities because of the difficulty in obtaining timely and exhaustive information on the traffic input to the ATM network. Conventional CAC schemes that utilize either

capacity estimation or buffer thresholds suffer from some fundamental limitations. One of the limitations is the difficulty of obtaining complete statistics on input traffic to a network (as detailed above). As a result, it is not easy to accurately determine the equivalent capacity or effective thresholds for multimedia high-speed networks in various bursty traffic flow conditions.

Conventional schemes provide optimal solutions only under steady state. A control scheme that dynamically regulates traffic flow according to changing network conditions, however, requires a good understanding of network dynamics. The rationale and principles underlying the nature and choice of thresholds or equivalent capacity under dynamic conditions is incoherent. Networks are forced to make decisions based on incomplete information so that the decision process is full of uncertainty. Thus, because of the unpredictable statistical fluctuations of the system, these control schemes will always be subject to decision errors, which degrade performance.

Traditional analytical methods require complicated calculations and exhaustive simulations to derive relationships between input traffic and service quality and cannot adapt to unexpected traffic.

New CAC schemes proposed in [42] and [43] were highlighted earlier on in this section. However, these control schemes require computation of convolution, which is a difficult proposition in a real time reactionary situation. The computationally simple approximate expression suggested in [44] renders the gain arising out of statistical multiplexing ineffective and thus, becomes inappropriate when statistical multiplexing is significant.

Theoretical analyses of the limitations of conventional schemes suggest that in order for CAC to be effective in the context of today's high-speed networks, the policies need to be able to react fast in real time and most importantly, need to have accurate information about the incident traffic characteristics.

Admission control is invoked when the connection is setup. The admission decision should, therefore, be quick so as to minimize the connection setup delay. Computational complexity in the admission decision procedure should be minimized.

In addition, from the viewpoint of switching node cost, extra hardware and software control should be avoided.

The traffic contract between the incident connection and the network should be an accurate reflection of the traffic characteristics to be expected from that connection. This will enable the controller to make accurate decisions, and network resources will not be under utilized. Either that, or a mechanism should be devised whereby the traffic characteristics are predicted on the basis of incident traffic behaviour. This is where a neural network, with its self-learning capability to characterize the relationship between input traffic and system performance, can be used to enhance the performance of conventional CAC controllers.

In summary, this chapter introduced the concept of Call Admission Control. The factors that go into the network making an admission decision were highlighted. The steps that are taken from the establishment of a traffic contract by an incident connection request to the admission control decision factoring in the QoS demands on each node in the requested network path was detailed. A brief overview of the work done in this field of research was surveyed and salient features outlined. The inherent limitations with conventional CAC schemes were explored and possible alternate suggestions listed. A viable alternative to circumvent the prevailing shortcomings lies in the use of neural network for predicting traffic behaviour and dynamic bandwidth allocation. Theoretical analysis of neural networks and the role they play in the call admission decision process will be performed in the following chapters.

Chapter 4

Neural Call Admission Controllers

Neural networks, with their ability to perform time-series predictions, are used to enhance the performance on conventional CAC controllers. This chapter provides further insights into how neural networks can be incorporated into conventional CAC policies. Prior to that, the theory behind neural networks will be elucidated upon. The study of neural networks is fast gaining popularity, due to the fact that it attempts to model the brain that is found in all living creatures, which enables them to perform complicated tasks that seem to have complex reasoning. Neural networks, which consist of a large number of densely connected simple processing units often called artificial neurons or nodes, are classified by their network architectures, node characteristics and learning algorithms. Of the various types of networks being developed, Radial Basis Function (RBF) neural networks – which have been gaining popularity due to their simple structure that avoids unnecessary and lengthy calculations – will be surveyed. Special focus will be dedicated to the recently developed, MRAN and GAP networks and their special processing capabilities.

4.1 Radial Basis Function Neural Networks

In spite of the fact that modern digital computers have made great leaps in both speed and processing power, there are certain tasks that may not be performed sufficiently

well for digital computers due to the complexities associated with the problem at hand. Such tasks include optimization, pattern recognition, generalization and classification. Neural networks can solve these problems more easily than conventional digital computers.

A neural network is generally a multiple input multiple output non-linear mapping circuit, which can learn an unknown non-linear input output relation from a set of examples. It consists of many neurons connected to each other; each neuron is a multiple input single output non-linear circuit. The connection strengths between neurons are called weights, and changing the set of weight values called the weight vector can modify the neural network input-output relation. The simplest three-layered neural network consists of an input layer, a hidden layer and an output layer. Each layer is a group of many neurons and the output of a neuron in one layer is the input to all the neurons in the next layer. In the operating phase, the user sets the values of the input neurons, and the neural network produces the output values. In the training phase, the user simultaneously sets the input and the desired output values. Then all the weight values are modified according to a selected learning equation.

The networks are designed to mimic the function of the biological neurons in the brain. Even though the individual biological neurons operate at much slower rates than artificial silicon neurons, the brain can efficiently handle computationally complicated tasks, such as image processing. The ability of learning and generalizing is one that the artificial neuron has long been struggling to come to terms with. In trying to model the brain, the network architectures of the artificial neural networks have seen enhancements over time, from the simple perceptron, to the more complex multi-layer feed-forward networks (MFNs) and recurrent networks. The Radial Basis Function Neural Network (RBFNN) is one such network that has been developed. It can be viewed as a special case of the multilayer feed-forward network, when only the topologies of the networks are considered. However, RBFNN node

characteristics and the training algorithms differ greatly from MFNs. The former uses radial basis functions, while the latter usually uses sigmoidal functions for their nodes.

The radial basis function networks were originally developed for data interpolation in multi-dimensional space. It can also be seen in signal processing and literature on neural networks. It is known as a general-purpose method for approximating nonlinear mappings. Associated with the RBFNN, is a body of techniques for training the network, such as data clustering, gradient descent and recursive least squares.

In an RBF network, data is sent to the input layer following which it passes through a set of basis functions in the hidden layer – where they are expanded into hidden unit space. This hidden layer consists of computing units or hidden neurons (as they are more commonly known). The output of the RBF network is a linear combination of the outputs from its hidden units as depicted in Figure 4.1.

Figure 4.1 – Radial Basis Function Network Model

Among the various radial basis functions, the Gaussian function is highly preferred due to its ability in generalizing global mapping and refining local features without significantly altering the mapping already learned. The Gaussian function's output monotonically decreases with distance from its centre. It, thus, tends to have a response that is finite and more local. The multi-quadratic function has a global and infinite response. Both the position and the shape of the Gaussian function are more flexible and can be adjusted, as compared with other functions like the thin-plate-spline function.

RBF networks learn as a result of many presentations of a prescribed set of training examples. One complete presentation of the entire training set during the network learning process is called an *epoch*. The learning process can either be batch driven or sequential. In batch learning, the network parameters are updated after the presentation of all the training patterns that constitute an epoch. The network parameters are maintained on an epoch-by-epoch basis until the parameters of the network stabilize and the average network output error over the entire training set converges to some minimum value. In sequential learning, the network parameters are adjusted after the presentation of each training pattern. Another factor for consideration is, whether the number of hidden neurons is fixed *a priori*, or is increased as the training progresses. Users are often unable to determine correctly the number of hidden neurons required for the network to reach a desired level of performance. This difference in the construction of RBF networks has also led to the development of several different training algorithms.

While there are many contributions to the field of deriving batch-learning algorithms, for purposes of this section, mild attention will be drawn towards the work done in sequential learning algorithms. Platt [54] and Kadirkamanathan & Niranjan [55] deserve a special mention for their contribution to this field. Platt proposed a Resource Allocation Network (RAN) and Kadirkamanathan proposed an enhancement to obtain a

Resource Allocation Network via an Extended Kalman Filter (RANEKF). Both are algorithms that exhibit sequential learning and start with no computational units. The number of hidden units need not be selected *a priori*. As each input is received, a new hidden unit is allocated, based on the ‘novelty’ of the input. A new pattern is considered novel if that input pattern is far away from the existing centres and if the error between the network output and the desired output is large. If no additional units are added, the parameters of the existing neurons, such as centres, widths and weights, are updated. RAN and RANEKF differ in the manner in which the parameters are updated. Platt uses a Least Means Square (LMS) algorithm, while Kadirkamanathan and Niranjan use an Extended Kalman Filter. They showed that using the EKF improves the rate of convergence and further reduces the network complexity. The shortcoming is an increase in computational complexity. Another drawback of this approach is that the hidden units cannot be removed in both RAN and RANEKF once they are allocated. This can result in networks that are larger than required. To overcome this, a new sequential learning algorithm, known as MRAN, was developed and used to train a Gaussian based RBF network. The following section will elaborate more on MRAN – Minimal Resource Allocation Network – and its successor, GAP.

4.2 Minimal Resource Allocation Network (MRAN)

The Minimal Resource Allocation Network (MRAN) is a recently developed sequential learning algorithm for function approximation and time series prediction using a minimal radial basis function neural network [56]-[58]. In the classical approach to radial basis functions (RBF) network implementation, the basis functions are usually chosen as Gaussian and the number of hidden units (i.e., the centres and widths of the Gaussian functions) is fixed a priori based on the properties of the input data. The weights connecting the hidden and output units are estimated by a linear least squares method.

The MRAN learning algorithm succeeds the Resource Allocation Network (RAN) and the Resource Allocation Network via Extended Kalman filter (RANEKF) research initiatives. The MRAN overcomes the problems stated above by incorporating the growth criteria of RAN networks and concurrently running a neuron pruning strategy based on the relative contribution of each hidden unit to the overall network output. The resulting topology leads towards a minimal topology for RAN.

The primary disadvantage of using RAN and RANEKF is that the appended neurons are not removable. Neurons that were previously active could be redundant to the network performance during the mature stages in the life of a neural network. However, these neurons still persist in the RAN and RANEKF networks. This drawback has been aptly addressed in the MRAN. The MRAN learning algorithm allocates a new hidden neuron based on the novelty of observation and when the need arises, prunes those hidden units, which have insignificant contributions to the outputs of the network. Therefore, when applied to time varying dynamic systems, the architecture of MRAN will be adjusted automatically on line to fit closely the dynamics of the observation data.

In the MRAN algorithm, the network starts with no hidden units. As each input-output training data is received, the network is built up based on certain growth criteria. The algorithm adds hidden units, as well as adjusts the existing network, according to the data received.

The hidden layer consists of an array of neurons (ϕ_1 to ϕ_k) connected to the output by n connection weights (α_1 to α_n). The output of the hidden layer is the vector $\phi_k(\mathbf{x})$ with m inputs $\mathbf{x}$, (x_1 to x_m). The overall network response is:

$$f(x) = \alpha_0 + \sum_{k=1}^K \alpha_k \phi_k(x) \quad (1)$$

Where α_k is the weight connecting the k^{th} -hidden unit to the output unit. α_0 is the bias term. Here, K represents the number of hidden neurons in the network. α_k is a Gaussian function given by,

$$\phi_k(x) = \exp\left(-\frac{1}{(\sigma_k)^2} \|x - \mu_k\|^2\right) \quad (2)$$

Where μ_k is the centre and σ_k is the width of the Gaussian function. $\| \cdot \|$ denotes the Euclidean norm. The network begins with no hidden units. As each input-output training data (x_n, y_n) is received, the network is built up based on certain growth criteria [4]. The algorithm adds hidden units, as well as adjusts the existing network, according to the data received. Only when all these criteria are met, is a new hidden node added to the network. Each new hidden unit added to the network will have the following parameters associated with it:

$$\alpha_{k+1} = e_n, \mu_{k+1} = x_n \text{ And } \sigma_{k+1} = k \|x_n - \mu_{nr}\|. \quad (3)$$

The overlap of the responses of the hidden units in the input space is determined by k , the overlap factor. When an input to the network, does not meet the criteria for a new hidden unit to be added, the network parameters $w = [\alpha_0, \alpha_1, \mu_1^T, \sigma_1, \dots, \alpha_K, \mu_K^T, \sigma_K]^T$ are adapted in correspondence to a particular strategy.

The algorithm also incorporates a pruning strategy, which is used to prune hidden nodes that do not contribute significantly to the output of the network, or are too close to each other.

The MRAN adopts the basic idea of RANEKF and augments it with a pruning strategy to remove the hidden neurons that have insignificant contribution to the total output of

the network. This results in a minimal RBF network. The RANEKF itself was based on the RAN, with a difference in the adaptation of the parameters. Both the RAN and RANEKF had only two growth criteria and no pruning. The addition of the pruning strategy in the learning algorithm may result in oscillations in the transition of the number of hidden neurons. To overcome the oscillations and make the transition smooth, an additional growth criterion based on the Root Mean Square (RMS) error of the network output over a sliding window was introduced in the MRAN. This reduces the effect of measurement noise in the data. The addition of a third growth criteria and a pruning strategy resulted in an algorithm, which obtained a smaller network size, without compromising the accuracy of the results.

The pruning is done by observing the output of each of the hidden nodes for a period of time, and then removing the node that has not been contributing a significant output for that period. Consider the output, o_k of the k^{th} -hidden unit:

$$o_k = \alpha_k \exp\left(-\frac{1}{(\alpha_k)^2} \|x - \mu_k\|^2\right) \quad (4)$$

A node is pruned, if the output of that node falls below a threshold value for M consecutive inputs.

The pruning strategy is not just used to prune hidden nodes that do not contribute significantly to the outcome of the network, but also those that are too close to each other. The former is done by observing the output of each of the hidden nodes for a period of time and then removing the hidden node that has not been contributing a significant output for that period. The latter is performed if the distance between two hidden units are found to be closer than the threshold value and the difference between their width parameter is less than the threshold value. The two units are combined to form a single unit and the weights of the two units are added together to form the weight of the new unit. When pruning takes place, the dimensions of the EKF are

reduced accordingly, to suit the reduced network. The pruning criteria ensures that the network does not have hidden units that may have been active initially but subsequently end up contributing little to the network output. It also ensures that the hidden units do not form clusters that are too close together. If that happens, they are replaced by a single neuron, which performs the combined function of the previous two neurons. The detection and removal of such hidden units during the learning process ensures that a more parsimonious network topology is realized.

The pruning strategy can be summarized as follows:

If a centre's normalized contribution to the output for a certain number of consecutive inputs is found to be below a threshold value, that centre is pruned

If two centres are found to be close to each other, as defined by a threshold value, the two centres are combined into a single centre

The dimensions of the EKF are adjusted and the next input is evaluated.

Like the RANEKF and RAN, the learning process for MRAN involves the allocation of new hidden units as well as adjusting the network parameters. In all the three algorithms, the network begins with no hidden units. As each observation, consisting of an input-output training data, is received and processed, the network is built up based on certain growth criteria. The algorithm adds hidden units, as well as adjusts the existing network, according to data received.

The sequential learning algorithm for MRAN, can be summarized as follows.

1. Obtain an input and calculate the network output and the corresponding errors e_n and e_{rmsn}
2. Create a new Radial Basis Function centre

- if the error e_n exceeds a minimum threshold value
- the rms. error e_{rmsn} of the network for a series of post data has been above a certain threshold value and
- the new input is sufficiently far from the existing centres.

3. If condition (2) is not met, adjust the weights and widths of the existing RBF network using the Kalman Filter Algorithm.

The MRAN has achieved significant performance enhancements in a wide array of problems with dynamic specifications. Since the application of MRAN to traffic management of ATM networks, particularly in Call Admission Control, has not been the focus of any previous research initiatives, MRAN was designated as the neural network of choice for purposes of this research.

4.3 MRAN Applications

MRAN has been successfully applied to address problems in different areas such as function approximation, pattern classification, time-series prediction and system identification. MRAN has been used previously to approximate a static function that consists of a summation of six exponential functions. It has also been applied to some benchmark problems and its performance compared with previous learning schemes for both RBF neural networks and MFNs.

The first of the benchmark problems was the problem of the Hermite polynomial [57]. The Hermite polynomial was first employed for the data-modelling problem by using the Bayesian interpolation approach and was adopted by Kadirkamanathan and Niranjan for function approximation using both RAN and RANEKF. It was shown that the MRAN achieved the smallest approximation error with the most compact network size, when compared with both RAN and RANEKF. The MRAN was also used to approximate two continuous functions with multiple inputs and multiple outputs.

The MRAN was also used to solve the benchmark problems presented in the PROBEN1 database. PROBEN1 is a collection of benchmark problems for neural network learning in the realm of pattern classification and function approximation including a set of rules and conventions for carrying out benchmark testing with these or similar problems. There are 15 datasets from different domains in PROBEN1 – these represent realistic problems suited for supervised learning. The MRAN performance for two function approximation problems (*hearta* and *flare*) and one pattern classification problem (*cancer*) from this dataset as evaluated. From the results of the *hearta* problem, it was found that, compared to MFN, MRAN produced a smaller approximation error with a more compact network. It was also observed that the data was only required to be presented to MRAN once, as compared with 40 epochs for the MFN. The magnitude of the number of free network parameters used by MRAN was also one or two orders less than that of the MFN. For the *flare* problem, the approximation error of MRAN was similar to that of the MFN, but achieved with a smaller network. The MFN required 50 epochs, as compared with one epoch for MRAN. The network parameters used for MRAN were also considerably smaller. Comparison of the results for the *cancer* problem was similar to those observed for *flare*.

As the benchmark problems addressed above are static problems that do not deal with time, the MRAN was applied in dynamic problems for non-linear systems. These included non-linear time series prediction, non-linear signal processing and non-linear system identification problems.

The highly successful track record and the vast diversity of applications that MRAN has been incorporated in prompted its use in the enhancement of ATM traffic management. MRAN has also been enhanced and its latest version is called GAP – which will be detailed in the following section.

4.4 Growing and Pruning (GAP) Neural Network

Significance is defined as a neuron's statistical contribution to the overall performance of the network. In the Growing and Pruning (GAP) RBF algorithm [59], Sundararajan, Huang and Saratchandran have used this significance in growing and pruning strategies. For growing, apart from the distance criteria used in MRAN, GAP uses a criterion based on the significance of the neuron. A new neuron will be added only if its significance is more than the chosen learning accuracy. If during training, the significance of the neuron becomes less than the learning accuracy, then the neuron will be pruned. Further, for both growing and pruning, it is shown that one needs to check only the nearest neuron (based on the Euclidean distance to the latest input data) for its significance. If the input data does not require a new hidden neuron to be added, then the parameters of only the nearest neuron are adjusted, resulting in a reduction in the overall computation and thereby increasing learning speed. The GAP RBF algorithm is truly sequential and can be used for online learning in real time applications where the training observations are sequentially (one by one) presented and discarded after being learned. The learning (parameter adjustment, network growing or pruning) is conducted whenever a new observation is presented.

4.4.1 Growing Criterion

The learning process of GGAP-RBF involves the allocation of new hidden neurons as well as adaptation of network parameters. The RBF network begins with no hidden neurons. As inputs are received sequentially during training, some of them may initiate new hidden neurons based on a growing criterion as follows:

$$\|x_n - u_{nr}\| > \varepsilon_n$$

$$\|e_n\|_q \left(\int_X \exp \left(-\frac{q \|x - x_n\|^2}{k^2 \|x_n - \mu_{nr}\|^2} \right) p(x) dx \right)^{1/q} > e_{\min} \quad (1)$$

(1)

where x_n is the latest input received, μ_{nr} is the center of the hidden neuron nearest (in the Euclidean distance sense) to x_n . $e_{\min}$ is the expected approximation accuracy and n is a threshold to be selected appropriately.

When a new neuron is added, the parameters associated with it are given by

$$\begin{aligned} \alpha_{K+1} &= e_n \\ \mu_{K+1} &= x_n \\ \sigma_{K+1} &= \kappa \|x_n - \mu_{nr}\| \end{aligned} \quad (2)$$

where $e_n = y_n - f(x_n)$ and $\hat{e}$ is an overlap factor that determines the overlap of the responses of the hidden neurons in the input space.

This is the statistical contribution of neuron k to the overall output of the RBF network and we define this as the “significance” of a specified neuron k , $E_{sig}(k)$, and is given by

$$E_{sig}(k) = \lim_{n \rightarrow +\infty} E_q(k) = \|\alpha_k\|_q \left(\int_X \exp \left(-\frac{q \|x - \mu_k\|^2}{\sigma_k^2} \right) p(x) dx \right)^{1/q} \quad (3)$$

Remark: The first criterion ensures that a new neuron is only added if the input data is sufficiently far from the existing neurons. The second criterion ensures that the significance of the newly added neuron obtained by substituting the first equation (23) in the second equation above is greater than the required approximation accuracy e_{min} . Also the second criterion subsumes the novelty criterion used in RAN, RANKEF and MRAN where the condition $e_n < e_{min}$ is obviously implied in our enhanced growing criterion based on the significance of newly intentionally added neuron.

$$\|e_n\|_q \left(\int_X \exp\left(-\frac{q\|x - x_n\|^2}{\kappa^2\|x_n - \mu_{nr}\|^2}\right) p(x) dx \right)^{1/q} > e_{min} \quad (4)$$

4.4.2 Pruning Criterion

If the significance of neuron k is less than the approximation accuracy e_{min} , neuron k is insignificant and should be removed otherwise, neuron k is significant and should be retained. Given the approximation accuracy e_{min} , neuron k will be pruned if

$$E_{sig}(k) = \lim_{n \rightarrow +\infty} E_{ave}(k) = \|\alpha_k\|_q \left(\int_X \exp\left(-\frac{q\|x - \mu_k\|^2}{\sigma_k^2}\right) p(x) dx \right)^{1/q} < e_{min} \quad (5)$$

The above condition implies that after learning each observation, the significance for all neurons should be computed and checked for possible pruning. This will be a computationally intensive task. But it is shown in the next subsection that only the nearest neuron is possibly insignificant and need to be checked for pruning and there is no need to compute the significance of all neurons in the network.

4.4.3 Nearest Neuron for Parameter Adjustment and Pruning

In order to increase the learning speed further it can be shown that, (instead of adjusting the parameters and conducting pruning checking for all neurons after each observation), one needs only to adjust parameters of the neuron nearest (in the Euclidean distance sense) to the most recently received input if no new neuron is added and only needs to check the nearest (most recently adjusted) neuron for pruning. It is neither necessary to adjust the parameters for all neurons nor necessary to check all neurons for possible pruning. At any time instant, only the single nearest neuron needs to be adjusted or needs to be checked for pruning. The rationale for this is given as follows.

Compared to the Gaussian function, its first and second derivatives will approach zero faster when $x > \sigma$. Thus, in the gradient vector expressed in the Extended Kalman Filter, all elements except the following will approach zero quickly when $x > \sigma$.

$$\phi_{nr}(x_n), \phi_{nr}(x_n) \frac{2\alpha_{nr}}{\sigma_{nr}^2} (x_n - \mu_{nr})^T, \phi_{nr}(x_n) \frac{2\alpha_{nr}}{\sigma_{nr}^3} \|x_n - \mu_{nr}\|^2 \quad (6)$$

This would dramatically increase the learning speed by removing the “curse of dimensionality” of RANEKF and MRAN but without losing learning performance in most cases. In fact, when an observation is presented, RANEKF and MRAN needs to calculate Kalman gain vector and error covariance matrix by using a $3K$ gradient vector, where K is the number of neurons obtained so far. When K becomes large, this would result in a larger learning time and possibly make ordinary computers incapable of learning because of lower memory configuration, etc. However, since only one neuron’s parameters are adjusted at any time, GAP-RBF only needs to do simple matrix operation on a very small 3×3 matrix no matter how many neurons have been obtained. Suppose that after sequentially learning n observations, a RBF network with K neurons has been obtained. Obviously all these K neurons should be significant since

insignificant neurons would have been pruned after learning the n -th observation. If a new $(n + 1)$ -th observation (x_{n+1}, y_{n+1}) arrives and the growing criteria is satisfied, a new significant neuron $K + 1$ will be added. The parameters of all the rest neurons remain unchanged and those neurons will remain significant after learning the $(n+1)$ -th observation, and the new added neuron is also significant, thus, pruning checking need not be done after a new neuron is added. If a new observation (x_{n+1}, y_{n+1}) arrives and the growing criteria is not satisfied, no new neuron will be added and only the parameters of the nearest neuron will be adjusted. Since the parameters of all the neurons except for the nearest one remain unchanged, those neurons except for the nearest one will remain significant after learning the $(n+1)$ -th observation. After the parameters of the nearest neuron are adjusted, if the nearest neuron becomes insignificant, it should be removed. That means, if only the parameters of the nearest neuron are adjusted after each observation during sequential learning, one needs to check only whether the nearest neuron becomes insignificant after adjustment. As for the parameter adjustment, the adjustment will be done for the parameters of the nearest neuron using the EKF algorithm. In EKF algorithm, the Kalman gain vector computation becomes dramatically simpler because we are adjusting only one neuron. Thus, a new simple efficient generalized growing and pruning RBF algorithm suitable for multi-input multi-output applications and with any sampling density function $p(x)$ is summarized below.

The GGAP- RBF Algorithm:

For each observation (x_n, y_n) presented to the network, where $x_n \in X \subseteq R^l$

Compute the overall network output

$$f(x_n) = \sum_{k=1}^K \alpha_k \exp\left(-\frac{1}{\sigma_k^2} \|x_n - \mu_k\|^2\right) \quad (7)$$

where K is the number of hidden neurons.

Calculate the parameters required in the growth criterion.

$$\begin{aligned} \varepsilon_n &= \max \left\{ \varepsilon_{\max} \gamma^n m \varepsilon_{\min} \right\}, (0 < \gamma < 1) \\ e_n &= y_n - f(x_n) \end{aligned} \quad (8)$$

Apply the criterion for adding neuron

If

$$\|x_n - \mu_{nr}\| > \varepsilon_n \text{ and } \|e_n\|_q \left(\int_X \exp \left(-\frac{q \|x - x_n\|^2}{\kappa^2 \|x_n - \mu_{nr}\|^2} \right) p(x) dx \right)^{1/q} > e_{\min}$$

allocate a new hidden neuron $K+1$ with

$$\begin{aligned} \alpha_{K+1} &= e_n \\ \mu_{K+1} &= x_n \\ \sigma_{K+1} &= \kappa \|x_n - \mu_{nr}\| \end{aligned}$$

Else

Adjust the network parameters $\alpha_{nr}, \mu_{nr}, \sigma_{nr}$ for the nearest neuron only.

Check the criterion for pruning the adjusted hidden neuron:

If

$$E_{sig}(nr) = \|\alpha_{nr}\|_q \left(\int_X \exp \left(-\frac{q \|x - \mu_{nr}\|^2}{\sigma_{nr}^2} \right) p(x) dx \right)^{1/q} < e_{\min}$$

Remove the nr -th hidden neuron

Reduce the dimensionality of EKF

Endif

Endif

GAP has been shown to perform very competitively when compared against MRAN, RANEKF and RAN in standard benchmark problems such as California Housing and Mackey-Glass chaotic time series prediction. The GGAP-RBF algorithm can be used for any arbitrary input sampling distribution. The key difference between MRAN, RANEKF and GAP is that the latter reduces the computation complexity by only adjusting the parameters of the nearest neuron every time instead of all neurons without losing learning performance.

4.5 Neural Networks for Traffic Management of ATM Networks

As mentioned in the previous chapter, inadequate knowledge of the incident traffic characteristics and the dynamic nature of ATM networks render conventional traffic management schemes ineffective.

Neural networks are able to address issues where conventional traffic management has been found to be deficient. Conventional schemes are limited to making admission control decisions based on QoS requests from incident traffic sources. The QoS requests do not entirely factor in the dynamic nature of changing traffic conditions in the network. Lack of knowledge of random time varying traffic characteristics in an ATM network is one of the major obstacles in the path of making an optimal decision.

Conventional schemes would benefit from more accurate predictions of traffic conditions in the network. Admission control decisions that factor in these conditions would result in better network utilization. Neural networks are used in call admission

control schemes to predict the random time varying traffic conditions in the network based on past data. Other methods of prediction are generally linear in nature and are not able to adapt to the non-linear randomly varying nature of ATM traffic. The ability of a neural network to adaptively learn about the underlying non-linear traffic conditions prevalent in the ATM network and to generalize from the observations made is very useful for traffic control in the network. The network uses the predictions of traffic rate to make better-informed call admission control decisions. The self-learning capability of the neural network is used to characterize the relationship between input traffic and system performance. Neural network adaptability is used to learn the relation between offered traffic and service quality from observed data. The neural network learns how to adapt to unexpected traffic as a non-linear function from noisy data. Traffic control using neural networks can adapt to changes in traffic characteristics and to the addition of new communication services. Neural network capability of function learning can also be used for prediction. Such capability of a neural network could be used in various functions for cell transfer, call control and network control level.

In [12], a simple 1-5-1 back-propagation neural network has been used to predict Variable Bit Rate traffic accurately. Based on the prediction results obtained from the neural network, a dynamic bandwidth allocation system for ATM has been proposed. Similarly, in [60] a neural network facilitated cell loss rate estimation method for real time CAC in ATM networks has been proposed. In general, neural networks can be used in a network for the dedicated prediction of a particular parameter or they can be distributed for multi-parameter prediction.

In the hierarchical traffic control system, each control function strongly depends on the performance of the others. A small change in the control function requires changes in other control functions. However, the complete analysis of such a complicated traffic control system is not easy for traditional CAC approaches. Then the adaptability of the neural network is used to integrate the whole traffic control system. Changes in one control function are absorbed by adaptability in other functions. In other words,

efficient use of adaptability in one control function requires adaptability in the other control functions. The integration of distributed neural networks with interconnections was proposed by Hiramatsu [61] to integrate the whole ATM traffic control system and to optimize network throughput. Moreover, this adaptive traffic control concept could be extended to off line network design.

Figure 4.2 –Neural Networks in a traffic control system

Figure 4.2 illustrates a typical application of neural networks in ATM congestion control scheme. The inputs to the neural controller are $n(k)$, $n(k-1)$ and $u(k-1)$, $u(k-2)$ where at sample k , $n(k)$ is the number of cells in the buffer within the ATM multiplexer and $u(k)$ is the feedback control signal produced by the controller.

$u(k)$ is used to control the source-coding rate, as defined by the equation:

$$C(k) = C_0 u(k)$$

Where $C(k)$ = controlled coding rate at sample k

$C_o(k)$ = maximum uncontrolled coding rate of the source

The neural network predicts the control signal $u(k)$ using the above mentioned inputs and also based on the performance criterion based on a function of $[(n_d - n(k))^2 - (u_d - u(k))^2]$. Here the first term $n_d - n(k)$ represents the deviation of the number of the cells in the buffer with its maximum capacity and the second term $u_d - u(k)$ tries to keep the control signal change not large.

The inherent traffic conditions in the network are thus taken into consideration and the neural networks used to predict the future traffic rate and also control the congestion in the network.

The benefits of using neural networks for traffic control are:

Adaptive learning: A detailed mathematical understanding about the underlying network to be controlled is not required since neural networks can learn from observations or examples during the course of network operation.

High Computation rates: Due to the massive parallelism of the hardware implementation

Generalization from learning: A neural network has the ability to generalize learning to what has never been seen. This is particularly useful for learning a dynamic environment for congestion control in ATM networks where observations may be incomplete, delayed or partially available.

High degree of robustness or fault tolerance: due to the nature of distributed processing.

When incorporated into a network, the neural network creates a decision function by learning the behaviour of the operating multiplexer. When the characteristics of the multiplexed calls change, the neural network correctly modifies its nonlinear function

by interpolating or extrapolating from the past experienced results. Initially all the weights of the neural network are set to near optimal values. Then these initial weights are adjusted to the optimal values through training with real traffic patterns. This method is suitable to a situation where the traffic characteristics rapidly change because the neural networks can adaptively control by learning the relationship between the offered traffic characteristics and the service qualities from the results of actual network operation.

With these capabilities, there is definitely a need for neural networks in the management of traffic on ATM networks. Fuzzy logic and neural network systems have been widely employed to deal with CAC related problems in the ATM networks. An adaptive neural network admission controller for dynamic bandwidth allocation has also been proposed in [62]. Extensive research on approaches such as the ones described above and others have been done to investigate the viability of using neural networks in the resolution of traffic management problems. Habib [63] has conducted research in this area and neural networks have been proposed to provide a flexible high performance solution to queue management in ATM networks. Comparative studies done in [64] to [67] have shown that the proposed neural network approaches significantly improve system performance compared with conventional approaches.

However, there are certain drawbacks in the use of neural networks. The main problems of adaptive ATM QoS control methods using neural networks were the exponentially wide range of the output target and the real time training data sampling. But new practical techniques to overcome these problems may open new neural network applications [68]. However, realizing large neural networks presents some engineering difficulties such as determining the number of neurons in each hidden layer, ascertaining the number of hidden layers in a feed-forward neural network and ensuring the quality of near optimal solutions in feedback neural networks.

Nevertheless, the pros outweigh the cons and intense research is being presently conducted in the development of applications of neural networks to global network management and the integration of multiple levels of control. Theoretical analysis of the growing and pruning methodology used in the MRAN and GAP suggests that these neural networks when deployed to aid in ATM traffic management will stand in as effective predictive tools for more optimal admission control decisions.

In summary, this chapter helped to introduce the concepts behind neural networks and their working. Special focus was brought to the RBF neural networks and their sequential learning algorithms. This led to the introduction of the RAN and RANEKF neural networks and the enhanced MRAN and GAP networks. The applications of these neural networks were briefly highlighted. The utility of neural networks to overcome the shortcomings of conventional CAC and traffic management schemes was outlined – leading up to the applicability of MRAN and GAP in resolving ATM traffic management woes. The next chapter describes the steps taken to setup the simulation exercise and to investigate the performance of ATM Call Admission Control using OPNET.

Chapter 5

OPNET Simulation Setup for ATM Networks and Call Admission Control Schemes

The chapters presented hitherto have covered material that is very relevant to the subject matter of the research undertaken. The chapters henceforth will detail the approach taken to setup and execute the simulations necessary to ratify the research objective. The performance results will be presented and relevant deductions drawn. Specifically, this chapter will walkthrough the various aspects of the simulation setup and execution. OPNET Radio/Modeler v8.0 is a complex simulation tool and for reasons of comprehension and continuity in future research efforts, it was considered important to elucidate in detail the steps taken to setup and execute the simulations. This chapter will provide insights, among other areas, into OPNET's hierarchical modelling structures, the Application & Profile configuration setup, the configuration of the various elements on the network diagram, capturing of link throughput and incorporation of the neural networks into the prevailing conventional CAC scheme.

5.1 Relevant features of OPNET Radio / Modeler v8.0

The network model design, construction and analysis was done using the advanced network simulation tool – Optimized Network Engineering Tools (OPNET) Modeler / Radio ver 8.0.

OPNET has various product offerings – each targeting a separate customer segment. In its entirety, OPNET's offering is a vast software package with an extensive set of features designed to support general network modelling and to provide specific support for particular types of network simulation projects.

Specifically, OPNET Modeler is primarily used by academic and Research and Development facilities for investigative and research purposes. It is a comprehensive software environment for modelling, simulating and analyzing the performance of communication networks, computer systems and applications, and distributed systems, to provide insights into the performance and behaviour of existing or proposed networks, systems and processes. Besides offering various tools for setup and investigative purposes, Modeler also offers a development kit for researches to customize OPNET's standard model offerings to suit individual research needs. The development environment includes tools for model design, simulation, data collection and analysis. OPNET employs a unique modelling methodology that provides a powerful and flexible environment for model builders. It has detailed protocol models, an extensive library of application and vendor specific device models and the ability to import traffic topology information.

OPNET can be used as a platform to develop models of a wide range of systems. Some examples of possible applications include Standards-based LAN and WAN performance modeling, Internet work planning, research and development in communications architectures & protocols, distributed sensor and control networks,

“on-board” systems, resource sizing, Proto-C, mobile packet radio networks and satellite networks

The OPNET Modeler tools (Figure 5.1) fit into three categories: model development, simulation execution and results analysis. The model development tools include the network editor, node editor, process editor, packet format editor and parameter editor. The network, node and process editors are the three main editors used by model developers. The network editor defines or a change network topology models and is used to construct network models. The node editor defines or changes data models and is used to create models of nodes. The process editor defines or changes control flow models.

Figure 5.1 – OPNET Modeler / Radio Version 8.0 Common Desktop Environment

The simulation execution category in OPNET Modeler is very advanced and provides an integrated suite of varied features. This category includes the following tools: probe editor and simulation tool. The probe editor allows the user to attach probes to places that will be monitored and allows the user to see the dynamic operation of the network. The simulation tool defines and runs simulation studies using models constructed with the OPNET Modeler editors.

OPNET Modeler divides the majority of model specifications into a set of three environments called modelling domains. This differs from the approach taken by most modelling frameworks, which use a single paradigm to specify all aspects of the system. The contrast between the OPNET approach and the single paradigm approach is most evident when considering the problem of modelling both a systems' behaviour and its structure. Communication networks and distributed systems typically encompass a wide range of technologies ranging from low level communications hardware to high level decision making software. A successful system modelling effort must represent each of these subsystems and their interactions at a level of detail that is sufficient to obtain valid predictions of performance and behaviour. The nature of these subsystems varies significantly from level to level. For instance, a communications protocol and a real time operating system are fundamentally different types of entities than communications links or packet buffers. As a result, single paradigm frameworks must be "stretched" to develop adequate models. OPNET's approach to modelling is based on three paradigms that specifically target the distinct levels identified in a communications network. This allows the structure of models to closely resemble that of real world systems and to be more intuitive.

OPNET possesses several features that are immensely beneficial for the network designer:

The tool provides for a three layer modelling hierarchy.

The OPNET ATM model suite (AMS) described in [9] supports many of the characteristics of ATM networks. The model suite provides support for signalling, call setup and tear down, segmentation and reassembly of cells, cell transfer, traffic management and buffer management.

OPNET's Process Editor expresses process models in a language called Proto-C, which is specifically designed to support development of protocols and algorithms. Proto-C is based on a combination of state transition diagrams, a library of high level commands known as Kernel Procedures and the general facilities of the C or C++ programming language. This systematic flow along with the general level of familiarity with C /C++ makes understanding of available code and the addition of new code easier.

Performance evaluation and trade off analysis require large volumes of simulation results to be interpreted. OPNET includes a sophisticated tool for graphical presentation and processing of simulation output.

Specialized in communication networks and information systems. OPNET provides many constructs relating to communications and in-formation processing, providing high leverage for modelling of networks and distributed systems.

Graphical specification. Wherever possible, models are entered via graphical editors. These editors provide an intuitive mapping from the modelled system to the OPNET model specification.

Flexibility to develop detailed custom models. OPNET provides a flexible, high-level programming language with extensive support for communications and distributed systems. This environment allows realistic modelling of all communications protocols, algorithms, and transmission technologies.

Automatic generation of simulations. Model specifications are automatically compiled into executable, efficient, discrete-event simulations implemented in the C programming language. Advanced simulation construction and configuration techniques minimize compilation requirements.

Application-specific statistics. OPNET provides numerous built-in performance statistics that can be automatically collected during simulations. In addition, modellers

can augment this set with new application-specific statistics that are computed by user-defined processes.

Interactive analysis. All OPNET simulations automatically incorporate support for analysis via a sophisticated interactive “debugger”.

Several positive features available in OPNET Modeler imminently qualified it as the designated simulation tool to setup and validate the research objectives.

5.2 A Review of the Hierarchical Model Architecture in OPNET

OPNET provides a comprehensive development environment for modeling and performance-evaluation of communication networks and distributed systems. The package consists of a number of tools, each one focusing on particular aspects of the modeling task. These tools fall into three major categories that correspond to the three phases of modeling and simulation projects: *Specification*, *Data Collection and Simulation*, and *Analysis*

These phases are necessarily performed in sequence. They generally form a cycle, with a return to Specification following Analysis. Specification is actually divided into two parts: initial specification and re-specification, with only the latter belonging to the cycle, as illustrated in the following figure.

Figure 5.2 - Simulation Project Cycle

OPNET Modeler offers a stable of technologies and protocols to build network topologies from. These range from IP to Wireless to proprietary elements from Cisco, 3COM and other corporations. The ATM Modelling Suite (AMS) is a very comprehensive set of ATM network elements comprising switches, routers, links, traffic sources and sinks. Across all its protocols offerings, Modeler conforms to a very strict hierarchical architecture. Every element, irrespective of the protocol to which it subscribes, is broken down to three levels at a very minimum:

The Network Domain

The Node Domain

The Process Domain

The Process Domain is intricately linked to the codes that implement protocol-specific behavioural traits. The Node and Process Domain could be in turn linked to stand alone child process and external files that perform specific functions whose instances can be triggered from various points of call anywhere in the Node and Process Domain.

The *Network Domain*'s role is to define the topology of a communication network. The communicating entities are called nodes and the specific capabilities of each node are defined by designating their model. Within a single network model, there may be many nodes that are based on the same node model; the term "node instance" is used to refer to an individual node, in order to distinguish it from the class of nodes sharing the same model. Most nodes require communicating with some or all other nodes to perform their function in a network model. Several different types of communication link architectures are provided to interconnect nodes that communicate with each other. To break down complexity and to simplify network protocols and addressing, many large networks make use of an abstraction known as a sub network. A sub network is a subset of a larger network's devices that forms a network in its network models. Different types of communication links, depending on the type of sub network, can

then connect these sub networks. The larger sub network can then be viewed as a network of its sub networks. This abstraction can then be carried out at many levels. In other words, one can form networks of sub networks and so on. At the bottom of this hierarchy, the lowest level sub network is composed only of nodes and links, but no other sub networks.

The *Node Domain* provides for the modelling of communication devices that can be deployed and interconnected at the network level. In OPNET terms, these devices are called nodes and in the real world they may correspond to various types of computing and communicating equipment such as routers, bridges, workstations, terminals, mainframe computers, file servers, fast packet switches, satellites and so on. Node models are expressed in terms of smaller building blocks called modules. Some modules offer capability that is substantially predefined and can only be configured through a set of built in parameters. These include simple message generators and various transmitters and receivers allowing a node to be attached to communicate links in the network domain. Other modules, called processors, queues and external system modules, are highly programmable, their behaviour being prescribed by an assigned process model. A node model can consist of any number of modules of different types. Essentially, the modelling paradigm selected for the Node Domain was designed to support general modelling of high-level communication devices.

The *Process Domain* facilitates the programming of the tasks that the node modules execute. Queue and processor modules are user-programmable elements that are key elements of communication nodes. The tasks that these modules execute are called processes. A process can in many ways be thought of as similar to an executing software program, since it includes a set of instructions and maintains a state memory. Processes in OPNET are based on process models. The relationship between process model and process is similar to the relationship between a program and a particular session of that program running as a task. The process-modelling paradigm of OPNET supports the concepts of process groups. A process group consists of multiple processes that execute within the same processor, queue or external system module. When a

simulation begins, each module has only one process, termed the root process. This process can later create new processes, which can in turn create others as well. When a process creates another one, it is termed the new processes' parent; the new process is called the child of the process that created it. Processes that are created during the simulation are referred to as dynamic processes.

Only one process may be executing at any time. A process is considered to be executing when it is progressing through new instructions that are part of its process model. When a process begins execution it is said to be invoked. A process that is currently executing can invoke another process in its process group to cause it to begin executing. When this happens, the invoking process is temporarily suspended until the invoked process blocks. A process blocks by indicating that it has completed its processing for its current invocation. Once the invoked process has blocked, the invoking process resumes execution where it had left off, in a manner similar to the procedure-call mechanism in a programming language such as C.

Processes in OPNET are designed to respond to interrupts and/or invocations. Interrupts are events that are directed at a process and that may require it to take some action. They may be generated by sources external to a process group, by other members of a process group, or by a process for itself. Interrupts typically correspond to events such as messages arriving, timers expiring, resources being released, or state changes in other modules. Once a process has been invoked due to an interrupt, it may invoke other processes in the group and these may in turn invoke other processes, etc. An interrupt's processing is completed when the first process that was invoked blocks.

OPNET's Process Editor expresses process models in a language called Proto-C, which is specifically designed to support development of protocols and algorithms. Proto-C is based on a combination of state transition diagrams (STDs), a library of high level commands known as Kernel Procedures and the general facilities of the C and C++ programming languages. A process model's STD defines a set of primary modes or

states that the process can enter and, for each state, the conditions that would cause the process to move to another state. The condition needed for a particular change in state to occur and the associated destination states are called a transition. Proto-C models allow actions to be specified at various points in the finite state machine. The actions are extremely general in nature since they are expressed as C or C++ statements. In addition, because Proto-C is focussed on modelling protocols and algorithms, it provides an extensive library of over 300 Kernel Procedures that can be invoked to perform commonly needed actions. Kernel Procedures are grouped into packages of related functions.

In a typical simulation setup exercise, the top down approach is taken. The exercise would commence with identifying the protocol(s) to be used for the project. Once a protocol has been isolated, for example ATM, using OPNET's palette of network elements from its ATM Model Suite (AMS) Figure 5.2, the network level configuration is finalized.

The network level configuration includes customizing the parameters for every element on the network diagram. The parameter setup would largely govern the behaviour of the individual elements and the network on the whole.

Once the network configuration has been completed, each node instance can be dissected into its own node level composition – which comprises of processors such as message generators, traffic sources, sinks, queues, buffers, etc. A sample node model is depicted in Figure 5.3

Each element on the node domain is made up of individual process models consisting of finite state machines, the codes governing the behaviour of each state and the interfaces between each state. At the lowest level, the behaviour of processes (protocols, algorithms, applications) is specified using finite state machines and extended high-level languages. Each of the node modules therefore has their own process maps defined by FSMs. Since processes always create dynamic processes, the root process must be involved in originating them. The first generation dynamic processes are those that are created by the root. These may in turn create new processes, which would be referred to as second generation. With respect to a process that creates it, a dynamic process is referred to as a child process; conversely the creating process is referred to as the parent process. The root process is the ultimate “grandparent” of all processes. The root process and dynamic child processes can in turn make reference to external files for the definition and execution of specific functions. Modeler provides the capability of customizing Process and Node models – the Process and Node user interfaces provide options to compile the customized models upon completion. A sample process model is depicted in Figure 5.4.

External files are executable Proto-C or C++ files. These are customized programs written in OPNET customized versions of C / C++. As previously stated, external files hold stand-alone functions or procedures that can be referenced from parent or child process models. These can be customized to include C/C++ code and compiled by OPNET’s inbuilt compiler.

Customized processes, nodes and external files can be stored in user-defined directories and the simulations can be defined to point to specific directories to retrieve tailor-made models.

An understanding of the hierarchy discussed above and the subsequent dissection of the various parameters is essential for a better appreciation of the basis of implementation

Figure 5.6– ATM Network Model

The network model consists of a few primary element types listed below.

Application Configuration Panels

Profile Configuration Panels.

Traffic sources and Data Servers

ATM switches

Links

The following sections will provide an understanding of the role that each of these element types play and the steps taken to configure them.

Once the network topology has been constructed, the network traffic models were created next. In general, there are two techniques for representing network traffic: the first technique is to import traffic as “routed background traffic” and the second technique is to model application traffic by setting up various application attributes. The second technique was adopted in this project.

OPNET provides global objects for defining profiles and applications. The advantage of using a global object is that once they have been defined, they can be re-used across the entire network topology. These global objects are portable entities that are defined independent of each other and of other objects. The following sections describe how these global entities were used to model traffic in the project network model.

5.3.1 Application Configuration Panel

Figure 5.7 - Application Model Hierarchy

Figure 5.7 depicts the hierarchy that was adopted in constructing the application profiles that were defined in the project.

OPNET provides global objects for defining Applications. Using the Application Configuration Panel, the characteristics of the various applications in use by the end-users connected to the network were defined. The Application Configuration Panel is a portable entity and the application defined therein can be used anywhere within the context of the overall topology. To keep the simulation setup simple, while retaining a sense of realistic traffic and applications, three application instances were defined: 1. EMAIL 2. FTP and 3.VOICE. The details of these definitions are listed in the latter parts of this section.

The profiles were constructed using different application definitions. For each application defined, usage parameters such as start time, duration and repeatability were specified. For every instance, application specific parameters were set. These parameters will define the behavioural traits of each application instance. The parameters are relevant to the application type selected and this was put into overall context of the traffic characteristics needed for the simulations.

The application profiles used were amended using some of the number of common network applications, which are native to the software. Besides the three application types that were customized for this project, the others available in the Standard Network Applications suite include Rlogin, Video Conferencing, Database, HTTP, Print, etc.

A multi-step process was adopted to setup the nodes to use the applications for traffic generation. The process necessarily has to commence with setting up the applications in the Application configuration panel.

Statistical distributions were used to define the randomness of parameters such as Inter-request time, File size, Send Inter-arrival time and others. The choice of statistical distributions was comprehensive and the end selection was made while putting into perspective the requirement for the traffic characteristics needed for the simulations.

5.3.2 Profile Configuration Panel

Profiles were used to describe the activity patterns of the workstations on the network. Similar to the Application Configuration Panel, OPNET has global objects that can be used to define the behavioural profiles of end-users.

Profiles describe the activity patterns of a user or a group of users in terms of the applications used over a period of time. Several different profiles can be defined to run on a given LAN or workstation. Profiles can execute repeatedly on the same node. Profiles contain a list of applications. Within a profile, the applications can be configured to run at the same time, sequentially in a specified order or one after another randomly.

Using the Profile Configuration Panel, the parameters were set to define typical end user profiles. Typical profiles would include multi-tasking Application instances (defined in the Application Configuration Panel). The duration of the profile can vary from instance to instance – including the start and end times as well as the repeatability of any particular instance. For a particular profile, applications can be multi-tasked and run simultaneously with different start and end times. Profiles can be defined to differentiate the behavioural patterns of an engineer as opposed to a manager and so on.

The Profile Definition Object was used to define all the profiles in the project. Only profiles that have been defined in the Profile Definition Object can be applied to

workstations or the LANs defined on the project and only applications that have been defined in the Application Definition Object can be used in profile definitions.

5.3.3 Scenarios, Profiles and Applications

In order to categorically differentiate the performance of the conventional Call Admission Control schemes compared to the MRAN and GAP based CAC schemes, the approach taken was to simulate the performance of the CAC methods under varied circumstances. The scenarios were constituted a matrix of varied incident traffic against different traffic Quality of Service request.

To elucidate further, three traffic scenarios were simulated. Each of the traffic scenarios above were simulated under circumstances where the workstations request for three different levels of Quality of Service – High, Medium and Low as listed in Table 5.1

Attribute	QoS = HIGH	QoS = OPT	QoS =
PCR (Mbps)	0.5	0.1	0.05
MCR (Mbps)	0.5	0.1	0.05
SCR (Mbps)	0.5	0.1	0.05
MBS	50	10	5
Requested QoS	CBR	CBR	CBR
PpCDV (milliseconds)	10	10	10
Max CTD	200	200	200
Cell Loss Ratio	3 e-7	3 e-7	3 e-7

Table 5.1 – Summary Parameter setup for Quality of Service levels

Traffic Scenario 1 – The incident traffic consisted of profiles wherein the application under use was FTP only. The traffic scenario was simulated under three conditions ranging from low to medium to high QoS traffic contracts. Table 5.1 depicts at a summary level the parameters used to represent Traffic Scenario 1. For an expanded list

of parameters used to depict traffic scenario 1 (including parameters to represent Low and Optimal QoS conditions), please refer to Appendix B on Parameter Setup.

Profile – Profile 1	
Attribute	Value
Profile Name	Profile 1
Application	FTP
Start Time (seconds)	Uniform (100, 110)
Inter-repetition time (seconds)	Constant (300)
Application - FTP	
Attribute	Value
Command Mix (Get/Total)	50%
Inter-request time (seconds)	Exponential (100)
File size (bytes)	Triangular (100, 200)
QoS - HIGH	
Attribute	Value
PCR (Mbps)	0.5
MCR (Mbps)	0.5
SCR (Mbps)	0.5
MBS	50
Requested QoS	CBR
PpCDV millisecs)	10
Max CTD	200
Cell Loss Ratio	3 e-7

Table 5.2 – Summary Parameter setup for Traffic Scenario 1

Traffic Scenario 2 – The incident traffic consisted of profiles wherein the applications under use were FTP and Email only. The traffic scenario was simulated under three conditions ranging from low to medium to high QoS traffic contracts. Table 5.2 depicts at a summary level the parameters used to represent Traffic Scenario 2. For an expanded list of parameters used to depict traffic scenario 2 (including parameters to represent Low and Optimal QoS conditions), please refer to Appendix B on Parameter Setup. Traffic Scenario 2 uses two Profiles – Profile 1 as defined above and Profile 2 as summarized below.

Profiles – Profile 1 (as in Table 5.1) and Profile 2	
Attribute	Value
Profile Name	Profile 2
Application	EMAIL
Start Time (seconds)	Uniform (5, 10)
Inter-repetition time (seconds)	Exponential (300)
Application – FTP (as in Table 5.1) and EMAIL	
Attribute	Value
Send Inter-arrival Time	Exponential (180)
Send Group Size	Constant (3)
Receive Inter-arrival Time	Exponential (60)
Email size (bytes)	1000
QoS – HIGH	
Attribute	Value
PCR (Mbps)	0.5
MCR (Mbps)	0.5
SCR (Mbps)	0.5
MBS	50
Requested QoS	CBR
PpCDV millisecs)	10
Max CTD	200
Cell Loss Ratio	3 e-7

Table 5.3 – Summary Parameter Setup for Traffic Scenario 2

Traffic Scenario 3 – The incident traffic consisted of profiles wherein the applications under use were FTP, Email and Voice only. The traffic scenario was simulated under three conditions ranging from low to medium to high QoS traffic contracts. Table 5.3 depicts at a summary level the parameters used to represent Traffic Scenario 3. For an expanded list of parameters used to depict traffic scenario 3 (including parameters to represent Low and Optimal QoS conditions), please refer to Appendix B on Parameter Setup. Traffic Scenario 3 uses three Profiles – Profile 1 as defined in Table 5.1, Profile 2 as defined in Table 5.2 and Profile 3 as summarized below.

Profiles – Profile 1 (as in Table 5.1), Profile 2 (as in Table 5.2) and Profile 3	
Attribute	Value
Profile Name	Profile 3
Application	VOICE
Start Time (seconds)	Uniform (5, 10)
Inter-repetition time (seconds)	Exponential (300)
Application – FTP (as in Table 5.1), EMAIL (as in Table 5.2)	
and VOICE (PCM Quality Speech)	
QoS – HIGH	
Attribute	Value
PCR (Mbps)	0.5
MCR (Mbps)	0.5
SCR (Mbps)	0.5
MBS	50
Requested QoS	CBR
PpCDV millisecs)	10
Max CTD	200
Cell Loss Ratio	3 e-7

Table 5.4 – Summary Parameter Setup for Traffic Scenario 3

5.3.4 ATM Call Admission Control in OPNET

ATM Management module – Builds and updates ATM routing table. Provides route recommendations to other ATM processes.

ATM Translation module – Receives incoming ATM cells from the network. Translates VPI/VCI values for outgoing cells. Forwards cells to appropriate ATM module.

The Call Admission Control calculating routine in OPNET was amended by the incorporation of MRAN and GAP. The CAC routine in OPNET is located in an external file named ams_atm_support_v3.ex.c. ATM traffic control functions are explicitly modelled in OPNET. Call Admission Control is based on the user specified traffic contract.

The Call Admission Control algorithm in OPNET stipulates that the sum of the PCRs is:

Less than or equal to the maximum allocated bandwidth for the CBR category

The sum of SCRs for the VBR category

The sum of MCRs for the ABR category

These definitions are in the external file named above.

MRAN and GAP were introduced as schemes to perform dynamic allocation of bandwidth, which was aimed at optimizing the CAC decisions.

The process model named ATM Translation module receives incoming ATM cells from the network and translates VPI/VCI values for outgoing cells. It forwards cells to appropriate ATM module. The ATM Translation process model was modified so as to enable it to also calculate the data rate / throughput for the links that it forwards the incoming cells to. The rate of data transmission was calculated and stored to a local file dynamically during the course of the simulation. The data rate values were used as staggered inputs to the neural network schemes MRAN and GAP.

By using the values of data rate as input, the neural network schemes predict the cell rate on the various VCIs in the network. This predicted cell rate is used for every Call Admission Control decision. It removes the need for the network to make static admission control decisions based on pre-determined values of PCR, SCR and MBS.

As explained in Chapter 4, the MRAN and GAP algorithms do not need offline data for training purposes. The MRAN and GAP learning algorithms are online in nature wherein they allocate a new hidden neuron based on observations made and when the need arises, shed those neurons that have insignificant contributions to the outputs of

the network. Thus, when introduced in the critical path of traffic data rate prediction in the dynamic time varying ATM network, the architecture of the neural networks will adjust automatically on line to adapt closely to the dynamics of the observation data. MRAN and GAP are used in ATM networks for prediction of the incident traffic rate with the aim of subsequently using the predicted values for more effective call admission control. Since the neural networks are used for prediction, no classification or class labels are required.

The two neural network schemes MRAN and GAP were defined as separate stand-alone functions in the external file `ams_atm_support_v3.ex.c`. The admission control algorithms are defined in function `ams_atm_vp_supports_traffic()`. The function calls for `mran()` and `gap()` were defined in the function `ams_atm_vp_supports_traffic()`. The data rates tapped and written to local files were used as inputs to the neural network schemes. The output from the neural network functions was written to a variable that would replace the static traffic contract indicators in the existing admission control algorithms. The use of the `mran()` and `gap()` functions was mutually exclusive. The program was devised in such a manner so as to enable the switching of algorithms from conventional to `mran()` to `gap()`. Once setup, without any change to the programming logic, simulations can be run for the conventional scheme followed by either of the neural network schemes.

The following statistics were collected for every simulation run for each of the three modes – conventional, MRAN and GAP.:

Call Blocking Ratio

ATM Cell Delay

ATM Cell Delay Variation

Throughput/Utilization

In summary, this chapter was aimed at providing insights into the process modelling that was done to amend the admission control schemes in OPNET in order to incorporate the neural networks MRAN and GAP. The capabilities of OPNET modeller software were summarized. Insights were provided on the utility of the modelling hierarchical structure and how it was used and amended in order to achieve the goals of this project. The network, node and process model setup specific to the project were explained. Critical parameter setup activities including the Application and Profile panel configuration were discussed. Lastly, the inclusion of the neural network schemes in the critical path of the admission control algorithm was elucidated. The following chapter will provide a summary of the results collected and compare the differences between the statistics collected across the conventional and neural network influenced admission control simulations.

Chapter 6

Quantitative Performance Evaluation of Neural CAC Schemes

The previous chapter described the hierarchical model structure in OPNET and how it was used to model the simulation setup for the project at hand. Specifically, insights were provided into the methodologies used to create and amend the network, node and process models. The parameter setups constitute an important segment of the project and were explained at length. These setups go a long way towards influencing, understanding and analysing the results. This chapter will focus on providing a layout of the results corresponding to the traffic scenarios driven by the parameter setups detailed previously. The results will be summarily explained and conclusions corresponding to the individual result sets will be drawn.

The statistics collected for illustrating the variances in performance between the neural and conventional schemes are as follows:

Call Blocking Ratio – Ratio of number of calls rejected to the total number of calls attempted. Calls could get rejected due to several reasons. Details about the reason for rejection and the node that rejected the call can be found in the simulation log.

Cell Delay – Represents the end-to-end delay of ATM cells received by the ATM layer in this node. Measured from the time an ATM cell is sent from the source ATM layer to the time it is received by the ATM layer in the destination node.

Cell Delay variation – represents the average variances in the cell delays.

Point-to-Point Throughput – Represents the total number of bits/packets sent on a link between two nodes in the network.

6.1 Traffic Scenario 1 – Application: FTP

The incident traffic consisted of profiles wherein the application under use was FTP only. The traffic scenario was simulated under three conditions ranging from low to medium to high QoS traffic contracts (as explained in Chapter 5 Section 5.3.3). The incident traffic was modelled to begin at 100 seconds and end at 300 seconds.

Figure 6.1.1 – Call Blocking Ratio – Medium QoS

Observation: Figure 6.1.1 plots Call Blocking Ratio as a function of time for Traffic Scenario 1 and a medium QoS traffic contract. In summary, it can be inferred that the neural CAC schemes – namely GAP and MRAN respectively – provide a lower Call Blocking Ratio compared to the conventional scheme. The number of calls blocked under the conventional CAC scheme hovers between 80-90 %. The Call Blocking Ratio for the network under the influence of the MRAN neural network CAC scheme is considerably less while the CAC performance under the influence of GAP is even better. While the Call Blocking Ratio under GAP CAC is considerably better during the course of the simulation, it was observed that at the initiation of traffic incidence, momentarily the Call Blocking Ratio peaks at a 100%. This is attributed to the time taken for the neural network to learn the pattern of input traffic.

Figure 6.1.2 – Call Blocking Ratio – High QoS

Observation: Figure 6.1.2 plots Call Blocking Ratio as a function of time for Traffic Scenario 1 and a high QoS traffic contract. The Call Blocking Ratio under the conventional scheme is very high compared to the neural CAC schemes – namely MRAN and GAP respectively. Under Traffic Scenario 1 and a high QoS traffic

contract, the conventional CAC schemes are observed to be ineffective with a Call Blocking Ratio of a 100%. The call blocking ratio patterns for the neural network scheme remain largely unchanged from a medium QoS traffic contract scenario with the performance under the effect of GAP superseding the MRAN driven CAC schemes. This is true because the incoming traffic patterns remain unchanged. The neural network schemes make CAC decisions based on predicting network behaviour based on existing patterns. Conventional schemes are based on QoS traffic contracts – since the QoS demands are higher in the second simulation, the Call Blocking Ratio under the conventional scheme is higher.

Call Blocking Ratio is the most relevant statistic to understand Call Admission Control. In summary, Figures 6.1.1 and 6.1.2 highlight comprehensively that conventional CAC schemes under medium and high QoS traffic contracts do not perform as well as CAC schemes driven by neural network engines. In addition, CAC decisions under the influence of GAP were found to be better than those made under the influence of MRAN.

Figure 6.1.3 – ATM Cell Delay – Medium QoS

Observation: Figure 6.1.3 plots ATM Cell Delay as a function of time for Traffic Scenario 1 and a medium QoS traffic contract. For Traffic Scenario 1 and an medium QoS traffic contract, the end to end delay taken for an ATM cell to traverse the path from the source node to the destination node was found to average at a value marginally less than 0.005 seconds under the Conventional CAC scheme. The cell delay in the network under the MRAN-CAC scheme was observed to fluctuate between 0.004 and 0.005 seconds – for the most part less than the delay experienced under the conventional scheme. The cell delay in the network under the GAP-CAC scheme was observed to fluctuate between 0.004 and 0.006 seconds before settling at 0.005 seconds. The initial variation is attributed to the overhead introduced for the neural network to learn the networks behavioural pattern thus contributing to a delay in the transmission of cells.

Figure 6.1.4 – ATM Cell Delay – High QoS

Observation: Figure 6.1.4 plots ATM Cell Delay as a function of time for Traffic Scenario 1 and a High QoS traffic contract. It is observed that under these conditions,

the cell delay experienced by the conventional scheme is more than that experienced by GAP and MRAN respectively. For Traffic Scenario 1 and a high QoS traffic contract, the end to end delay taken for an ATM cell to traverse the path from the source node to the destination node was found to average at a value close to 0.006 seconds increasing marginally towards the end of the simulation. The cell delays experienced under the MRAN and GAP driven CAC schemes were observed to mirror the medium QoS simulation conditions. Unlike the previous simulation, the cell delay experienced in this simulation under conventional schemes was unequivocally higher than the performance under the neural network schemes.

Figure 6.1.5 – ATM Cell Delay Variation – Medium QoS

Figure 6.1.6 – ATM Cell Delay Variation – High QoS

Observation: Figure 6.1.5 plots ATM Cell Delay Variation as a function of time for Traffic Scenario 1 and an medium QoS traffic contract while Figure 6.1.6 plots ATM Cell Delay Variation as a function of time for Traffic Scenario 1 and a high QoS traffic contract. In both cases, it is observed that the variations experienced by the GAP scheme are greater than that experienced by the MRAN and conventional schemes respectively. The ATM Cell Delay variation indicates the variances experienced in average cell delay. As observed in Traffic Scenario 1 under both medium and high traffic conditions, the variances in the cell delay experienced during the conventional CAC schemes are negligible. The GAP-CAC cell delay fluctuated significantly, which is reflected in the swinging GAP-CAC Cell Delay variation. The Cell Delay Variation statistic should be viewed in conjunction with the Cell Delay statistic.

Figure 6.1.7 – Traffic Throughput on the backbone link – Medium QoS

Observation: Figure 6.1.7 plots Traffic Throughput on the backbone link as a function of time for Traffic Scenario 1 and a medium QoS traffic contract. The throughput detected under the conventional scheme is much less to that obtained under the GAP and MRAN schemes and is close to the zero line on the plot. Link 1 is the backbone link between the two central switches in the network model depicted in Figure 5.6. The throughput on the central link under the conventional schemes is observed to be at an average of 85 KBps. The throughput on the central link under the MRAN-CAC schemes is observed to peak at 18000 KBps while gradually decreasing throughout the rest of the simulation. The throughput on the central link under the GAP-CAC scheme was found to hover between the 11000-18000 KBps range.

Figure 6.1.8 – Traffic Throughput on the backbone link – High QoS

Observation: Figure 6.1.8 plots Traffic Throughput on the backbone link as a function of time for Traffic Scenario 1 and high QoS traffic contract. The throughput detected under the conventional scheme is much less to that obtained under the GAP and MRAN schemes and is close to the zero line on the plot. Link 1 is the backbone link between the two central switches in the network model depicted in Figure 5.6. The throughput results under Medium and High QoS traffic contracts were found to be similar.

The total throughput for the duration of the simulation is obtained by calculating the area under the graph. The throughput statistic mirrors the extent of call blocking executed by the three schemes. Since the ratio of calls blocked under the conventional scheme is high, the throughput on the central link is correspondingly very low. Similarly, as the call blocking done under the neural network driven CAC schemes is lower compared to the conventional schemes, the throughput is correspondingly high. Notably, since the incoming traffic remains unchanged in the face of changing QoS traffic contracts, the throughput under the neural network schemes remains unchanged.

6.2 Traffic Scenario 2 – Applications: FTP and Email

The incident traffic consisted of profiles wherein the applications under use were FTP and Email only. The traffic scenario was simulated under three conditions ranging from low to medium to high QoS traffic contracts (as explained in Chapter 5 Section 5.3.3). The incident traffic was modelled to begin at 100 seconds and end at 300 seconds.

Figure 6.2.1 – ATM Call Blocking Ratio – Medium QoS

Observation: Figure 6.2.1 plots Call Blocking Ratio as a function of time for Traffic Scenario 2 and a medium QoS traffic contract. In summary, it can be inferred that the neural CAC schemes – namely GAP and MRAN respectively – provide a lower Call Blocking Ratio compared to the conventional scheme. For Traffic Scenario 2 and a medium QoS traffic contract, the number of calls blocked under the conventional CAC scheme hovers between 80-100 %. The Call Blocking Ratio for the network under the influence of the MRAN neural network CAC scheme is considerably less while the

CAC performance under the influence of GAP is even better. While the Call Blocking Ratio under GAP CAC is considerably better during the course of the simulation, it was observed that at the initiation of traffic incidence, the Call Blocking Ratio peaks at 80%. This is attributed to the time taken for the neural network to learn the pattern of input traffic.

Figure 6.2.2 – ATM Call Blocking Ratio – High QoS

Observation: Figure 6.2.2 plots Call Blocking Ratio as a function of time for Traffic Scenario 2 and a high QoS traffic contract. The conventional Call Blocking Ratio is observed to be very high (=1). In contrast, the Call Blocking Ratios under MRAN and GAP were lower respectively and found to follow a decreasing trend. In summary, it can be inferred that the neural CAC schemes – namely GAP and MRAN respectively – provide a lower Call Blocking Ratio compared to the conventional scheme. Under Traffic Scenario 2 and a high QoS traffic contract, the conventional CAC schemes are observed to be ineffective with a Call Blocking Ratio of a 100%. The call blocking ratio patterns for the neural network scheme remain largely unchanged from a medium QoS traffic contract scenario with the performance under the effect of GAP superseding

the MRAN driven CAC schemes. This is true because the incoming traffic patterns remain unchanged. The neural network schemes make CAC decisions based on predicting network behaviour based on existing patterns. Conventional schemes are based on QoS traffic contracts – since the QoS demands are higher in the second simulation, the Call Blocking Ratio under the conventional scheme is higher.

Call Blocking Ratio is the most relevant statistic to understand Call Admission Control. In summary, Figures 6.1.1 and 6.1.2 highlight comprehensively that conventional CAC schemes under medium and high QoS traffic contracts do not perform as well as CAC schemes driven by neural network engines. In addition, CAC decisions under the influence of GAP were found to be better than those made under the influence of MRAN.

The neural network driven CAC schemes are found to perform better than the conventional schemes under both traffic scenarios. Traffic scenario 2 inflicts more traffic on the network compared to traffic scenario 1. The neural network schemes handle increased traffic load with marginal reduction in performance. On the other hand, changes in QoS traffic contracts significantly change the call admission decisions made by the conventional schemes, even though the incident traffic remains unchanged.

Figure 6.2.4 – ATM Cell Delay – High QoS

Figure 6.2.4 plots ATM Cell Delay as a function of time for Traffic Scenario 2 and a high QoS traffic contract. The cell delays detected for the GAP and MRAN schemes were found to be higher than those detected under the conventional schemes.

The ATM cell delay statistic reflects the average time for cells to reach from source to destination. It reflects the networks behaviour under heavy traffic conditions. It reflects the networks response to varying traffic conditions. For the networks driven by neural network CAC schemes, the cell delay statistic can be extrapolated to explain the time taken for the neural network to adjust to varying traffic conditions and to learn the networks behavioural patterns. The simulation results above indicate that MRAN and specifically, GAP, introduce an overhead in adapting to varying traffic conditions resulting in a higher cell delay. Within a traffic scenario, the cell delay statistics (and correspondingly, cell delay variations) will not differ greatly, since they are not affected by changing QoS traffic contracts. However, when compared across traffic scenarios, the differences explain the different approaches adopted across the conventional and neural network schemes. The neural network schemes predict traffic patterns based on

Figure 6.2.3 – ATM Cell Delay – Medium QoS

Observation: Figure 6.2.3 plots ATM Cell Delay as a function of time for Traffic Scenario 2 and a medium QoS traffic contract. The cell delays detected for the GAP and MRAN schemes were found to be higher than those detected under the conventional schemes. For Traffic Scenario 2 and an medium QoS traffic contract, the end to end delay taken for an ATM cell to traverse the path from the source node to the destination node was found to average at a value marginally less than 0.004 seconds under the Conventional CAC scheme. The cell delay in the network under the MRAN-CAC scheme was observed to fluctuate between 0.005 and 0.008 seconds. The cell delay in the network under the GAP-CAC scheme was observed to fluctuate between 0.007 and 0.011 seconds before settling around 0.008 seconds. The initial variation is attributed to the overhead introduced for the neural network to learn the networks behavioural pattern thus contributing to a delay in the transmission of cells.

existing incident traffic behaviour. Therefore, in Traffic Scenario 2 where the incident traffic load is heavier when compared to Traffic Scenario 1, the cell delays experienced under the neural network schemes are higher – indicating poorer network performance. The cell delays experienced under the conventional CAC schemes are by and large not affected. This is because the conventional schemes make their call admission decisions based on the traffic contract parameters. Once the call admission decision is made, the network is essentially buffered against traffic congestion, thus providing a stable cell delay statistic.

Figure 6.2.5 – ATM Cell Delay Variation – Medium QoS

Figure 6.2.6 – ATM Cell Delay Variation – High QoS

Observation: Figure 6.2.5 plots ATM Cell Delay Variation as a function of time for Traffic Scenario 2 and an medium QoS traffic contract while Figure 6.2.6 plots ATM Cell Delay Variation as a function of time for Traffic Scenario 2 and a high QoS traffic contract. In both cases, it is observed that the variations experienced by the GAP scheme are greater than that experienced by the MRAN and conventional schemes respectively. The ATM Cell Delay variation indicates the variances experienced in average cell delay. As observed in Traffic Scenario 1 under both medium and high traffic conditions, the variances in the cell delay experienced during the conventional CAC schemes are negligible. The GAP-CAC cell delay fluctuated significantly, which is reflected in the swinging GAP-CAC Cell Delay variation. The Cell Delay Variation statistic should be viewed in conjunction with the Cell Delay statistic.

Figure 6.2.7 – ATM Traffic Throughput on Link 1 – Medium QoS

Observation: Figure 6.2.7 plots Traffic Throughput on the backbone link as a function of time for Traffic Scenario 2 and a medium QoS traffic contract. The throughput detected under the conventional scheme is much less to that obtained under the GAP and MRAN schemes and is close to the zero line on the plot. The throughput on the backbone link under the GAP scheme begins to increase with a time lag when compared to the throughput under the MRAN scheme. Link 1 is the backbone link between the two central switches in the network model depicted in Figure 5.6. The throughput on the central link under the conventional schemes is observed to be close 25 Kbps. The throughput on the central link under the MRAN-CAC and GAP-CAC schemes is observed to peak at 80000 Kbps.

Figure 6.2.8 – ATM Traffic Throughput on Link 1 – High QoS

Observation: Figure 6.2.8 plots Traffic Throughput on the backbone link as a function of time for Traffic Scenario 2 and a high QoS traffic contract. The throughput detected under the conventional scheme is much less to that obtained under the GAP and MRAN schemes and is close to the zero line on the plot. . The throughput on the backbone link under the GAP scheme begins to increase with a time lag when compared to the throughput under the MRAN scheme. Link 1 is the backbone link between the two central switches in the network model depicted in Figure 5.6. The throughput results under Medium and High QoS traffic contracts were found to be similar.

The total throughput for the duration of the simulation is obtained by calculating the area under the graph. The throughput statistic mirrors the extent of call blocking executed by the three schemes. Since the ratio of calls blocked under the conventional scheme is high, the throughput on the central link is correspondingly very low. Similarly, as the call blocking done under the neural network driven CAC schemes is

lower compared to the conventional schemes, the throughput is correspondingly high. Since the incoming traffic remains unchanged in the face of changing QoS traffic contracts, the throughput under the neural network schemes remains unchanged. The time lag between the GAP and MRAN throughput is attributed to the overhead introduced by greater computation needed in the GAP neural network. When compared across the traffic scenarios, the throughput is notably higher since the incoming traffic is much higher in Traffic Scenario 2 when compared to Traffic Scenario 1.

6.3 Traffic Scenario 3 – Applications: FTP, Email and Voice

The incident traffic consisted of profiles wherein the applications under use were FTP Email and Voice. The traffic scenario was simulated under three conditions ranging from low to medium to high QoS traffic contracts (as explained in Chapter 5 Section 5.3.3). The general observations made in Traffic Scenario 3 compare well when extrapolated from Traffic scenarios 1 and 2 and will not be dealt with in detail. In line with this, the results collected for Traffic Scenario 3 under Heavy QoS traffic conditions have not been presented. The incident traffic was modelled to begin at 100 seconds and end at 300 seconds.

Figure 6.3.1 – ATM Call Blocking Ratio – Medium QoS

Observation: Figure 6.3.1 plots Call Blocking Ratio as a function of time for Traffic Scenario 1 and a medium QoS traffic contract. In summary, it can be inferred that the neural CAC schemes – namely GAP and MRAN respectively – provide a marginally lower Call Blocking Ratio compared to the conventional scheme. For Traffic Scenario 3 and a medium QoS traffic contract, the Conventional CAC Call Blocking Ratio was found to be higher than the Call Blocking Ratio experienced under the MRAN and GAP-CAC schemes respectively.

Figure 6.3.2 – ATM Cell Delay – Medium QoS

Observation: Figure 6.3.2 plots ATM Cell Delay as a function of time for Traffic Scenario 3 and a medium QoS traffic contract. The cell delays detected for the GAP and MRAN schemes were found to be much higher than those detected under the conventional schemes. Given the plot range, the conventional cell delay graph is close to the zero line. For Traffic Scenario 3 and a medium QoS traffic contract, the cell delays experienced by the network under the three schemes were similar. The MRAN-CAC cell delay tended to increase towards the end of the simulation. The high cell delays experienced by the neural network CAC schemes are attributed to increased traffic load.

Figure 6.3.4 – ATM Cell Delay Variation – Medium QoS

Observation: Figure 6.3.4 plots ATM Cell Delay Variation as a function of time for Traffic Scenario 3 and a medium QoS traffic contract. It is observed that the variations experienced by the GAP scheme are greater than that experienced by the MRAN and conventional schemes respectively. The Cell Delay Variation mirrors the behaviour observed for the Cell Delay statistic. The cell delay variations are minimal across the three schemes during the course of the simulation and this is reflected in the graph. As the simulation progresses, the cell delay experienced by the MRAN-CAC schemes tends to increase and this is reflected in the Cell Delay Variation graph.

6.4 Summary of Performance Results

Simulation results under a typical simulation run for the three traffic scenarios indicate that the GAP and MRAN controllers perform better than the conventional call admission controller in making more effective call blocking decisions while maintaining a high quality of service. The neural CAC schemes consistently provide a better Call Blocking Ratio when compared to the conventional schemes across the three traffic scenarios simulated. Similarly, the neural controllers were found to provide better throughput through the network for the three traffic and QoS scenarios simulated. The cell delays experienced by the network for traffic scenarios 1 and 3 were found to be approximately the same for the neural controllers and the conventional schemes. For traffic scenario 1, under a high QoS stipulation, the cell delay experienced by the neural controllers was actually found to be better than the conventional CAC schemes. The conventional scheme was observed to have less cell delay compared to the neural controllers for traffic scenario 2 under a high QoS stipulation. The cell delay variation experienced by the neural controllers was found to be greater than the conventional schemes for all scenarios simulated. For further details, please refer to Tables 6.4.1-6.4.5.

The main performance results of GAP and MRAN given in the previous sections are summarized in the tables below.

Traffic Scenario 1 was parameterized to depict incident varying multiplexed bursty, variable bit rate and multiplexed traffic profiles consisting of the FTP application only. The traffic scenario was simulated under three conditions ranging from low to medium to high QoS traffic contracts. The results collected for Traffic Scenario 1 are summarized in Table 6.4.1-2.

Traffic contract	Call Blocking Ratio			ATM Throughput (Kbps/sec)		
	GAP	MRAN	Conv	GAP	MRAN	Conv
Medium	5.01E-01	5.64E-01	8.04E-01	8.87E+03	8.05E+03	8.53E+01
High	5.01E-01	5.64E-01	1.00E+00	8.87E+03	8.05E+03	8.53E+01

Table 6.4.1 – Overall Performance Comparison of the Call Admission Controllers – Traffic Scenario 1

Traffic contract	ATM Cell Delay (secs)			ATM Cell Delay Variation (secs)		
	GAP	MRAN	Conv	GAP	MRAN	Conv
Medium	4.96E-03	4.26E-03	4.75E-03	8.50E-06	1.74E-06	1.00E-07
High	4.96E-03	4.26E-03	6.05E-03	8.50E-06	1.74E-06	1.00E-07

Table 6.4.2 – Overall Performance Comparison of the Call Admission Controllers – Traffic Scenario 1

Traffic Scenario 2 was parameterized to depict incident varying multiplexed bursty, variable bit rate and multiplexed traffic profiles consisting of FTP and Email applications. The traffic scenario was simulated under three conditions ranging from low to medium to high QoS traffic contracts. The results collected for Traffic Scenario 2 are summarized in Table 6.4.3-4.

Traffic contract	Call Blocking Ratio			ATM Throughput (Kbps/sec)		
	GAP	MRAN	Conv	GAP	MRAN	Conv
Medium	5.13E+01	4.85E+01	9.11E+01	9.96E+03	1.13E+04	2.26E+01
High	5.13E+01	5.90E+01	1.00E+02	9.96E+03	1.13E+04	2.26E+01

Table 6.4.3 – Overall Performance Comparison of the Call Admission Controllers – Traffic Scenario 2

Traffic contract	ATM Cell Delay (secs)			ATM Cell Delay Variation (secs)		
	GAP	MRAN	Conv	GAP	MRAN	Conv
Medium	8.62E-03	7.40E-03	3.79E-03	1.17E-05	5.01E-06	6.29E-09
High	8.62E-03	7.40E-03	3.79E-03	1.17E-05	5.01E-06	6.29E-09

Table 6.4.4 – Overall Performance Comparison of the Call Admission Controllers – Traffic Scenario 2

Traffic Scenario 3 was parameterized to depict incident varying multiplexed bursty, variable bit rate and multiplexed traffic profiles consisting of FTP, Email and Voice applications. The traffic scenario was simulated under three conditions ranging from low to medium to high QoS traffic contracts. The results collected for Traffic Scenario 1 are summarized in Table 6.4.5.

Traffic contract	Call Blocking Ratio			ATM Cell Delay (secs)			ATM Throughput (Kbps/sec)		
	GAP	MRAN	Conv	GAP	MRAN	Conv	GAP	MRAN	Conv
Medium	5.13E+01	4.85E+01	5.85E+01	3.95E-03	4.26E-03	3.95E-03	9.96E+03	1.13E+04	2.26E+01

Table 6.4.5 – Overall Performance Comparison of the Call Admission Controllers – Traffic Scenario 3

The results can be summarized as follows:

- GAP and MRAN driven call admission control result in better Call Blocking Ratios when compared to the conventional schemes under heavy and regular incident traffic scenarios. For traffic scenario 1 (FTP only) under a high QoS traffic contract, GAP-CAC resulted in about 50% of the total number of calls being blocked, compared to approximately 56% for MRAN-CAC and 100% for Conventional-CAC. For traffic scenario 2 (FTP and Email only) under a high QoS traffic contract, GAP-CAC resulted in about 51% of the total number of

calls being blocked, compared to approximately 60% for MRAN-CAC and 100% for Conventional-CAC. For traffic scenario 3 (FTP, Email and Voice) under an medium QoS traffic contract, GAP-CAC resulted in about 51% of the total number of calls being blocked, compared to approximately 49% for MRAN-CAC and 59% for Conventional-CAC. A similar trend in the Call Blocking Ratio was observed for a Medium QoS traffic contract as illustrated in Tables 7.1 – 7.3.

- The throughput on a sample backbone link of the network when subject to the neural call admission controllers was found to be much higher than that of the network under the conventional CAC scheme. The neural network schemes result in higher traffic throughput owing to better call blocking ratios. For traffic scenario 1, the backbone link throughput under GAP and MRAN was found to be approximately 100 times greater than the conventional scheme. For traffic scenarios 2 and 3, the backbone link throughput under GAP and MRAN was found to be approximately 400 and 500 times greater than the backbone link throughput for the network subject to the conventional admission controller.
- The average ATM cell delays experienced by the network under GAP and MRAN for traffic scenarios 1 and 3 were found to be in the same range as that experienced by the conventional schemes. The average ATM cell delay experienced by the GAP and MRAN networks in traffic scenario 2, however, 2.27 and 1.95 times that experienced by the network when subject to conventional CAC. This is attributed to the delays introduced by neural computation and the traffic parameterization for scenario 2.

6.5 Cumulative Distribution Function Analysis of Call Blocking Ratio and ATM Cell Delay

This section provides the performance comparison based on a detailed statistical analysis through the Cumulative Distribution Function (CDF) of the Call Blocking Ratio and the ATM Cell Delay statistics. The previous sections detail the results collected for a typical simulation run across three scenarios, each evaluated for three QoS traffic contracts. The simulations under traffic scenario 1 and traffic scenario 2 as highlighted in Section 6.1-6.2 for a High QoS and Medium QoS traffic contract respectively were executed 20 times under varying traffic generation seeds. The traffic generation seeds can be amended in the Application and Profile configuration panels in OPNET. By varying the traffic generation parameters, the simulation setup is observed under varied random incident traffic conditions. The objective was to verify the robustness of the simulation conditions and the hypothesis. The Call Blocking Ratio and Cell Delay statistics were collected for the two traffic scenarios and used to plot the Cumulative Distribution Function.

The CDF of a statistic presents the likelihood of occurrence of the statistic's ordinate values. It shows the accumulated probability mass of all ordinates less than or equal to a particular ordinate. In the calculation of a CDF, proportions for each ordinate value in the original statistic are computed. The entries of the CDF are constructed by positioning the distinct ordinate values of the original statistic in increasing order on the abscissa, with one entry for each value. The CDF value for the initial entry is simply the probability mass of the corresponding ordinate. In summary, the CDF of a statistic STAT can be read to determine the probability p_x of that statistic being less than or equal to a given value α . In summary:

$$P(STAT \leq \alpha) = p_x \dots\dots\dots (1)$$

The performance comparison is done based on the Call Blocking Ratio statistic under Traffic Scenario 1 (High QoS traffic contract) and Traffic Scenario 2 (Medium QoS traffic contract). The CDF of the ATM Cell Delay was also plotted for the Traffic Scenario 1 (High QoS traffic contract) and Traffic Scenario 2 (Medium QoS traffic contract).

The Cumulative Distribution (CDF) of the Call Blocking Ratio for GAP-CAC, MRAN-CAC and Conventional CAC for Traffic Scenario 1 (High QoS) are shown in Figure 6.4.1. Following this, the CDF for Call Blocking Ratio for Traffic Scenario 2 (Medium QoS) is depicted in Figure 6.4.2. The statistic Call Blocking Ratio, which is the single most relevant statistics in the investigation of Call Admission Control, was calculated and tabulated for each of the 20 simulation runs for each of the traffic scenarios 1 and 2.

Figure 6.4.1 – Cumulative Distribution of Call Blocking Ratio – Traffic Scenario 1 (High QoS)

Figure 6.4.1 shows the Cumulative Distribution of the Call Blocking Ratio for Traffic Scenario 1 with a High QoS requirement. It is clear that the GAP controller gives a

minimum-blocking ratio compared to MRAN and the conventional scheme respectively. The plot provides the probability that the Call Blocking Ratio is less than or equal to a certain given value (based on equation 1) for each of the three schemes and vice versa. For instance, the probability of achieving a Call Blocking Ratio less than or equal to 0.5 is zero for the conventional scheme, 0.19 for MRAN and 0.65 for GAP. Likewise, the conventional scheme was read to be likely to return a Call Blocking Ratio of 0.97 or less with a probability of 0.6, whereas for the same probability, the call blocking ratio in the MRAN and GAP based networks were determined to be equal to or less than 0.55 and 0.48 for the same probability. Figure 6.4.1 highlights that GAP is less likely to block calls compared to the MRAN and conventional controller respectively, for a traffic scenario 1 consisting of heavily multiplexed bursty FTP traffic while stipulating a high QoS requirement.

Figure 6.4.2 – Cumulative Distribution of Call Blocking Ratio – Traffic Scenario 2 – (Medium QoS)

Similarly, the Figure 6.4.2 shows the Cumulative Distribution of the Call Blocking Ratio for Traffic Scenario 2 with a Medium QoS requirement. It is clear that the GAP controller gives a minimum-blocking ratio compared to MRAN and the conventional scheme respectively. The plot provides the probability that the Call Blocking Ratio is less than or equal to a certain given value (based on equation 1) for each of the three schemes and vice versa. For instance, the probability of achieving a Call Blocking Ratio less than or equal to 0.7 is zero for the conventional scheme, 0.7 for MRAN and 1 for GAP. Likewise, the conventional scheme was read to be likely to return a Call Blocking Ratio of 0.93 or less with a probability of 0.9, whereas for the same probability, the call blocking ratio in the MRAN and GAP based networks was determined to be equal to or less than 0.76 and 0.64 for the same probability. Figure 6.4.2 highlights that GAP is less likely to block calls compared to the MRAN and conventional controller respectively, for a traffic scenario 2 consisting of heavily multiplexed bursty FTP traffic while stipulating an medium QoS requirement.

In addition to the Call Blocking Ratio, the CDF for ATM Cell Delay was also tabulated and plotted. For each of the 20 simulation cycles for Traffic Scenario 1 (High QoS) and Traffic Scenario 2 (Medium QoS), the average ATM Cell Delay was calculated and tabulated. The Cumulative Distribution (CDF) of the ATM Cell Delay GAP-CAC, MRAN-CAC and Conventional CAC for Traffic Scenario 1 (High QoS) is shown in Figure 6.4.3. The CDF for ATM Cell Delay under Traffic Scenario 2 (Medium QoS) is depicted in Figure 6.4.4.

Figure 6.4.3 - Cumulative Distribution of ATM Cell Delay – Traffic Scenario 1 - (High QoS)

Figure 6.4.3 shows the Cumulative Distribution of the ATM Cell Delay for Traffic Scenario 1 (High QoS). The figure highlights that for Traffic Scenario 1 with a High QoS traffic contract, the likelihood of getting a higher cell delay is greater when using the conventional controllers as opposed to the neural schemes. The plot provides the probability of getting a given value or less of the ATM Cell Delay (based on Equation 1) for each of the three neural controllers and vice versa. For instance, the probability of getting a Cell Delay of 0.006 seconds or less is 0.1 for the Conventional scheme, 0.65 for GAP and 1 for MRAN. Thus, the neural schemes are more likely to provide a lower cell delay than the conventional scheme for Traffic Scenario 1. Likewise, the conventional scheme will return an ATM Cell Delay of 0.0065 seconds or less with a probability of 0.5, while MRAN and GAP will return cell delays of at most 0.00520

seconds and 0.00575 seconds for the same probability. Figure 6.4.3 highlights that the conventional scheme is more likely to result in higher cell delays compared to the neural controllers respectively, for Traffic scenario 1 consisting of heavily multiplexed bursty FTP and Email traffic while stipulating a High QoS requirement.

Figure 6.4.4 - Cumulative Distribution of ATM Cell Delay Traffic Scenario 2 - (Medium QoS)

Figure 6.4.4 shows the Cumulative Distribution of the ATM Cell Delay for Traffic Scenario 2 (Medium QoS). In contrast to Figure 6.4.3, this figure highlights that for Traffic Scenario 2 with a medium QoS traffic contract, the likelihood of getting a higher cell delay is greater when using the neural controllers as compared to the conventional scheme. It is clear that the GAP controller results in higher cell delays

when compared to MRAN and the conventional scheme respectively. The plot provides the probability of getting a given value of ATM Cell Delay or less (based on equation 1) for each of the three neural controllers and vice versa. For instance, the probability of getting a Cell Delay of 0.007 seconds or less is zero for GAP, 0.4 for MRAN and 1 for GAP. Thus, under the conventional scheme the ATM Cell Delay in the network will always be less than 0.007 seconds while the cell delay in the GAP controlled network will never be less than the same value. Likewise, the conventional scheme will return at most an ATM Cell Delay of 0.0047 with a probability of 0.6 while MRAN and GAP will return at most cell delays of 0.0075 seconds and 0.0093 seconds for the same probability. Figure 6.4.4 highlights that GAP is more likely to result in higher cell delays compared to the MRAN and conventional controller respectively, for Traffic Scenario 2 consisting of heavily multiplexed bursty FTP and Email traffic while stipulating a medium QoS requirement.

In summary, the purpose of this chapter was to list the primary results collected during the project and focus on the specific observations made with probably corresponding explanations. Insights were provided into the main statistics collected and the variations in behaviour across the traffic scenarios simulated and the different QoS traffic contracts established. For each essential statistic collected, probably explanations were provided for the behaviour observed and possible conclusions drawn. The CDF for the Call Blocking Ratio and ATM Cell Delay statistics was also presented for Traffic Scenario 1 (High QoS) and Traffic Scenario 2 (Medium QoS) and analyzed. The following chapter will focus on overall conclusions drawn based on results collected and in light of the overall premises of the project. Possible areas for future work will be explored.

Chapter 7

Conclusions and Recommendations

7.1 Conclusions

In this thesis, the performance of recently developed radial basis function (RBF) networks GAP (Growing and Pruning) network and MRAN (Minimal Resource Allocation Network) for Call Admission control of ATM networks has been evaluated. The ability of GAP and MRAN to build networks based on the novelties of the input data and then accurately predict the traffic were seen to be very useful in making better call admission decisions when compared to conventional admission control schemes. GAP and MRAN use online learning and react quickly to dynamically changing ATM traffic flows. The radial basis function neural networks were used to devise an adaptive call admission control scheme. The performance of such schemes was studied using OPNET's simulation facilities under varying traffic scenarios.

The performance of the CAC schemes were observed under three main scenarios depicting heavy traffic situations of varying multiplexed traffic consisting of bursty, variable bit rate and multiplexed traffic (combining both speech and video signals) in the form of applications such as FTP, Email and Voice. For each scenario, the metrics were collected for three different traffic contracts of varying Quality of Service – labelled High, Medium and Low respectively.

The main conclusions can be drawn as follows:

- Three scenarios were simulated, each consisting of three levels of QoS requirements. For each scenario, the four performance metrics Call Blocking Ratio, ATM Cell Delay/Variation and ATM Throughput were collected. Results from a typical simulation run across all traffic scenarios and traffic contract requirements indicate that the neural CAC schemes consistently provide a better Call Blocking Ratio when compared to the conventional schemes. Similarly, the neural controllers were found to provide better throughput through the network for the three traffic and QoS scenarios simulated. The cell delays experienced by the network for traffic scenarios 1 and 3 were found to be approximately the same for the neural controllers and the conventional schemes. For traffic scenario 1, under a high QoS stipulation, the cell delay experienced by the neural controllers was actually found to be better than the conventional CAC schemes. The conventional scheme was observed to have less cell delay compared to the neural controllers for traffic scenario 2 under a high QoS stipulation. The cell delay variation experienced by the neural controllers was found to be greater than the conventional schemes for all scenarios simulated.
- The Cumulative Distribution Function was plotted for Call Blocking Ratio and ATM Cell Delay for Traffic Scenario 1 (High QoS) and Traffic Scenario 2 (Medium QoS). The Cumulative Distribution Function of the Call Blocking Ratio for both Traffic Scenario 1 and 2 indicate that the performance of the GAP based Call Admission Controller is better than the MRAN and conventional controller respectively in terms of making better informed and more efficient call admission decisions. The likelihood of achieving a low Call Blocking Ratio for the GAP controller is higher when compared to the MRAN and conventional controller. For instance, the probability of achieving a Call Blocking Ratio less than or equal to 0.5 is zero for the conventional scheme, 0.19 for MRAN and 0.65 for GAP. Likewise, the conventional scheme was read to be likely to return a Call Blocking Ratio of 0.97 or less with a probability of 0.6, whereas for the same probability, the call blocking ratio in the MRAN and GAP based networks were determined to be equal to or less than 0.55 and 0.48

for the same probability. This proves that the neural controllers make more effective call admission control decisions while maintaining the prescribed Quality of Service

- The Cumulative Distribution Function plots of the ATM Cell Delay highlight that for Traffic Scenario 1, the neural CAC based networks experience less cell delay when compared to the conventional schemes. In contrast, for Traffic Scenario 2 under a medium traffic contract stipulation, the network experiences higher cell delays when subject to the neural controllers as compared to the conventional schemes. The increased delay in the neural based CAC schemes for traffic scenario 2 is attributed to the time lost in additional computation needed by the neural schemes. For instance, with reference to Traffic Scenario 2, the probability of getting a Cell Delay of 0.007 seconds or less is zero for GAP, 0.4 for MRAN and 1 for GAP. Thus, under the conventional scheme the ATM Cell Delay in the network will always be less than 0.007 seconds while the cell delay in the GAP controlled network will never be less than the same value. Likewise, the conventional scheme will return at most an ATM Cell Delay of 0.0047 with a probability of 0.6 while MRAN and GAP will return at most cell delays of 0.0075 seconds and 0.0093 seconds for the same probability. The CDF for the Cell Delay statistic for Traffic Scenario 1 under a High QoS stipulation reads in favour of the neural controllers. This indicates that while the neural controllers make better call admission control decisions conclusively, they might cause greater cell delays when compared to the conventional schemes.
- The results from the simulation done across the three traffic scenarios for the three levels of QoS indicate that in all cases, the throughput realized on the central backbone link of the network for the neural controllers was higher than the conventional controller. This proves that the neural controllers are able to provide a high utilization of the network under heavily multiplexed bursty application traffic.
- The results from the simulation done across the three traffic scenarios for the three levels of QoS indicate that in all cases, the cell delay experienced by the

neural controllers was the same or higher than the conventional scheme. This is attributed to the overhead introduced by the additional computation that needs to be done by the neural controllers while making call admission decisions.

- The neural network admission control schemes are unaffected by changes in QoS traffic contracts. The admission control decisions are based on dynamic traffic prediction.
- Conventional schemes base their admission control decisions on QoS traffic contracts. As a result, regardless of the incident traffic, changes in the traffic contract parameters result in corresponding changes in admission control decisions.
- Under heavier incident traffic conditions, neural network driven admission control schemes result in higher cell delay. This is because of the overhead introduced in increased computation for traffic prediction.
- Based on the results obtained, the performance of GAP is found to be superior to MRAN that was in turn better compared to the results obtained from networks under the influence of conventional schemes. Detailed theoretical analyses are required to provide further insights and understand the underlying factors behind the superior performance of GAP and MRAN vis-à-vis conventional schemes.

7.2 Recommendations for Further Research

Given the juncture at which this research has been brought to at present, the following are further avenues for possible future work.

- Use of EMRAN: As the number of neurons in the neural network MRAN increases, its performance deteriorates due to the EKF. An improved version of the neural network called Extended MRAN (EMRAN) has been developed. It is recommended that EMRAN be deployed in the simulation setup to investigate the performance vis-a-vis MRAN, GAP and the conventional schemes.

- **Hardware implementation of the neural CAC schemes:** The neural network driven call admission control schemes and the use of OPNET have conclusively proven that MRAN and GAP are good candidates for performance enhancement of ATM traffic management. A physical implementation of the neural network should be undertaken to study the performance of the simulated setup under real circumstances. A DSP implementation of the GAP/MRAN neural network in an ATM network with appropriate tools to track the relevant metrics such as call blocking ratio should be done to study the reliability of the simulated results and deductions. Such an implementation will have multiple advantages. It will explore the adequacy of the proposed algorithms in terms of speed of operation, sensitivity to design parameters and integration with the available service provider hardware. In addition, such a hardware implementation will also provide further avenues for experimentation with larger networks and a wider set of arrival patterns and real world traces.
- **Analytical study:** The research conducted at present is limited to performance results based on simulated conditions. It is important to validate the simulated results by conducting a complete analytical study of the neural CAC schemes in the ATM network context.
- **Expansion of scope:** The scope of the research study may be expanded to incorporate a complete set of adaptive traffic management schemes. This would include cell, call and network level controllers. Special focus should be given to cell and call level traffic control. In addition, the neural network control algorithms could be modified to have faster responses and more intelligence to tackle congestion and call admission problems. Other neural network control methods such as Recurrent Radial Basis Functions may be introduced in order to have more compact networks.

- Fine-tuning of the neural threshold values: The MRAN and GAP algorithms contain several threshold values, which are selected based on experience. These values may be optimal for a particular problem, but not for others. A detailed study is recommended which would provide guidelines on the selection of these threshold and parameter values for various types of problems.

Author's Publications

M. Aiyar and N. Sundararajan, "Using MRAN for Call Admission Control of ATM Networks in ATM", *Proc. Of OPNETWORK'03*, Washington DC, Aug 2003.

M. Aiyar, N. Sundararajan, "RBF Neural Networks for Call Admission Control in ATM Networks", In preparation.

References

[1] ITU-T Recommendation I.150, B-ISDN Asynchronous Transfer Mode Functional Characteristics,

<http://www.itu.int/rec/recommendation.asp?type=items&lang=E&parent=T-REC-I.150-199902-I>

[2] ITU-T Recommendation I.311, B-ISDN General Aspects,
<http://www.itu.int/rec/recommendation.asp?type=items&lang=E&parent=T-REC-I.311-199608-I>

[3] ITU-T Recommendation I.321, B-ISDN Protocol Reference Model and its application,
<http://www.itu.int/rec/recommendation.asp?type=items&lang=E&parent=T-REC-I.321-199104-I>

[4] ITU-T Recommendation I.356 B-ISDN ATM Cell Transfer Performance,
<http://www.itu.int/rec/recommendation.asp?type=items&lang=E&parent=T-REC-I.356-200402-I!Amd1>

- [5] ITU-T Recommendation I.361, B-ISDN ATM Layer Specification, <http://www.itu.int/rec/recommendation.asp?type=items&lang=E&parent=T-REC-I.361-199902-1>.

- [6] Linawati and H. Mehrpour, "Delay Analysis of Self-Similar traffic in ATM networks", *5th IEEE International Conference on High Speed Networks and Multimedia Communications*, November 2002, pp.37-41.

- [7] K. Shiimoto, S. Chaki and N. Yamanaka, "A Simple Bandwidth Management Strategy Based on Measurements of Instantaneous Virtual Path Utilization in ATM networks", *IEEE/ACM Transactions on Networking*, Vol. 6, No. 5, October 1998, pp. 625-634.

- [8] D. Gaiti and G. Pujolle, "Performance Management Issues in ATM Networks: Traffic and Congestion Control", *IEEE/ACM Transactions of Networking*, Vol. 4, No. 2, April 1996, pp 249-257.

- [9] R. Jain, S. Kalyanaraman, S. Fahmy and R. Goyal, "Source Behaviour for ATM ABR Traffic Management: An Explanation.", *IEEE Communications Magazine*, November 1996.

- [10] S. Kalyanaraman, R. Jain, S. Fahmy, R. Goyal and B. Vandalore, "The ERICA Switch Algorithm for ABR Traffic Management in ATM Networks", *IEEE/ACM Transactions on Networking*, Volume 8, No.1, February 2000, pp. 87-98.

- [11] Y.K. Park and G. Lee, "Applications of Neural Networks in High Speed Communication Networks", *IEEE Communications Magazine*, October 1995, pp 68-74.

- [12] W.M. Moh, M.J. Chen, N.M. Chu and C.D. Liao, "Traffic Prediction and Dynamic Bandwidth Allocation over ATM: a Neural Network Approach". *Computer Communications*, Vol. 18, No. 8, August 1995, pp. 563-571.
- [13] Mil 3 Inc., OPNET Modelling Manual - OPNET Version 3.5, Vol. 1, 1997.
- [14] Mil 3 Inc., OPNET Models/Protocols Manual - OPNET Version 3.5, Vol. 1, 1997.
- [15] C. Xinjie, T.Y. Tan and K.R. Subramanian, "Source Traffic Modelling in OPNET", *OPNETWORK 1999 Proceedings*, May 1999.
- [16] R. Goyal, R. Jain, S. Fahmy and S. Narayanaswamy, "Modelling Traffic Management in ATM Networks with OPNET", *OPNETWORK 1998 Proceedings*, May 1998.
- [17] T. Suzuki, "ATM Adaptation Layer Protocol", *IEEE Communications Magazines*, Vol. 32, No. 4, April 1994, pp 80-83.
- [18] T.R. Henderson, "Design Principles and Performance Analysis of SSCOP: a New ATM Adaptation Layer Protocol", *ACM Sigcomm: Computer Communications Review (Special Issue on ATM)*, Vol. 25, No. 2, April 1995, pp. 47-59.
- [19] W.C. Lee, "Topology Aggregation for Hierarchical Routing in ATM networks.", *ACM Transactions on Modelling and Computer Simulation*, Vol. 10, Issue 4, October 2000, pp.295-325.

- [20] K. Srirarn and W. Whitt, "Characterizing Superposition Arrival Processes in Packet Multiplexers for Voice and Data". *IEEE Journal on Selected Areas of Communications*, Vol. SAC-4, Sept 1986, pp. 833-846.
- [21] B. Maglaris, D. Anastassiou, P. Sen, and J. D. Roberts, "Performance Models of Statistical Multiplexing in Packet Video Communications", *IEEE Transactions Communications*, Vol. 36, July 1988, pp 834-843.
- [22] D.P. Heyman and T.V. Laksman, "Source Models for VBR Broadcast-Video Traffic", *IEEE/ACM Transactions Networking*, Vol. 4, Feb 1996, pp. 40-48.
- [23] "ATM Forum Traffic Management Specification", Version 4.0, *ATM Forum*, April 1996.
- [24] E. Nordstrom, J. Carlstrom, O. Gallmo and L. Asplund, "Neural Networks for Adaptive Traffic Control in ATM Networks", *IEEE Communications Magazine*, October, 1995, pp. 43-49.
- [25] A.E. Eckberg "B-ISDN / ATM Traffic and Congestion Control", *IEEE Networks*, September 1992, pp. 28-37.
- [26] "Traffic Control and Congestion Control in B-ISDN", *ITU Recommendations Helsinki Vol. I. 371*, July 1993.
- [27] N. Shenoy, "A Distributed Flow Control Scheme for Data Traffic Improves Performance in ATM Networks", *IEEE International Conference on Telecommunications Proceedings, Greece*, 1998.

[28] J.J. Bae and T. Suda, "Survey of Traffic Control Protocols in ATM Networks", *GLOBECOM'91 Proceedings*, 1991, pp.300 1.1 – 300 1.6.

[29] Raj Jain. "Myths about Congestion Management in High Speed Networks.", *Internetworking: Research and Experience*, Vol. 3, 1992, pp. 101-113.

[30] S. Kamolphiwong, A.E. Karbowiak and H. Mehrpour, "Survey Flow Control in ATM Networks", *Computer Communications*, March 21, pp. 951-968.

[31] A. Pitsillides, and J. Lambert, "Adaptive Congestion Control in ATM based Networks: Quality of Service and High Utilization", *Computer Communications*, Vol. 20, March 1997, pp. 1239-1258.

[32] S. Jagannathan and J. Talluri, "Traffic Rate Control of ATM Networks: A Neural Network Approach", *Automatica*, Vol 3, 2002, pp. 815-9.

[33] S. Jagannathan and J. Talluri, "Adaptive Predictive Congestion Control of High-Speed ATM Networks", *IEEE Transactions on Broadcasting*, Vol. 48, No. 2, June 2002, pp.129-139.

[34] S. Jagannathan and A. Tohmaz, "Admission Control Design for High Speed Networks: A Hybrid System Approach", *Journal of High Speed Networks*, to appear in 2005.

[35] N. Yim and M.G. Hluchyi. "A Dynamic Rate Control Mechanism for Source Coded Traffic in a Fast Packet Network", *IEEE Journal on Selected Areas of Communications*, Vol. 9, no. 7, Sept 1991, pp. 1003-1012.

- [36] R. Jain, "Congestion Control in Computer Networks: Issues and Trends ", *IEEE Network Magazine*, May 1990, pp. 24-30.
- [37] M. Decina, T. Toniatti, P. Vaccari and L. Verri, "Bandwidth Assignment and Virtual Call Blocking in ATM Networks," *IEEE INFOCOM*, 1990, pp 881-888.
- [38] H. Esaki, K. Iwamura and T. Kodama "A Simple and Effective Call Admission Method for an ATM Network", *IEEE GLOBECOM*, 1990, pp. 300.5.1-300.5.6.
- [39] H. Esaki, "Call Admission Control Method in ATM Networks," *IEEE ICC*, 1992, pp. 354.4.1-354.4.6.
- [40] T. Kamitake and T. Suda, "Evaluation of an Admission Control Scheme for an ATM Network considering Fluctuations in Cell Loss Rate," *IEEE GLOBECOM*, 1989, pp 49.4.1-49.4.7.
- [41] B. Jabbari and F. Yenenolu, "An Upper Bound of Cell Loss Probability of Bursty Sources in Broadband Packet Networks," *IEEE ICC*, 1991, pp 23.3.1-23.3.6.
- [42] B. Jabbari and F. Yenenolu, "An Efficient Method for Computing Cell Loss Probability for Heterogeneous Bursty Traffic in ATM Networks", *J. Digital Analog Communications*, 1992, pp 39-48.
- [43] T. Murase, H. Suzuki, S. Sato and T. Takeuchi, "A Call Admission Control Scheme for ATM Networks Using a Simple Quality Estimation," *IEEE Journal on Selected Areas of Communications*., Vol 9, 1991, pp. 968-981.

- [44] R. Guerin, H. Ahmadi and M. Naghshineh, "Equivalent Capacity and its Application to Bandwidth Allocation in High-Speed Networks," *IEEE Journal on Selected Areas of Communications*, Vol 9, 1991, pp. 968-981.
- [45] A.J. Elwalid and D. Mitra, "Effective Bandwidth of General Markovian Traffic Sources and Admission Control of High Speed Networks," *IEEE INFOCOM*, 1993, pp. 256-265.
- [46] T.H. Lee, K.C. Lai and S.T. Duann, "Design of a Real-time Call Admission Controller for ATM Networks," *IEEE/ACM Transactions on Networking*, Vol. 4, No. 5, October 1996.
- [47] H. Saito, "Call Admission Control in an ATM Network Using Upper Bound Cell Loss Probability," *IEEE Trans. Commun.*, vol 9, Sept 1992, pp. 1512-1521.
- [48] C. Shim, I. Ryoo, J. Lee, and S. Lee, "Modelling and Call Admission Control Algorithm of Variable Bit Rate Video in ATM Networks", *IEEE Journal on Selected Areas of Communications*, Vol. 12, No. 2, February 1994, pp. 332-344.
- [49] O.O. Raif, "Asynchronous Transfer Mode Networks: Performance Issues", *Artech House Publishers*, 2nd Edition, October 1995.
- [50] B.E. Helvic, "Synthetic Load Generation for ATM Traffic Measurements", *Manuscript of paper Telektronikh*, Vol. 91, No. 2/3, October 1995, pp. 174 – 194.
- [51] K. Steenhaut and K. Degieter, W. Brissinck, and E. Dirkx, "Scheduling and Admission Control Policies: A Case Study for ATM", *Computer Networks and ISDN Systems*, 1997, pp. 539-554.

[52] N. Ghani, S. Nananukul, and S. Dixit, "ATM Traffic Management Considerations for facilitating Broadband Access", Nokia Research Centre; *IEEE Communications Magazine*, November 1998, pp. 98-104.

[53] H.W. Lee and J.W. Mark, "Capacity Allocation in Statistical Multiplexing of ATM Sources." *IEEE/ACM Transactions Networking* 3, No.2, 1995, pp. 139-151.

[54] J. Platt, "A Resource Allocating Network for Function Interpolation", *Neural Computation*, Vol. 3, Issue 2, 1991, pp.213-225.

[55] V. Kadirkamanathan and M. Niranjan, "A Function Estimation Approach to Sequential Learning with Neural Networks", *Neural Computation*, Vol 5, 1993, pp 954-975.

[56] L. Yingwei, N. Sundararajan, and P. Saratchandran, "Performance Evaluation of a Sequential Minimal Radial Basis Function (RBF) Neural Network Learning Algorithm", *IEEE transactions on Neural Network*, Vol. 9, No. 2, March 1998, pp. 308-318.

[57] L. Yingwei, N. Sundararajan and P. Saratchandran, "A Sequential Learning Scheme for Function Approximation using Minimal Radial Basis Function Neural Networks", *Neural Computation, Massachusetts Institute of Technology*, Vol. 9, June 1997, pp. 461 – 478.

[58] L. Yingwei, N. Sundararajan and P. Saratchandran, "Identification of Time Varying Nonlinear Systems using Minimal Resource Allocation Network Neural

Network Learning Algorithm", *M.Eng Thesis, School of Electrical and Electronic Engineering, NTU*, 1997.

[59] G.B. Huang, P. Saratchandran and N. Sundararajan, "An Efficient Sequential Learning Algorithm for Growing and Pruning RBF (GAP-RBF) Networks", *IEEE Transactions on Systems, Man and Cybernetics*, Part B, Vol. 34, No. 6, 2004, pp. 2284-2292.

[60] M. Masugi, "A Neural Network Approach to Cell Loss Rate Estimation for Call Admission Control Estimation in ATM Networks", *IEICE Transactions Communications*, Vol. E80 – B, No 3, March 1997, pp. 412-419.

[61] A. Hiramatsu, "Integration of ATM Call Admission Control and Link Capacity Control by Distributed Neural Networks", *IEEE Journal on Selected Areas of Communications*, Vol. 9, No. 7, September 1991, pp. 1131-1138.

[62] R. Bolla, F. Davoli, T. Parisini, and P. Maryni, "An Adaptive Neural Network Admission Controller for Dynamic Bandwidth Allocation", *IEEE Transactions on Systems, Man and Cybernetics*, Part B: Cybernetics, Vol. 28, No. 4, August 1998, pp. 592-600.

[63] I. Habib, "Applications of Neurocomputing in Traffic Management of ATM Networks", *Proceedings of the IEEE*, Vol. 84, No. 10, October 1996, pp. 1430-1441.

[64] R. G. Cheng, and C. J. Chang, "Neural Network Connection Admission Control for ATM Networks", *IEE Proceedings Communications*. Vol. 144, No. 2, April 1997, pp. 93-98.

[65] S. Omatu, M. Khalid and R. Yusof, "Neuro-Control and its Applications", *Springer-Verlag Telos*, October 1995.

[66] K.G. Cheng, C.J. Chang and L.F. Lin, "A QoS Provisioning Neural Fuzzy Connection Admission Controller for Multimedia High-Speed Networks", *IEEE/ACM Transactions on networking*, Vol. 7, No. 1, February 1999, pp. 111-121.

[67] P.L. Tien, J.M. Hah and M.C. Yuang, "Neural Network based Call Admission Control in ATM Networks with Heterogeneous Arrivals", *Computer Communications* 20, May 1997, pp. 732 – 740.

[68] A. Hiramatsu, "Training Techniques for Neural Network Applications in ATM", *IEEE Communications Magazine*, October 1995, pp. 58-67.

Appendix A - Traffic Scenario

Parameters Setup in OPNET

Traffic Scenario 1 – The incident traffic consisted of profiles wherein the application under use was FTP only. The traffic scenario was simulated under three conditions ranging from low to medium to high QoS traffic contracts.

Profile – Profile 1	
	Value
Profile Name	Profile 1
Application	FTP
Start Time Offset	Exponential (5)
Duration (seconds)	End of Profile
Repeatability	Once at Start Time
Operation Mode	Simultaneous
Start Time	Uniform (100, 110)
Duration (seconds)	End of Simulation
Inter-repetition	Constant (300)
Number of	None
Repetition Pattern	Serial
Application - FTP	
Attribute	Value
Command Mix	50%
Inter-request time	Exponential (100)
File size (bytes)	Triangular (100,
Type of service	Best effort (0)
Back-End Custom	Not used
QoS - HIGH	
Attribute	Value
Incoming PCR	0.5
Outgoing PCR	0.5
CDVT	Maximum
Incoming MCR	0.5
Outgoing MCR	0.5
Incoming SCR	0.5
Outgoing SCR	0.5
MBS – Incoming	50

MBS – Outgoing	50
Requested QoS	CBR
PpCDV Incoming	10
PpCDV Outgoing	10
Max CTD	200
Max CTD	200
Cell Loss Ratio	3 e-7
Cell Loss Ratio	3 e-7
QoS - MEDIUM	
Attribute	Value
Incoming PCR	0.1
Outgoing PCR	0.1
CDVT	Maximum
Incoming MCR	0.1
Outgoing MCR	0.1
Incoming SCR	0.1
Outgoing SCR	0.1
MBS – Incoming	10
MBS – Outgoing	10
Requested QoS	CBR
PpCDV Incoming	10
PpCDV Outgoing	10
Max CTD	200
Max CTD	200
Cell Loss Ratio	3 e-7
Cell Loss Ratio	3 e-7
QoS - LOW	
Attribute	Value
Incoming PCR	0.05
Outgoing PCR	0.05
CDVT	Maximum

Incoming MCR	0.05	PpCDV Incoming	10
Outgoing MCR	0.05	PpCDV Outgoing	10
Incoming SCR	0.05	Max CTD	200
Outgoing SCR	0.05	Max CTD	200
MBS – Incoming	5	Cell Loss Ratio	3 e-7
MBS – Outgoing	5	Cell Loss Ratio	3 e-7
Requested QoS	CBR		

Table Appendix A – Parameter setup for Traffic Scenario 1

Traffic Scenario 2 – The incident traffic consisted of profiles wherein the applications under use were FTP and Email only. The traffic scenario was simulated under three conditions ranging from low to medium to high QoS traffic contracts.

Profile – Profile 1			
Attribute	Value	QoS - HIGH	
Profile Name	Profile 1	Attribute	Value
Application	FTP	Incoming PCR	0.5
Start Time Offset	Exponential (5)	Outgoing PCR	0.5
Duration (seconds)	End of Profile	CDVT	Maximum
Repeatability	Once at Start Time	Incoming MCR	0.5
Operation Mode	Simultaneous	Outgoing MCR	0.5
Start Time	Uniform (100, 110)	Incoming SCR	0.5
Duration (seconds)	End of Simulation	Outgoing SCR	0.5
Inter-repetition	Exponential (300)	MBS – Incoming	50
Number of	None	MBS – Outgoing	50
Repetition Pattern	Serial	Requested QoS	CBR
Profile – Profile 2		PpCDV Incoming	10
Attribute	Value	PpCDV Outgoing	10
Profile Name	Profile 2	Max CTD	200
Application	EMAIL	Max CTD	200
Start Time Offset	Uniform (5, 10)	Cell Loss Ratio	3 e-7
Duration (seconds)	End of Profile	Cell Loss Ratio	3 e-7
Repeatability	Once at Start Time	QoS - MEDIUM	
Operation Mode	Simultaneous	Attribute	Value
Start Time	Uniform (100, 110)	Incoming PCR	0.1
Duration (seconds)	End of Simulation	Outgoing PCR	0.1
Inter-repetition	Exponential (300)	CDVT	Maximum
Number of	None	Incoming MCR	0.1
Repetition Pattern	Serial	Outgoing MCR	0.1
Application - FTP		Incoming SCR	0.1
Attribute	Value	Outgoing SCR	0.1
Command Mix	50%	MBS – Incoming	10
Inter-request time	Exponential (100)	MBS – Outgoing	10
File size (bytes)	Triangular (100,	Requested QoS	CBR
Type of service	Best effort (0)	PpCDV Incoming	10
Back-End Custom	Not used	PpCDV Outgoing	10
Application – EMAIL		Max CTD	200
Attribute	Value	Max CTD	200
Send Inter-arrival	Exponential (180)	Cell Loss Ratio	3 e-7
Send Group Size	Constant (3)	Cell Loss Ratio	3 e-7
Receive Inter-	Exponential (60)	QoS - LOW	
Email size (bytes)	Constant (1000)	Attribute	Value
Type of service	Best effort (0)	Incoming PCR	0.05
Back-End Custom	Not used	Outgoing PCR	0.05

CDVT	Maximum	Requested QoS	CBR
Incoming MCR	0.05	PpCDV Incoming	10
Outgoing MCR	0.05	PpCDV Outgoing	10
Incoming SCR	0.05	Max CTD	200
Outgoing SCR	0.05	Max CTD	200
MBS – Incoming	5	Cell Loss Ratio	3 e-7
MBS – Outgoing	5	Cell Loss Ratio	3 e-7

Table Appendix B– Parameter setup for Traffic Scenario 2

Traffic Scenario 3 – The incident traffic consisted of profiles wherein the applications under use were FTP, Email and Voice only. The traffic scenario was simulated under three conditions ranging from low to medium to high QoS traffic contracts.

Profile – Profile 1		Command Mix	50%
Attribute	Value	Inter-request time	Exponential (100)
Profile Name	Profile 1	File size (bytes)	Triangular (100,
Application	FTP	Type of service	Best effort (0)
Start Time Offset	Exponential (5)	Back-End Custom	Not used
Duration (seconds)	End of Profile	Application – EMAIL	
Repeatability	Once at Start Time	Attribute	Value
Operation Mode	Simultaneous	Send Inter-arrival	Exponential (180)
Start Time	Uniform (100,110)	Send Group Size	Constant (3)
Duration (seconds)	End of Simulation	Receive Inter-	Exponential (60)
Inter-repetition	Exponential (300)	Email size (bytes)	Constant (1000)
Number of	None	Type of service	Best effort (0)
Repetition Pattern	Serial	Back-End Custom	Not used
Profile – Profile 2		Application – VOICE (PCM Quality	
Attribute	Value		
Profile Name	Profile 2	QoS - HIGH	
Application	EMAIL	Attribute	Value
Start Time Offset	Uniform (5, 10)	Incoming PCR	0.5
Duration (seconds)	End of Profile	Outgoing PCR	0.5
Repeatability	Once at Start Time	CDVT	Maximum
Operation Mode	Simultaneous	Incoming MCR	0.5
Start Time	Uniform (100, 110)	Outgoing MCR	0.5
Duration (seconds)	End of Simulation	Incoming SCR	0.5
Inter-repetition	Exponential (300)	Outgoing SCR	0.5
Number of	None	MBS – Incoming	50
Repetition Pattern	Serial	MBS – Outgoing	50
Profile – Profile 3		Requested QoS	CBR
Attribute	Value	PpCDV Incoming	10
Profile Name	Profile 2	PpCDV Outgoing	10
Application	VOICE	Max CTD	200
Start Time Offset	Uniform (5, 10)	Max CTD	200
Duration (seconds)	End of Profile	Cell Loss Ratio	3 e-7
Repeatability	Once at Start Time	Cell Loss Ratio	3 e-7
Operation Mode	Simultaneous	QoS - MEDIUM	
Start Time	Uniform (100, 110)	Attribute	Value
Duration (seconds)	Exponential (60)	Incoming PCR	0.1
Inter-repetition	Exponential (300)	Outgoing PCR	0.1
Number of	None	CDVT	Maximum
Repetition Pattern	Serial	Incoming MCR	0.1
Application - FTP		Outgoing MCR	0.1
Attribute	Value	Incoming SCR	0.1

Outgoing SCR	0.1	CDVT	Maximum
MBS – Incoming	10	Incoming MCR	0.05
MBS – Outgoing	10	Outgoing MCR	0.05
Requested QoS	CBR	Incoming SCR	0.05
PpCDV Incoming	10	Outgoing SCR	0.05
PpCDV Outgoing	10	MBS – Incoming	5
Max CTD	200	MBS – Outgoing	5
Max CTD	200	Requested QoS	CBR
Cell Loss Ratio	3 e-7	PpCDV Incoming	10
Cell Loss Ratio	3 e-7	PpCDV Outgoing	10
QoS - LOW		Max CTD	200
Attribute	Value	Max CTD	200
Incoming PCR	0.05	Cell Loss Ratio	3 e-7
Outgoing PCR	0.05	Cell Loss Ratio	3 e-7

Table Appendix C – Parameter Setup for Traffic Scenario 3