Connection Admission Control in Multiservice Networks: Simplicity versus Efficiency

Teck Kiong Lee

Submitted in total fulfillment of the
requirements of the degree of

Doctor of Philosophy

November 2002

Department of Electrical and Electronic Engineering

The University of Melbourne

Abstract

This thesis addresses the fundamental issue of efficient admission control operations for the purpose of maximizing network utilization subject to meeting Quality of Service (QoS) requirements. The aims of this thesis are twofold. Firstly, to provide two novel Connection Admission Control (CAC) frameworks; and secondly, to investigate how complex a CAC scheme needs to be in order to achieve a certain network efficiency level. Accordingly, this thesis investigates simplicity versus efficiency tradeoffs for various CAC schemes and provides practical recommendations. We conduct and report on an intensive comparative investigation of the two CAC frameworks' performance using different realistic traffic traces under both homogeneous and heterogeneous traffic streams scenarios.

The two novel CAC frameworks are named Model and Histogram based frameworks, and they contain CAC schemes that share common traits in their admission control algorithms. In addition, some of these schemes are customizable to the network provider's traffic control requirements.

The Model-based framework comprises CAC schemes that use traffic models to aid in making admission decisions. We have incorporated the traditional Gaussian and Effective Bandwidth traffic models into this framework. The Gaussian CAC scheme is based on the central limit theorem. However, we have observed that the aggregate traffic instead converges to Gaussian 'slowly'. When the number of connections is below a certain threshold, the aggregate traffic will be non-Gaussian. Hence, we propose an enhanced version of the Gaussian CAC scheme, which considers the total number of connections, and the resulting multiplexing gain effect, in its service

bandwidth computation. To gauge the level of traffic aggregation, various Gaussian boundaries for different traffic genres are derived.

We also introduce the measurement-based counterparts, i.e., measurement-based CAC (MBCAC), for the traditional Gaussian and the enhanced Gaussian CAC schemes. Other than the peak rate, all other relevant traffic statistics are measured in real-time. Some MBCAC schemes include an Adaptive Feedback Control Mechanism (AFCM) that adapts these schemes to changing traffic load conditions.

The traditional Effective Bandwidth CAC scheme is based on the concept that the effective bandwidth of the aggregate traffic stream is equal to the sum of the effective bandwidth of the individual traffic streams. In other words, this CAC scheme does not consider the presence of other neighboring connections in the link. Hence within this framework, we compare the performance of this scheme against other schemes that consider multiplexing instead, i.e., the traditional Gaussian and the enhanced Gaussian CAC schemes.

All schemes within the model-based framework require at least one a-priori traffic information, and this is the performance margin's multiple factor table specific to a traffic genre. We investigate the effects of using a default set of tables, on connections belonging to the same traffic genres but whose traffic statistics are not closely matched.

The Histogram-based CAC framework is made up of different modules, with each module containing different techniques with common functionality. These modules also include the adaptive feedback controller – AFCM. The schemes created within this framework are all measurement-based, thus only the user-declared peak rate parameter is required. By 'mixing and matching' techniques taken from every module, an MBCAC scheme can be constructed specially for use in a network with certain traffic control requirements. Admission decision is supported by a novel procedure of 'Available bandwidth' evaluation. Using past arrival traffic statistics and three

fundamental CAC algorithms, the bandwidth values required to service the established connections whilst still meeting the QoS requirement, are computed. Based on the choice of AFCM techniques and the instantaneous traffic load condition, an overall spare/available bandwidth value is then derived.

To obtain statistics on past arrival traffic, the histogram-based framework records the amount of arrival work into a set of traffic histograms. Each histogram holds a collection of traffic load measured from consecutive windows with each window having a fixed time-frame. Hence, the framework maintains traffic load records across multiple time-scales. To ensure the available bandwidth is evaluated accurately, different histogram update techniques are used. These update techniques, ranging from no update to complete updates, vary in complexity and storage requirement.

Another module within the histogram-based framework contains a technique to increase link utilization through easing a constraint that is imposed on the algorithm that computes the service bandwidth values.

Overall, the model and histogram based frameworks can be used to test various traffic control strategies, and in particular, to focus on the simplicity versus efficiency tradeoff issues. That is, if significant complexity does not improve efficiency substantially, simpler admission control methods should be used instead.

Declaration

This is to certify that:

(i) The thesis comprises only my original work towards the PhD,

(ii) Due acknowledgement has been made in the text to all other material used,

(iii) The thesis is less than 100,000 words in length, exclusive of tables, maps, bibliographies and appendices.

Teck Kiong Lee

Acknowledgements

This thesis is dedicated especially to my family and wife for their unwavering love, trust and support.

In addition, I would like to whole-heartedly thank my supervisor, Professor Moshe Zukerman, for giving his time and knowledge so generously during our numerous insightful discussions. This thesis is a culmination of our steadfast research efforts.

I would also like to thank my friends for their constant encouragements.

Table of Contents

List of Figures

List of Tables

List of Acronyms and Abbreviations

AAL	ATM Adaptation Layer
ABR	Available Bit Rate
AFCM	Adaptive Feedback Control Mechanism
ATM	Asynchronous Transfer Mode
AWF	Adaptive Weight Feedback
AWP	Adaptive Warming-up Period
B-ISDN	Broadband Integrated Services Data Networking
BO	Buffer Occupancy
CAC	Connection Admission Control
CBR	Constant Bit Rate
CDV	Cell Delay Variation
CDVT	Cell Delay Variation Tolerance
CIF	Common Interchange Format
CLCP	Cell Loss Conservative Period
CLP	Cell Loss Priority
CLR	Cell Loss Ratio
CTD	Cell Transfer Delay
DiffServ	Differentiated Services
EB	Effective Bandwidth
eGA	Enhanced Gaussian
EU	Exact histogram Update
FIFO	First-In First-Out
GA	Gaussian
GCRA	Generic Cell Rate Algorithm
GOP	Group of Pictures

IETF	Internet Engineering Task Force
IntServ	Integrated Services
IP	Internet Protocol
ISP	Internet Service Providers
ITU-T	International Telecommunications Union – Telecommunications
IWF	Inter Working Function
LEOS	Low Earth Orbit Satellite
LO	Link Occupancy
LRD	Long Range Dependence
MBCAC	Measurement-based Connection Admission Control
MBS	Maximum Burst Size
MCR	Minimum Cell Rate
MP	M/Pareto
MPEG	Motion Picture Experts Group
NMS	Network Management System
nrt-VBR	Non-real-time Variable Bit Rate
NT	Network data Traffic
NU	No histogram Update
OLFB	Overall Link Free Bandwidth
PCR	Peak Cell Rate
PNNI	Private Network-Network Interface
PRA	Peak Rate Allocation
PRAFB	PRA Free Bandwidth
PVC	Permanent Virtual Connection
QoS	Quality of Service
RCBR	Renegotiated Constant Bit Rate
REM	Rate Envelope Multiplexing

RFC	Request For Comments
RM	Resource Management
RS	Rate Sharing
RSFB	RS Free Bandwidth
rt-VBR	Real-time Variable Bit Rate
SCR	Sustainable Cell Rate
SD	Standard Deviation
SSQ	Single Server Queue
SVC	Switched Virtual Connection
TCP	Transmission Control Protocol
TM	Traffic Management
UBR	Unspecified Bit Rate
UNI	User-Network Interface
UPC	Usage Parameter Control
VBR	Variable Bit Rate
VC	Virtual Connection
VP	Virtual Path
VT	Video Traffic
WP	Warming-up Period
WWW	World Wide Web
ZU	ZT histogram Update

1 Introduction

1.1 Quality of Service

Quality of Service (QoS) is a generic term given to certain characteristics associated with providing service at a network access point, and it is specified in term of a set of parameters. In other words, the term QoS basically refers to the packet delivery service provided by the network operator, and characterized by traffic parameters such as packet delay rates, and packet loss rate.

In the early days of networking, the concept of QoS did not really exist because delivering packets to their destination was the first and foremost concern. During this time, the Transmission Control Protocol/Internet Protocol (TCP/IP) underlying mechanisms evolved to make the most efficient use of this paradigm. Congestion management and differentiation of services were not critical issues. The principal interest was simply keeping the traffic flowing, the network links up, and the routing system stable.

Since that initial period, not much has changed other than the fact that the same problems have been amplified significantly. In addition to these problems, the networking community now faces many more complex issues pertaining to policy, scaling, and stability. Only recently has the community seen a surge in research interest in the areas of QoS.

In the commercial Internet environment, QoS can be a competitive mechanism that provides a more distinguished service than those presently offered by the Internet Service Providers (ISPs). ISPs are collectively seen by many users to be the same, except for the services they provide. Hence, ISP operators generally view QoS as a valuable service that will give them a competitive edge, and also provides them with an additional source of revenue.

We consider a connection-oriented multiservice network that is Asynchronous Transfer Mode (ATM) based. This ATM model is developed by the ATM Forum and presented in the Anchorage Accord [ATM96a]. This is a milestone document which comprises sixty specifications including foundation specifications for building an ATM infrastructure, and expanded feature specifications for enabling migration to ATM multiservice networks. The Accord establishes criteria to ensure interoperability of ATM products and services between current and future specifications.

Even though ATM networks are nowadays used mainly for its excellent packet transport technology, the ATM service classes are still relevant to current IP research. In this thesis, we focus on providing ATM QoS to already established connections through the traffic control function called Connection Admission Control (CAC). The ideas and concepts expressed here are now being translated for use in future IP services.

The problem of providing IP Integrated Services [BCS94] within an ATM network is looked at by Garrett et al. [GB98] and expressed in the Request For Comments (RFC) 2381. The authors considered the service types, parameters and signaling elements needed for service interoperation. Figure 1-1 shows the service mapping and Virtual Connection (VC) management functions located in a network edge device which acts as both IP router and ATM interface. The Inter Working Function (IWF) abstractly summarizes the tasks that are not executed by IP or ATM; and these tasks are segregated into the control and data planes. These mappings serve to provide effective end-to-end QoS for IP traffic that traverses ATM networks.

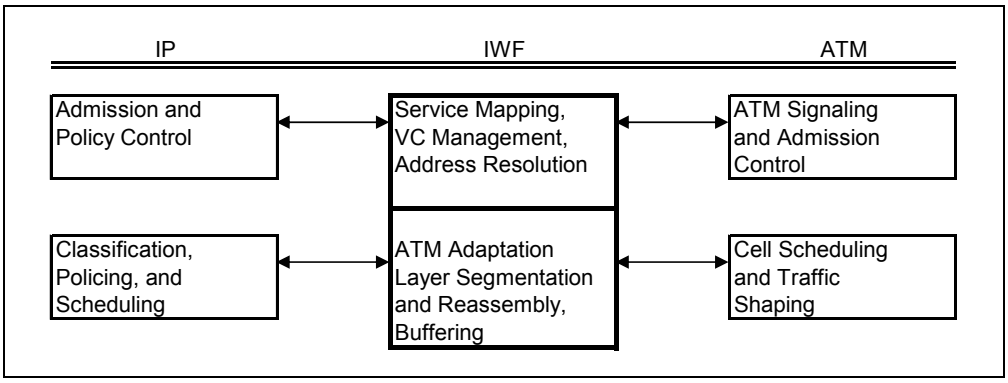


Figure 1-1. Edge device functions showing the IWF.

To ensure the success of delivering QoS to already established connections, the ATM network providers need to control traffic because of its limited link capacity. The primary responsibility of the network traffic control function is to promote network efficiency and avoids traffic congestion so that the overall network performance does not degenerate. That is, the bandwidth demand imposed by transporting one form of application data should not adversely impact the capability to efficiently transport other traffic in the network. For example, the transport of bursty traffic should not introduce an excessive amount of jitters into the transportation of other traffic such as constant bit rate, and real-time video or audio applications.

To deliver this stability, the ATM Forum [ATM99] has defined the following set of functions to be used independently or in conjunction with one another to provide for traffic management and control of network resources:

- **Connection Admission Control (CAC)** – Actions taken by the network during connection set-up phase to determine whether a connection request can be accepted or rejected.

- **Usage Parameter Control (UPC)** – Actions taken by the network to monitor and control traffic and to determine the validity of ATM connections and the associated traffic transmitted into the network. The primary purpose of UPC is to protect the network from traffic

misbehavior that can adversely impact the QoS of already established connections. UPC detects violations of negotiated traffic parameters and takes appropriate actions by either tagging cells as CLP = 1 or discarding cells altogether.

- **Cell Loss Priority (CLP) Control** – If the network is configured to distinguish the indication of the CLP bit, the network may selectively discard cells with their CLP bit set to 1 in an effort to protect traffic with cells marked as a higher priority (CLP = 0). Different strategies for network resource allocation may be applied, depending on whether CLP = 0 or CLP = 1 for each traffic flow.

- **Traffic Shaping** – ATM devices may control traffic load by implementing leaky-bucket traffic shaping to control the rate at which traffic is transmitted into the network. A standardized algorithm called Generic Cell Rate Algorithm (GCRA) is used to provide this function.

- **Network-resource Management** – Allows the logical separation of connections by Virtual Path (VP) based on their service criteria.

- **Frame Discard** – A congested network may discard traffic at the ATM Adaptation Layer (AAL) frame level, rather than at the cell level, in an effort to maximize discard efficiency.

- **ABR Flow Control** – Available Bit Rate (ABR) flow-control protocol to adapt subscriber traffic rates in an effort to maximize the efficiency of available network resources. ABR flow control also provides a feedback mechanism to re-route traffic around a particular node whenever loss or congestion events are detected, or when the traffic contract is in danger of being violated as a result of a local connection admission control decision. With the feedback mechanism, an intervening node signals back to the originating node that it no longer

is viable for a particular connection and hence can no longer deliver the committed QoS requirements.

1.2 Connection Admission Control

An ATM network aims to efficiently utilize its limited resources whilst still meeting the desired QoS requirements. However, due to the unpredictable statistical fluctuations of the traffic flows in the network, congestion may occur and the desired QoS requirements may not be met.

In this thesis, QoS is assured through the use of various CAC schemes within an ingress ATM switch. Because the emphasis is on the effectiveness of these schemes, the ATM switch is simplified to be a simple buffered or non-buffered Single Server Queue (SSQ). Even though we consider the CAC function as the only traffic control tool for the SSQ, the results presented here will give network providers the underlying principles as well as specific traffic control guidelines on how to implement a QoS-compliant multiservice network.

CAC is specified in [ATM99, ITU00d] as a set of actions taken by the network to decide whether a new connection is accepted or rejected. It is a preventive congestion control mechanism that ensures a certain desired level of QoS can be experienced by all connections including the newly admitted. To be efficient, a CAC has to accurately predict the amount of traffic load that may be submitted in the near future by all connections, whenever it makes a decision to admit or reject a request to set-up a new connection.

Typically, when a new connection request arrives, and the route between the origin node and the destination node is established, all network bottlenecks along the end-to-end route are checked to make sure that sufficient capacity is available for the new connection without sacrificing the QoS requirements. If sufficient capacity is available, the new connection is admitted, otherwise the new connection is rejected and another route may be considered instead. In

another scenario, instead of an individual new connection request, it may be a capacity request for an aggregate of traffic flows.

The various CAC schemes considered in this dissertation differ in the way they compute the available capacity values. Since the connections considered here are typically Variable Bit Rate (VBR) applications, the available capacity values will vary with time, including the capacity required by the new connection. Therefore, the CAC decisions are in fact made under uncertain environment. Consequently, a conservative CAC scheme will be less efficient because it over-allocates the required resources but as a result it is more likely to meet QoS requirements, while a more efficient and daring CAC scheme may be at risk of not meeting the QoS requirements.

1.3 Focus of this Thesis

This thesis addresses the fundamental issue of efficient admission control operations for the purpose of ensuring maximum network utilization for the network providers and guaranteed level of QoS for the established connections. The aims of this thesis are: (1) to provide two novel CAC frameworks; and (2) to investigate how complex a CAC scheme needs to be in order to achieve a certain network efficiency level. Accordingly, this thesis investigates simplicity versus efficiency tradeoffs for various CAC schemes and provides practical recommendations.

In the later part of this thesis, we conduct and report on an intensive comparative investigation of the performance of the two CAC frameworks under both homogeneous and heterogeneous traffic scenarios. Different realistic traffic traces recorded from network data and video sources are used in these studies.

It is worth noting that although we consider CAC in the context of ATM service classes, the ideas and concepts expressed here are general across most multiservice networks, including networks providing IP services.

There are basically two approaches to CAC [BJS99, JSD97, KQ98]:

- **Model-based Approach** – This approach computes the amount of network resources required to support a set of flows based on the a-priori traffic characteristics provided.

- **Measurement-based Approach** – This approach relies mainly on taking measurements of actual traffic load in order to derive relevant traffic statistics that are used to aid in making admission decisions.

The admission control literature is already quite extensive. In this dissertation, we propose and investigate two novel CAC frameworks that are named Model and Histogram based frameworks. These frameworks contain CAC schemes that share common traits in their admission control algorithms. Some of the CAC schemes proposed here are customizable to the network provider's traffic control requirements. In other words, these frameworks enable many new schemes to be constructed. We believe these frameworks are both novel and practical enough to merit attention.

The Model-based framework comprises CAC schemes that use traffic models to aid in making admission decisions. In addition to making certain traffic behavior/modeling assumptions, these schemes also require a-priori traffic information. During an admission decision process, an equivalent bandwidth value equal to the minimal amount of bandwidth required to service the established connections whilst still meeting the QoS requirement, is computed by the traffic model. From this equivalent bandwidth value, a spare/available bandwidth value is then derived. A new connection is admitted only if there is adequate spare bandwidth to service that connection's peak rate.

Amongst all traffic models proposed in the literature, the most popular is the Effective Bandwidth traffic model by Kelly [Kel91]. An alternative model is the Gaussian traffic model by Addie [Add98]. We incorporate these traditional traffic models into the model-based CAC framework. The

traditional Gaussian CAC scheme is based on the central limit theorem, and when applied to heavy traffic processes, whereby the resulting aggregate stream is highly aggregated, this aggregate traffic can be accurately modeled by a Gaussian traffic process

However, we have observed that the aggregate traffic instead converges to Gaussian 'slowly'. In other words, when the number of established connections is below a certain threshold unique to that type of traffic, the aggregate traffic stream will be lightly aggregated and hence exhibits non-Gaussian behavior. Nevertheless, the aggregate traffic stream will begin to exhibit Gaussian behavior as the number of connections increases over the threshold. Hence, a CAC scheme based purely on the Gaussian model is not applicable to this lightly aggregated traffic stream because the model considers the traffic to be Gaussian even though the total number of connections is below the threshold. As a result of this observation, we propose an enhanced version of the Gaussian model-based CAC scheme, which is also efficient if the traffic does not exhibit Gaussian behavior. It is efficient because it considers the total number of established connections, and the resulting multiplexing gain effect, into its equivalent bandwidth computations. Hence, with this additional process, the enhanced Gaussian CAC scheme is able to model the statistical behavior of the aggregate traffic stream more accurately.

To gauge the level of traffic aggregation, various Gaussian boundaries for different traffic genres, i.e., network data and video, are derived through empirical-based studies. These boundaries are expressed in term of the number of homogeneous connections required in order to attain a level of aggregation that will ensure the aggregate traffic is Gaussian.

In addition to these model-based CAC schemes, we also introduce the measurement-based counterparts, i.e., MBCAC, for the traditional Gaussian and the enhanced Gaussian CAC schemes. These MBCAC schemes require minimal a-priori traffic information to be provided. Some MBCAC schemes include an Adaptive Feedback Control Mechanism (AFCM) that adapts these

schemes to changing traffic load conditions. By configuring certain AFCM parameters, the network providers can customize these schemes according to their traffic control requirements.

All CAC and MBCAC schemes within the model-based framework require at least one a-priori traffic information, and this is the performance margin's multiple factor look-up table specific to a traffic genre. We investigate the effects of using a default table unique to a traffic genre, on connections transmitting work belonging to the same traffic genre but whose traffic statistics are not closely matched to that default multiple factor values. The motivation behind this study is that with a default set of tables for different traffic genres, the CAC and MBCAC schemes will be more easily deployable in a network, and it will also greatly simplify the use of these schemes.

The Histogram-based CAC framework is made up of different modules, with each module containing different techniques with common functionality. These modules also include the adaptive feedback controller – AFCM, which protects the network whenever an MBCAC scheme fails to meet the QoS requirement, either because the scheme tries to be too daring, or when the traffic exhibits unpredictable behavior, or both. The schemes created within this framework are purely measurement-based. Hence, except for the user-declared peak rate, no other traffic information is provided. By 'mixing and matching' techniques taken from every module, an MBCAC scheme can be constructed specially for use in a network with certain traffic control demands. Hence, many customized MBCAC schemes can be created within this framework.

When a new connection request arrives, the MBCAC scheme will make an admission decision supported by a novel procedure of 'Available bandwidth' evaluation. Using past arrival traffic statistics and three fundamental CAC algorithms [RMV96], i.e., Peak Rate Allocation, Rate Envelope Multiplexing, and Rate Sharing, the bandwidth values required to service established connections whilst still meeting the QoS requirement, are computed. Based

on the choice of AFCM techniques and the instantaneous traffic load condition, an overall spare/available bandwidth value is then derived. If the new connection's peak rate is less than the derived available bandwidth value, the new connection will be admitted. Otherwise, it will be rejected.

To obtain statistics on past arrival traffic, the histogram-based framework records the amount of arrival work into a set of traffic histograms. Each histogram holds a collection of traffic load measured from consecutive windows with each window having a fixed time-frame. Hence, the framework maintains traffic load records across multiple time-scales. To ensure the available bandwidth is evaluated accurately, different histogram update techniques are used.

The histogram update techniques, ranging from no update to complete updates, vary in complexity and storage requirement. For example, the most complex update technique involves measuring and then storing the amount of work submitted by a connection; and this process is repeated for all established connections. In other words, we maintain a depository that records the numerical amount of bandwidth consumed by a connection, for every connection. Hence, a per-flow traffic statistics can be computed for any connections.

Another module within the histogram-based framework contains a technique to increase link utilization through easing a constraint that is imposed on the algorithm that computes the service bandwidth values.

The AFCM is a generic component that can be used by a variety of different MBCAC schemes within both model and histogram based frameworks. It provides an additional control layer to the MBCAC schemes. In addition, it is simple to implement and imposes very minimal storage and computing demands on the network switches. It is basically a collection of two inter-dependent modules, i.e., (1) Prudence level policy module – which uses an active parameter to adapt the MBCAC scheme to changing traffic conditions,

and (2) Load and traffic measurements module – which compares the traffic load against a choice of different threshold values. The latter module contains techniques that basically issue advance QoS breach warnings to the prudence level policy module, based on the levels of traffic loading.

Overall, the model and histogram based CAC frameworks can be used to test various traffic control strategies, and in particular, to focus on the simplicity versus efficiency tradeoff issues. That is, if significant complexity does not improve efficiency substantially, simpler admission control methods should be used instead.

1.4 Sub-division of this Thesis by Chapter

This thesis is divided into 6 chapters. We begin in chapter 2 by addressing the various CAC schemes currently evolving in the literature, as well as the research issues still outstanding. In chapter 3, we introduce the two novel CAC frameworks. The first framework is made up of CAC and MBCAC schemes based on various traffic models, while the second framework is purely made up of MBCAC schemes that use real-time measurements of the arrival traffic load. In chapter 4, we outline the simulation methods used in our studies, including the assumptions that are made. In chapter 5, we report and discuss the results of the intensive comparative investigation of the performance of the two CAC frameworks.

Below is a brief overview of the thesis content:

- **Chapter 2: Quality of Service via Traffic Control** – We briefly introduce how connection admission control is used as a network traffic management tool. We look at the protocol and explore the way in which they work. In addition, we present an overview of the current state-of-the-art admission control schemes, as well as the research issues still outstanding.

- **Chapter 3: Connection Admission Control** – We describe our ***first framework*** – Model-based CAC framework, which is made up of CAC and MBCAC schemes based on various traffic models. These traffic models use a variety of a-priori traffic parameters that describe the statistical behavior of a traffic source. This framework is an extension of our research published in [LZ00, LZ99a, LZ99b, LZ99c, LZA01, LZC99].

  Next, we describe our ***second framework*** – Histogram-based CAC framework, which is made up of different modules, with each module containing different techniques with common functionality. The MBCAC schemes created within this framework are purely measurement-based. By 'mixing and matching' techniques taken from every module, an MBCAC scheme can be constructed specially for use in a network with certain traffic control requirements. Hence, many customized MBCAC schemes can be created within this framework. This framework is an extension of our research published in [LZ98, LZA01, ZL98a, ZL98b].

- **Chapter 4: Simulation Methodology** – We introduce our simulation methods and the assumptions that are made. We also outline the traffic sources used in our studies. The aim is to lay the groundwork for the performance studies that are conducted and reported in chapter 5.

- **Chapter 5: Comparative Performance Studies** – We report and discuss the results of the intensive comparative investigation of the performance of the two CAC frameworks. Accordingly, we investigate simplicity versus efficiency tradeoffs for various CAC and MBCAC schemes and provide practical recommendations.

1.5 Contributions of this Thesis

Below is a list of the original contributions of this thesis, along with the relevant chapters where the contributions are first discussed. In addition,

relevant publications are shown alongside. Some contributions span multiple publications as a result of the techniques' generic properties, hence making it applicable to different CAC schemes.

- Development of our own admission control philosophy based on observations and literature reviews. The philosophy provides the foundation for effective CAC procedures and specifies practical admission control methodologies for QoS assurance in a multiservice network (Chapter 3, [LZ99c, LZA01, ZL98b]).

- Formulation of two CAC frameworks containing a variety of CAC and MBCAC schemes. Some MBCAC schemes are customizable to specific traffic control requirements (Chapter 3, [LZ99c, LZA01, ZL98b]).

- Development of a framework for model-based CAC and MBCAC schemes that use either a-priori or measured traffic parameters to help determine the amount of bandwidth required by the aggregate traffic stream. To compute the bandwidth value, these schemes make certain traffic behavior/modeling assumptions (Chapter 3, [LZ00, LZ99a, LZ99b, LZ99c, LZC99]).

- Development of the enhanced Gaussian CAC and MBCAC schemes within the model-based framework. These schemes consider the total number of established connections, and the resulting multiplexing gain effect, into its service bandwidth computations. This additional process will enable the schemes to model the statistical behavior of the aggregate traffic stream with increased accuracy (Chapter 3, [LZ99c]).

- Investigation of a threshold that will distinguish aggregate traffic stream between non-Gaussian and Gaussian behavior, for a particular traffic genre. This threshold is expressed in term of the number of homogeneous connections. From our study, a variety of Gaussian

boundaries are derived for a range of traffic genres (Chapter 4, [LZ99c]).

- Investigation of the effects of using a generic set of traffic parameters unique to a traffic genre, to compute the bandwidth required to service connections belonging to the same traffic genre but whose traffic statistics are not closely matched to that generic values (Chapter 5).

- Development of a framework for histogram-based MBCAC schemes that use traffic statistics derived from past traffic load to help determine the amount of bandwidth required by the aggregate traffic stream (Chapter 3, [LZ98, LZA01, ZL98a, ZL98b]).

- Development of a procedure to evaluate 'Available bandwidth' through the use of past arrival traffic statistics and three fundamental CAC algorithms (Chapter 3, [LZ98, LZA01, ZL98a, ZL98b]).

- Development of a variety of different traffic histogram update techniques that vary in complexity and storage requirement. These update techniques, ranging from no update to complete updates, are used to ensure the service bandwidth values are computed accurately whenever connections depart from the network (Chapter 3, [LZ98, LZA01, ZL98a, ZL98b]).

- Investigation of the effects of relaxing the cell loss rate L constraint on MBCAC schemes within the histogram-based framework. This technique basically increases link utilization by easing a constraint that is imposed on the algorithm that computes the service bandwidth values (Chapter 5, [LZ98, LZA01, ZL98b]).

- Development of an adaptive feedback control mechanism that enables the admission decision process to be adaptive to varying traffic load conditions. Depending on the traffic load, this feedback controller will change the admission decision behavior from being conservative to

daring, and vice versa. This feedback controller is a generic component that can be used by a variety of different MBCAC schemes. It is basically a collection of two inter-dependent modules, i.e., Prudence level policy module and Load and traffic measurements module (Chapter 3, [LZ98, LZ99b, LZA01, LZC99, ZL98a, ZL98b]).

- Investigation of the effects of using different 'prudence level policy' techniques to adapt the MBCAC schemes to varying traffic load conditions (Chapter 5, [LZ98, LZA01, ZL98a, ZL98b]).

- Investigation of the effects of using different 'load and traffic measurements' techniques to provide advance QoS breach warnings to the prudence level policy module (Chapter 5, [LZ98, LZA01, ZL98a, ZL98b]).

- Investigation of the performance of all CAC and MBCAC schemes within the model and histogram based frameworks. In these intensive comparative studies, the schemes are subjected to both homogeneous and heterogeneous traffic streams made up of traffic sources using realistic network data and video traces (Chapter 5, [LZ00, LZ98, LZ99a, LZ99b, LZ99c, LZA01, LZC99, ZL98a, ZL98b]).

1.6 Publications arising from this Thesis

The list below summarizes the papers published as a result of the work presented in this thesis:

[LZ98] Teck Kiong Lee and Moshe Zukerman. 'Simple Measurement-based Connection Admission Control for Heterogeneous Traffic Sources.' International Conference on Telecommunications (ICT) '98: Bridging East and West Through Telecommunications. In *Proceedings of ICT '98*, F. N. Pavlidou (ed.), Aristotle University, Thessaloniki, vol. 1, pp. 518 - 522. Porto Carras, Chalkidiki, Greece. 22 - 24 June 1998.

[ZL98a] Moshe Zukerman and Teck Kiong Lee. 'A Measurement-based Connection Admission Control for ATM Networks.' IEEE International Conference on ATM (ICATM) '98. In *Proceedings of ICATM '98: IEEE International Conference on ATM*, P. Lorenz (ed.), ISDN 0-7803-4982-2, pp. 140 - 144. Colmar, France. 22 - 24 June 1998.

[ZL98b] Moshe Zukerman and Teck Kiong Lee. 'A Framework for Real-time Measurement-based Connection Admission Control in Multi-service Networks.' IEEE Global Telecommunications Conference (GLOBECOM) '98. In *Proceedings of IEEE GLOBECOM '98*, H. Bradlow (ed.), IEEE Communications Society, ISDN 0-7803-4984-9, pp. 2983 - 2988. Sydney, N.S.W., Australia. 8 - 12 November 1998.

[LZ99a] Teck Kiong Lee and Moshe Zukerman. 'An Efficiency Study of Different Model-based and Measurement-based Connection Admission Control Techniques Using Heterogeneous Traffic Sources.' IEEE ATM Workshop '99. In *Proceedings of IEEE ATM Workshop '99*, H. Terada et al. (eds.), Institute of Electronics, Information and Communication Engineers (IEICE) Japan and IEEE Communications Society, ISDN 4-88552-164-5, pp. 89 - 94. Kochi, Japan. 24 - 27 May 1999.

[LZC99] Teck Kiong Lee, Moshe Zukerman and Fraser Cameron. 'Utilization Comparisons for Several Admission Control Schemes

under Realistic Traffic Conditions.' Seventh International Federation for Information Processing (IFIP) Workshop on Performance Modelling and Evaluation of ATM Networks. In *Proceedings of the Seventh IFIP Workshop on Performance Modelling and Evaluation of ATM Networks*, D. D. Kouvatsos (ed.), IFIP Working Group 6.3: Performance of Communication Systems. Antwerp, Belgium. 28 - 30 June 1999.

[LZ99b] Teck Kiong Lee and Moshe Zukerman. 'Practical Approaches for Connection Admission Control in Multiservice Networks.' IEEE International Conference on Networks (ICON) '99. In *Proceedings of ICON '99*, J. Indulska and A. Seneviratne (eds.), IEEE Computer Society, ISDN 0-7695-0243-1, pp. 172 - 177. Brisbane, Queensland, Australia. 28 September – 1 October 1999.

[LZ99c] Teck Kiong Lee and Moshe Zukerman. 'Efficiency Comparisons between Different Model-based and Measurement-based Connection Admission Control Schemes under Heavy Traffic.' IEEE Global Telecommunications Conference (GLOBECOM) '99. In *Seamless Interconnection for Universal Services, GLOBECOM '99*, R. Sampaio-Neto and E. de Souza e Silva (eds.), IEEE Communications Society, ISDN 0-7803-5796-5, vol. 2, pp. 1629 - 1633. Campinas, Rio de Janeiro, Brazil. 5 - 9 December 1999.

[LZ00] Teck Kiong Lee and Moshe Zukerman. 'Connection Admission Control Techniques With and Without Real-time Measurements.' Institute of Electronics, Information and Communication

Engineers (IEICE) Transactions on Communications: IEICE/IEEE Joint Special Issue on Recent Progress in ATM Technologies, T. Takahashi (ed.), ISDN 0916-8516, vol. E83-B, no. 2. 25 February 2000.

[LZA01] Teck Kiong Lee, Moshe Zukerman and Ronald G. Addie. 'Admission Control Schemes for Bursty Multimedia Traffic.' Conference on Computer Communications (INFOCOM), Twentieth Annual Joint Conference of the IEEE Computer and Communications Society. In *Proceedings of IEEE INFOCOM '01*, Daniel W. Repperger (ed.), IEEE Computer and Communications Societies, ISDN 0-7803-7016-3, vol. 1, pp. 478 - 487. Anchorage, Alaska, U.S.A. 22 - 26 April 2001.

[LZA02] Teck Kiong Lee, Moshe Zukerman and Ronald G. Addie. 'Admission Control Schemes for Bursty Multimedia Traffic.' *Submitted to IEEE/ACM Transactions on Networking (TON)*. August 2002.

2 Quality of Service via Traffic Control

2.1 Introduction

In recent years, there has been a steady growth in the development and deployment of Asynchronous Transfer Mode (ATM) networks. ATM is the well-defined and logical result of the past 10 years of Broadband Integrated Services Data Networking (B-ISDN) related research work. Today, the broadband technology rides on two realities, i.e., Internet Protocol (IP) services and an ATM infrastructure. Currently, IP and the World Wide Web (WWW) are changing our lives and the way we do business. ATM is the technology that provides the infrastructure necessary to meet this increasing demand on broadband services and applications. It aims to extend the global public telephone and data services of today into an all-encompassing global networking infrastructure for tomorrow.

One area of significant importance attracting high broadband-related research activities is traffic management, and one of its major mechanisms is traffic congestion control. The primary role of a network congestion control procedure is to protect the network and the user in order to achieve network performance objectives, and at the same time optimize the usage of network resources. Basically, the control procedure reacts to network congestion by minimizing its intensity, spread and duration. Through this congestion control

procedure, different levels of network performance can be provided for established connections whilst meeting the Quality of Service (QoS) requirements.

Congestion control procedure can be classified into two schemes:

- Preventive congestion control.
- Reactive congestion control.

Both schemes have advantages and disadvantages. In preventive control, a scheme is set-up to prevent the occurrence of congestion; whilst in reactive control, feedback information is relied upon for controlling the level of congestion. In ATM networks, a combination of these two control methods is currently used in order to provide effective congestion control and hence QoS for specific applications. For example, Constant Bit Rate (CBR) and Variable Bit Rate (VBR) services use preventive scheme, while Available Bit Rate (ABR) service is based on the reactive scheme.

The focus of this thesis is on a particular preventive congestion control procedure called Connection Admission Control (CAC) for CBR and VBR services. CAC is defined by the ATM Forum [ATM99] and the International Telecommunications Union – Telecommunications (ITU-T) [ITU00d] as a set of actions taken by the network during connection establishment phase in order to determine whether the connection can be progressed or should be rejected. Depending on the CAC function, a connection request is progressed only when: (1) sufficient resources are available at each successive network element for the purpose of establishing the connection in the network; (2) its desired QoS, traffic contract, and service category requirements can be met; and (3) the agreed QoS required by the existing connections can still be maintained.

Typically, when a user requests a connection to be set-up, the user will indicate the QoS requested per direction, the negotiated characteristics of the connection, and the type of service required.

The remainder of this chapter is as follows:

- Section 2.2 describes the ATM-layer QoS parameters used in the end-to-end network performance evaluation.

- Section 2.3 introduces the architecture for services provided at the ATM-layer. In section 2.3.1, the traffic parameters and descriptors used to characterize a connection traffic behavior are described; while in section 2.3.2, the service categories, e.g., CBR and VBR, defined by the ATM Forum to provide QoS for specific applications are elaborated further.

- Section 2.4 describes in detail the preventive congestion control function – CAC. The CAC function makes use of traffic contract information to compute admission decisions for CBR and VBR services. Alternative admission control schemes using information derived from real-time measurements of the arrival traffic are also discussed.

2.2 ATM-layer Quality of Service

Networks today not only must deliver services that their customers demand, they must also provide services when customers demand them and at a cost which meets the customers' expectations. Achieving this is about delivering QoS. It is an absolute concept, and it means reliably providing a service to a subscriber that matches their objectives. QoS is not about delivering a perfect service, nor is it simply about priority. A network that implements QoS can predict and guarantee the service that will be provided to any users. This may be a very high performance service or it may be best-effort service.

The network provider must take the bandwidth, circuits, and switches that comprise the network, and harness them to meet the user's goals. QoS is an end-to-end issue. It is measured by the end users from their own perspective, without regard to the state of the network. QoS may cover the throughput, the end-to-end delay, the delay variation, the data loss, or a combination of all these parameters. Each user will have an implicit or explicit contract with the network to deliver what is required.

The ATM-layer QoS is measured by a set of parameters characterizing the performance of a connection. Basically, these parameters quantify the end-to-end network performance. ITU-T Recommendation I.356 [ITU00c] refers to these parameters as network performance parameters. A network may support one or more performance objectives for each of the QoS parameters. QoS is negotiated amongst the networks and the end-systems for each direction of a connection. The network agrees to meet or exceed the negotiated QoS as long as the end-system complies with the negotiated traffic contract. QoS commitments are probabilistic in nature, and are intended to be only a first order approximation of the performance that the network expects to offer over the duration of the connection. In reality, QoS does varies over the duration of a connection since there is no limit to the connection's holding time, plus the network can only make admission decisions based on traffic information available at the time the new connection request arrives. In addition, transient events like uncontrollable impairments in transmission systems can cause short-term performance observations to be worse than the agreed QoS requirements. Hence, QoS commitments can only be evaluated over the long-term and over multiple connections with similar QoS commitments.

Table 2-1 lists the ATM-layer QoS parameters along with their commonly used acronyms. The last column provides an indication of whether the ATM Forum's specifications define a means for the user to negotiate the QoS parameter with a network. These specifications are: the Traffic Management (TM) Specification version 4.1 [ATM99], the User-Network Interface (UNI)

Signalling Specification Version 4.1 [ATM02a], and the Private Network-Network Interface (PNNI) Specification Version 1.1 [ATM02b].

It is observed that propagation delay dominates the fixed delay component in the wide area networks, while queuing behavior contributes to delay variations in heavily loaded networks. The effects of queuing strategy and buffer sizes dominate loss and delay variation performance in congested networks. A large single, shared buffer results in lower loss, but greater average delay and delay variation.

QoS parameter	QoS acronym	Negotiated?
Cell Delay Variation	Peak-to-peak CDV	Yes
Maximum Cell Transfer Delay	maxCTD	Yes
Cell Loss Ratio	CLR	Yes

Table 2-1. Quality of Service parameter terminology.

Below is the definition for the QoS parameters mentioned in Table 2-1:

- The Cell Delay Variation (CDV) parameter is defined as a measure of cell clumping, i.e., the difference in delay between successive cell arrivals. Cell clumping is of concern because if too many cells arrive too closely together, then it may cause the buffer to overflow. On the other hand, cell dispersion occurs if the network creates too great of a gap between cells, in which case the playback buffer would under-run. Standards define CDV using one-point and two-point measurement methods. A one-point measurement applies to CBR sources and determines the deviation from the nominal cell spacing. The two-point method measures the difference in cell spacing at an entry point and an exit point. An example of a two-point method is the peak-to-peak CDV associated with the CBR and VBR services, and determined by the probability α that a cell arrives late.

- The maximum Cell Transfer Delay (maxCTD) parameter specifies the accumulated delay between two measurement points for a specific virtual connection. It is the sum of coding, decoding, segmentation, reassembly, processing and queuing delays along the connection route.

- The Cell Loss Ratio (CLR) parameter is the value of negotiated CLR that the network agrees to offer as an objective over the lifetime of the connection. Basically, it measures the number of lost (i.e., not delivered) cells that do not reach the destination user. This parameter is defined for a connection as:

$$\text{Cell Loss Ratio} = \frac{\text{Lost Cells}}{\text{Total Transmitted Cells}}.$$

2.3 ATM Service Architecture

ATM has been conceived as a multiservice technology. The introduction of new service categories within the ATM layer makes ATM suitable for an unlimited range of applications. By using these categories as service building blocks, users have flexible access to the network resources and can achieve a good result in terms of performance and cost. In addition, network providers are able to share their network resources amongst different customers and meet their needs in a cost-effective manner. Furthermore, the concept of negotiating for each connection the expected behavior in terms of traffic and performance will enable users to better optimize their application requirements against the network capabilities. Given the presence of a heterogeneous traffic mix and the need to control the allocation of network resources, a much greater degree of flexibility and network utilization can be achieved by providing the selectable set of capabilities to the benefits of both users and network providers via the ATM-layer.

As the result of a major effort by many traffic management experts, the specification of the ATM-layer services is well documented by the ATM Forum and the ITU-T. The documents addressing the traffic management and

congestion control issues can be found in the ATM Forum TM 4.1 [ATM99] and the ITU-T Recommendation I.371 [ITU00d]. An ATM Service Category (the ATM Forum name) or ATM-layer Transfer Capability (the ITU-T name) is intended to represent a class of ATM connections that have homogeneous characteristics in terms of traffic pattern, QoS requirements, resource allocation method, and possible use of control mechanisms. Table 2-2 illustrates the definition of the ATM-layer services by the ATM Forum and ITU-T. These services are elaborated further in the section 2.3.2.

ATM Service Category (ATM Forum TM 4.1)	**ATM Transfer Capability (ITU-T I.371)**	**Typical use**
Constant Bit Rate (CBR)	Deterministic Bit Rate (DBR)	Real-time, QoS guarantees
Real-time Variable Bit Rate (rt-VBR)	(under development)	Statistical mux, real-time
Non-real-time Variable Bit Rate (nrt-VBR)	Statistical Bit Rate (SBR)	Statistical mux
Available Bit Rate (ABR)	Available Bit Rate (ABR)	Resource exploitation, feedback control
Unspecified Bit Rate (UBR)	(no equivalent)	Best effort, no guarantees
(no equivalent)	ATM Block Transfer (ABT)	Burst level feedback control

Table 2-2. Definition of ATM-layer services by the ATM Forum and ITU-T.

The ATM Service Architecture makes use of traffic control and congestion control procedures and parameters, to achieve its main aim of protecting the network in order to achieve high network performance. An additional role is to optimize the use of network resources. To meet these objectives, the set of functions forming the framework for managing and controlling traffic and congestion can be used in appropriate combinations.

The service category associates quality requirements and traffic characteristics to network behavior. It is intended to specify a combination of QoS commitment and traffic parameters that is suitable for a given set of applications. Functions such as CAC (see section 2.4) are made available within the ATM node equipment and are generally structured differently for each service category.

The following sections will define in detail:

- Section 2.3.1 – Traffic descriptors used to characterize a connection.
- Section 2.3.2 – Service categories used to provide QoS for specific applications.

2.3.1 Connection Traffic Parameters and Descriptors

The ATM traffic contract binds a network provider to a user by guaranteeing a specified QoS, if and only if, the user's packet flow conforms to a negotiated set of traffic parameters. These traffic parameters describe an inherent characteristic of a traffic source. Traffic parameters defined by the ATM Forum include Peak Cell Rate (PCR), Sustainable Cell Rate (SCR), Maximum Burst Size (MBS), and Minimum Cell Rate (MCR).

For a given connection, traffic parameters are grouped into a source traffic descriptor, which in turn is a component of a connection traffic descriptor. A source traffic descriptor is the set of traffic parameters of the traffic source. It is used during the connection establishment to capture the intrinsic traffic characteristics of the connection requested by a particular source.

The connection traffic descriptor specifies the traffic characteristics of the ATM connection. The connection traffic descriptor includes the source traffic descriptor, the Cell Delay Variation Tolerance (CDVT), and the conformance definition that is used to unambiguously specify the conforming cells of the connection. The CAC procedures will use the connection traffic descriptor to allocate resources and ensure network performance objectives can still be achieved once the connection is accepted.

The connection traffic descriptor can be declared as shown:

- A mandatory PCR in conjunction with a CDVT.
- An optional SCR in conjunction with a MBS.

Below is a list of the traffic parameters that captures intrinsic source traffic characteristics.

- Peak Cell Rate (PCR) is a mandatory traffic parameter that has to be declared explicitly or implicitly in a source traffic descriptor during the connection establishment and contract negotiation phase. PCR specifies the upper bound on the traffic that can be submitted by a connection. The PCR value should be the same along a given connection when negotiated and contracted at the connection establishment phase or modified and re-negotiated via signaling.

- Cell Delay Variation (CDV) is a means to determine the variance in cell delay, mainly intended to measure the amount of jitter. A connection's CDV may result in different transmission traffic characteristics from the original statistics declared by the traffic source. CDV is normally introduced when more than one connection is multiplexed at an input port. A connection's packet may be delayed by other connections' packets whilst being queued at the multiplexer input. In fact, Skliros [SKG92, Skl93] has even shown that CBR sources are not immune to CDV whenever a network statistically multiplex traffic together. CDVT traffic parameter, expressed in units of seconds, constrains the number of cells the user can send at the physical medium rate. This parameter normally cannot be specified by the user, but is set by the network.

- Sustainable Cell Rate (SCR) is a traffic parameter that characterizes a bursty, on-off traffic source. It defines the maximum allowable rate for a source in terms of the PCR and the MBS. It is equal to the ratio of the MBS to the minimum burst inter-arrival time. For VBR service, declaring only the PCR parameter may not be sufficient enough to understand the traffic characteristics. Hence, the optional SCR parameter may be declared in order to allow the network to provision the trunk capacity more efficiently. When used together with the MBS

parameter, the network will have a better understanding of the VBR traffic source and may multiplex the sources together to achieve multiplexing gain.

- Maximum Burst Size (MBS) is a traffic parameter that specifies the maximum number of cells that can be transmitted at the connection's PCR such that the maximum rate averaged over many bursts is no more than the declared SCR.

- Minimum Cell Rate (MCR) is a traffic parameter used only by the ABR service category. It specifies a rate at which the source may always transmit traffic.

From the definition of the five traffic parameters above, it is obvious that other than PCR, the remaining parameters are hard to evaluate. Typically, a new connection's traffic descriptor declared during the connection establishment phase is seldom detail and accurate. Take for example a workstation in a multi-tasking and multi-processing environment. At any moment, this workstation can be running one or more types of applications. Therefore it will be difficult, if not impossible, to accurately characterize the generated traffic. In other words, there is always a limit on the supply of information about the traffic of the requested connection. As shown by Rathgeb [Rat91, Rat93], not only are some of the traffic parameters, e.g., average connection duration and average burst rate, difficult to estimate, they are also extremely hard to police in some cases.

Theoretically, traffic descriptor is a powerful traffic management tool in a multiservice network. However, inaccurate traffic descriptor will result in an inefficient connection admission controller.

2.3.2 ATM Service Categories

To enable end users to fully utilize QoS for their specific applications, the ATM Forum has come up with a range of service categories employing simple acronyms relating to the bit rate and an implicit quality specification. Each service category definition includes terms that define the traffic contract parameters and QoS characteristics. Such service categories come about because the ATM Forum defines in great details the concept of QoS. The resulting complexity hence became a hindrance to the end users.

The ATM Forum TM 4.1 [ATM99] defines the following ATM-layer service categories:

- **Constant Bit Rate (CBR)** – It supports real-time applications requiring a fixed amount of capacity defined by the PCR, which is defined by the ITU-T [ITU00d] as the inverse of the minimum inter-arrival time between two successive cells. It is the maximum allowable rate at which cells can be transported by a connection. The PCR is the determining factor in how often cells are sent in relation to time in an effort to minimize jitters. For a CBR connection, PCR is the only traffic parameter that needs to be specified at connection request phase. CBR service category supports tightly constrained variations in delay. The term CBR normally refers to a constant bandwidth being assigned to a particular connection. For the whole duration of the connection, even if the allocated bandwidth is not utilized, the bandwidth will not be shared with other connections. Hence, valuable network capacity will be wasted if that connection remains idle most of the time.

  Examples of CBR applications will be any data/text/image transfer application that contains smooth enough traffic or for which the end-system's response time requirements justify occupying a fully reserved CBR channel. Typical applications are Videoconferencing, Interactive

audio (e.g., telephony), Audio/Video distribution (e.g., television, distance learning), and Audio/Video retrieval (e.g., video-on-demand, audio library). Normally, networks must allocate the peak rate to these types of sources.

- **Real-time Variable Bit Rate (rt-VBR)** – It supports time-sensitive applications, which require tightly bounded delay and delay variation requirements. These applications are characterized by the following traffic descriptors: PCR, SCR, and MBS. ITU-T [ITU00d] defines SCR as the rate that a bursty, on-off traffic source can send. It is an average allowable, long-term cell transfer rate for a specific connection. For MBS, it is defined as the maximum number of consecutive cells that a source can send at the peak rate. In other words, it is the maximum allowable burst size of cells that can be transmitted contiguously by a connection. The three parameters define a traffic contract in term of the worst-case source's traffic pattern, for which the network guarantees a specified QoS. Traffic streams from the rt-VBR sources are expected to be bursty and are delay sensitive.

  Examples of such bursty, delay variation sensitive sources are native ATM voice with bandwidth compression and silence suppression, and VBR video. For such applications, excessive delay in cell transmission will significantly reduce the quality of the received voice and video information. The advantage of this service category is that a network may statistically multiplex these types of traffic sources together to achieve network link efficiency.

- **Non-real-time Variable Bit Rate (nrt-VBR)** – It supports applications that have no constraint on delay and delay variation, but which still have variable-rate, bursty traffic characteristics. This service will guarantee a low CLR for traffic streams that complies with the traffic contract. The traffic contract is the same as that for rt-VBR.

Since non-real-time applications are supported, there is no delay bound associated with this service category.

Examples of such applications are packet data transfers, file transfers, and terminal sessions. As with rt-VBR traffic sources mentioned above, the network may also statistically multiplex these types of nrt-VBR traffic sources together to achieve network link efficiency

- **Available Bit Rate (ABR)** – It works in cooperation with sources that can change their transmission rate in response to rate-based network feedback used in the context of closed-loop flow control. It is defined by the ATM Forum as an ATM layer service category for which the limiting ATM layer transfer characteristics provided by the network may change subsequent to connection establishment. The aim of the ABR service is to dynamically provide access to capacity currently not in use by other service categories to users who can adjust their transmission rate in response to feedback. Hence, it attempts to fully explore the concept of statistical multiplexing by making use of the temporary available bandwidth released by the temporary idle CBR and VBR connections. ABR service does not provide bounded delay variation, hence it is not intended for real-time applications. ABR traffic sources are characterized by two traffic parameters. The first is the PCR, which is the maximum transmit rate; and the second parameter is the MCR, which is the minimum allowable rate at which cells can be transported along an ATM connection. In term of QoS, the network will provide very low CLR but no delay and delay variation guarantees. One major component of this service is a flow control mechanism that supports different types of feedback to control the source transmission rate in light of the varying network load conditions. The feedback mechanism employs a specific control cell called the Resource Management (RM) cell.

Examples of ABR applications are database archival, file transfer, web browsing and non-time-sensitive traffic. Another suitable application is LAN interconnection or internetworking services. These are typically run over router-based protocol stacks like TCP/IP that can easily vary their emission rate as required by the ABR rate control policy.

- **Unspecified Bit Rate (UBR)** – It is a best-effort service, which requires neither tightly constrained delay nor delay variation. Hence, it will only support applications without any real-time or time variance constraints. This service is signaled by the Best Effort Indicator bit in the ATM User Cell Rate Information Element. Generally, UBR provides no specific QoS or guaranteed throughput whatsoever. This traffic is therefore at risk since the network provides no performance guarantee. The Internet and LAN are examples of this type of best-effort delivery performance.

  Examples of UBR applications are IP over ATM, LAN emulation and Remote terminal (e.g., telecommuting). Other suitable applications include data/text/image file transfer submitted in the background of a workstation with minimal service requirements, Messaging, and Retrieval.

A summary of the attributes of these ATM-layer service categories is listed in Table 2-3. Table 2-4 shows the suitability of the ATM service categories for a number of commonly encountered applications.

		Guarantees			
Service Category	**Traffic Descriptor**	**Loss (CLR)**	**Delay Variance (CDV)**	**Bandwidth**	**Feedback Control**
CBR	PCR	Yes	Yes	Yes	No
rt-VBR	PCR, SCR, MBS	Yes	Yes	Yes	No
nrt-VBR	PCR, SCR, MBS	Yes	No	Yes	No
ABR	PCR, MCR, and behavior parameters	Yes	No	Yes	Yes
UBR	PCR	No	No	No	No

Table 2-3. ATM Forum service category attributes, QoS guarantees, and feedback usage.

Application	**CBR**	**rt-VBR**	**nrt-VBR**	**ABR**	**UBR**
Critical data	Good	Fair	Best	Fair	No
LAN interconnect	Fair	Fair	Good	Best	Good
WAN data transport	Fair	Fair	Good	Best	Good
Circuit emulation	Best	Good	No	No	No
Telephony	Best	Good	No	No	No
Videoconferencing	Best	Good	Fair	Fair	Poor
Compressed audio	Fair	Best	Good	Good	Poor
Video distribution	Best	Good	Fair	No	No
Interactive multimedia	Best	Best	Good	Good	Poor

Table 2-4. ATM Forum service categories associated with various common applications.

2.4 Connection Admission Control

The focus of this thesis is on preventive open-loop congestion controls for CBR and VBR services in a multiservice network. CAC is a preventive open-loop traffic congestion control function and it is defined by ITU-T as follows:

"The set of actions taken by the network at connection set-up phase, or during connection re-negotiation phase, in order to establish whether a Virtual

Channel or Virtual Path connection can be accepted."

ITU-T, 2000 [ITU00d]

When a user wants to transmit over a network, an end-to-end connection must first be set-up. The main objective of this procedure is to establish a path between the sender and the receiver, and this path may involve one or more switchers or routers. On each of these switches, resources have to be allocated to the new connection. If a new connection is accepted, bandwidth and/or buffer space in the switch is allocated for that connection. When the connection departs, all allocated resources will be released back to the network.

CAC is a function commonly implemented by software in ATM switches to determine whether to admit or reject connection requests. If connection negotiation is successful, then it is called a Traffic Contract. A connection request includes a set of traffic parameters, and either the ATM service category, requested QoS class, or the user specified QoS parameters. ATM switches use CAC to determine whether admitting the connection request at Permanent Virtual Connection (PVC) provisioning time or Switched Virtual Connection (SVC) connection origination time would violate the QoS already guaranteed to active connections. In other words, CAC admits the request only if the network can still guarantee QoS for all existing connections after accepting the request. For SVCs, each node performs CAC in a distributed manner. Connections are set up dynamically and terminated through signaling requests. However, for PVCs, CAC is performed at a centralized system called the Network Management System (NMS). PVCs are permanent or semi-permanent. This is because after they are configured and the connection established, they are not terminated until manual intervention. When a connection is accepted, the CAC will determine the policing and shaping parameters, the routing decisions, and the resource allocation. Network resources include trunk capacity and buffer space.

Two major QoS requirements are packet loss and packet delay. Furthermore, these requirements can be deterministic or statistical. Admission control schemes are designed to meet such request from new connections wanting a certain bound on either packet loss or delay. For a new connection with deterministic QoS requirements, the network guarantees a hard-bounded packet loss probability threshold or a maximum end-to-end packet delay. However, for a new connection with statistical QoS requirements, the network aims to provide a soft-bounded average packet loss probability, or an average end-to-end packet delay experience.

Typically, admission decision to accept or reject a new connection is made based on two considerations. The first consideration by the switch is that given the available trunk capacity, can it still meet the QoS requested by the new connection. The second consideration is that if the new connection is permitted into the trunk, will this new entry affect the QoS for the connections that are already established by the switch. Hence, admission decisions are dependent upon the switch state, the traffic behavior exhibited by all connections, and the QoS requested.

In general, admission control schemes can be classified either as:

- **Non-statistical Allocation** – Allocate more bandwidth than required to provide QoS guarantees for established connections, thereby resulting in network resources being under-utilized.

- **Statistical Allocation** – Under-allocate the bandwidth required by the connections, thereby necessitating additional precaution mechanisms to be used in conjunction with them.

For connection with deterministic QoS requirements, non-statistical allocation approach is used. An example is the Peak Rate Allocation (PRA) approach, which does not consider sharing bandwidth resource with other connections. In this approach, an amount of bandwidth equal to the peak transmission rate

for each connection is reserved. If the sum of the peak rates for every established connections plus the peak rate for the new connection is less than the trunk capacity, then the new connection is accepted into the link. Hence, PRA is a very simple approach to implement. However, the disadvantage of this simple approach is that in reality connections do not transmit at their peak rates all of the time, and therefore this will result in the low utilization of their reserved bandwidth and the overall trunk capacity.

For statistical QoS requirements, statistical allocation approach is used instead. An example is the Rate Envelope Multiplexing (REM) approach [RMV96]. This approach is based on the zero buffer approximation and it assumes the sharing of bandwidth resource with other connections, but not buffer space. Generally, statistical allocation approach allocates bandwidth that is less than the new connection's peak rate. In addition, the allocated bandwidth is not exclusively reserved for that new connection, instead it will be shared with all other established connections. Hence, admission decisions will result in the combined peak rates of all connections being greater than the trunk capacity. For bursty traffic sources, statistical allocation approach ensures efficient utilization of limited trunk capacity. The disadvantage of this approach is that QoS may be compromised as a result of its over-zealous admission policy. This stems from the inaccurate and difficult-to-measure statistical information of the traffic arrival process, for such information is used to make admission decisions.

To achieve high connection establishment rates, the CAC must be simple and fast. That is not to say that it must compromise its main objective of achieving maximum utilization whilst still guaranteeing QoS. How complex and accurate a CAC is depends on the connection traffic descriptor (see section 2.3.1), the service bandwidth estimation method, and the admission decision algorithm. The latter two requirements are closely inter-related because one cannot do without the other whenever a decision to admit or reject a new connection has to be made.

Lately, a lot of research has been done on the effectiveness of using the CAC as a network congestion controller for the purpose of guaranteeing QoS to the established connections [BJS00, BJS99, EP00, EP98, Flo96, JS97, JSD97, Kni97, KS99, PE96]. In sections 2.4.1 to 2.4.4, we will briefly describe some of the proposed admission control schemes available in the literature. Although some of these schemes may be used in the early network deployments, they should still be viewed as midway solutions to the problem.

In the list below, we begin by outlining some common and much generalized inadequacies amongst many admission control algorithms:

- When a traffic contract conformance-violation occurs from an established connection, policing actions are taken by a traffic congestion control function called the Usage Parameter Control (UPC), which is a set of actions taken by the network to monitor and control traffic. Its main purpose is to protect network resources from malicious as well as unintentional misbehavior, which can affect the QoS of other already established connections, by detecting violations of negotiated parameters and taking appropriate actions. Other than the negotiated PCR, the other traffic parameters such as SCR and MBS are difficult to police. In light of this difficulty, admission control schemes that rely on the latter two traffic parameters will have to implement a conservative admission policy so as not to compromise the QoS requirements.

- For the past few years, various CAC schemes have been proposed which require users to declare the characteristics of their traffic source. Various traffic parameters such as the PCR and SCR are required [FV90]. The disadvantage of such a requirement is the difficulties users have in accurately characterizing their traffic sources. This leads to a situation where users tend to over-estimate their traffic requirements in order to be cautious. The result of such over-estimation is the under-utilization of valuable link resource.

- In reality, detailed connection traffic descriptor that specifies the traffic characteristics of a connection is seldom accurate or available. This means that admission control schemes that rely on their users providing specific traffic parameters will be rendered ineffective when such parameters are not forthcoming. However, there is one traffic parameter that is always readily available from the user, and that is PCR. Compared to the other traffic parameters, PCR is a relatively easy value to measure. Therefore, admission control schemes that require more traffic parameters other than the PCR to be declared may not be robust enough to handle heterogeneous bursty traffic.

- Admission control schemes whose service bandwidth estimation algorithm relies solely on user-declared connection traffic descriptor may under-estimate the characteristics of the transmitted traffic. This is because the algorithm assumes the transmitted traffic will behave in a manner equal to the declared statistics.

- Some admission control schemes make admission decisions based on the effective bandwidth methodology. These generally assume that the computed effective bandwidth value would not change over time and is independent of other service classes. Consequently, these schemes are inadequate because the algorithm generally allocates bandwidth on a worst-case scenario.

In the following sections, we briefly discuss the salient features of each class of admission control schemes and review some of the proposed schemes within each category. The classification is based on the underlying principle that was used to develop the scheme.

2.4.1 Peak Rate Allocation

Peak Rate Allocation (PRA) scheme is an example of a non-statistical allocation method. Basically, this simple scheme entails reserving an amount of bandwidth equal to the peak rate for each source. Hence, if a source has an

average rate of m_j = 1 Mbps and a peak rate p_j = 3 Mbps, then PRA requires that 3 Mbps be reserved at the output port for the specific source, independent of whether the source transmits continuously at the peak rate.

One advantage of PRA is its simplicity in making admission decisions. A new connection is accepted if the sum of the peak rates of all the established connections including the new connection's peak rate p_{new} is less than the output link capacity C:

$$\sum_j p_j + p_{new} < C \,. \tag{2.1}$$

The other advantage is its ability to avoid packet loss. However, some packet loss may occur due to cell-scale effects. Packets belonging to a connection may be interleaved with other connections' packets. When packets belonging to a connection momentarily arrive faster than expected, the peak rate may be momentarily exceeded.

The disadvantage of PRA is that it does not exploit statistical multiplexing, which is the effect gained when many connections are multiplexed together. Hence, unless connections transmit at peak rates, the output link may be grossly under-utilized.

2.4.2 Effective Bandwidth

In the literature, numerous papers have been published on the effective bandwidth (or equivalent bandwidth or equivalent capacity) modeling of the arrival traffic in a network. Basically, the effective bandwidth approach views each connection at the queuing point in isolation, and then derives a real number c_j called the effective bandwidth of the connection such that the requested QoS is still satisfied. Thus, the CAC rule becomes very simple:

$$\sum_j c_j \le C \,,$$

where C is the output link capacity. A connection is admitted if there is available spare capacity, else it is rejected.

For a connection with an average rate SCR_j and peak rate PCR_j, the effective bandwidth lies between SCR_j and PCR_j, i.e., $SCR_j \leq c_j \leq PCR_j$. Typically, a connection's statistical traffic behavior and the congestion point's queuing property will affect the computed effective bandwidth value. For a switch with very small buffer space, the effective bandwidth will be close to the PCR; while for very large buffer space, the effective bandwidth will be close to the SCR.

For effective bandwidth to be useful it should have the following properties [RMV96]:

- **Additivity Property** – The effective bandwidth of the superposition of N streams is equal to the sum of the effective bandwidths of each stream.

- **Independence Property** – The effective bandwidth of a given traffic stream is only related to the statistical characteristics of that traffic stream and the network equipment, e.g., buffering capacity. It should not be dependent upon traffic characteristics of any other streams.

This approach is widely accepted and used because of its two inherent properties. In ATM technology, connections are set-up and torn down dynamically. Due to the additivity property of the effective bandwidth approach, whenever a connection is being set-up (or torn down), the CAC function can add (or subtract) the effective bandwidth of that connection from the total effective bandwidth.

However, it should be mentioned that due to the independence property, the approach can be more conservative than another approach that considers the statistical multiplexing of the other connections in the link. The logic is that true bandwidth needed to serve all the connections can be far less than the sum

of all the connections' effective bandwidth. Thus, having the independence property means that we cannot benefit from statistical multiplexing and henceforth we cannot achieve optimal efficiency. A justification for the argument that using the effective bandwidth approach can still achieve efficiency is based on certain mathematical arguments that show that the rate of the tail of unfinished work distribution under fractal traffic is not affected by traffic aggregation. However, Addie [Add98] has shown that actually the weight of the tail and not the rate of the tail is the significant factor in cell loss estimation.

Kelly [Kel91] showed that effective bandwidth exists for a GI/G/1 system with a constraint on tail probability and an M/G/1 system with constraints on mean workload. [EM93, GAN91, GH91] considered effective bandwidth for buffered network resources, while [Hui88, Kel91, Miy91] considered the bufferless case. In [KWC93, VW94, Whi93], results for the dominant negative exponential tails of certain non-Gaussian queues are obtained. [EM93] considered the effective bandwidth based on the stochastic fluid-flow model, while [KWC93] is based on the batch Poisson arrival process. In addition, both are based on the large deviation theory [TG97]. Zhang et al. [ZA94] discussed effective bandwidth for on/off sources with dependent and general distributions, while Kulkarni et al. [KGC94, KGC95] considered on/off sources with different priorities. In [AMS82], the overflow probability is approximated by solving an adequate system of differential equations that leads to a closed-form solution, while [BGRS94, NRSV91] provided a numerical solution.

Equivalent bandwidth method is also used by [DCLM89, Tur87, WKFR89] to allocate capacity based on source declarations and policing mechanisms. An extension of this method with some form of dependence on the current network state is looked at by [Jai95, LM94, Mit92]. Berger et al. [BW98] considered the case of network nodes using a priority-service discipline to support multiple classes of service. The authors derived multiple effective

bandwidths for a given connection, i.e., one for the priority level of that connection and one for each lower priority level.

In the equivalent bandwidth scheme by Guerin et al. [GAN91], the effective bandwidth is computed from the combination of two different approaches, one based on fluid-flow model [AMS82, Kos84, Mit88] and the other on an approximation of the stationary bit rate distribution. A connection is characterized based on a flow model in which the flow of bits is generated at a peak rate during the active period, whereas no bits are generated during the silent period. Assuming each source is characterized as a two-state on/off model, the duration at each state is exponentially distributed and independent of each other, the equivalent bandwidth c required by a source for a queue that corresponds to a cell loss rate ε, is given by:

$$c \cong p\frac{y - B + \sqrt{(y-B)^2 + 4B\rho y}}{2y}, \qquad (2.2)$$

where $y = ln(1/\varepsilon)d(1 - \rho)p$, p is the source's peak rate, B is the switch's buffer size, ρ is the source utilization, i.e., probability that the source is in an active state, and d the average duration of the active period. The total bandwidth C of n multiplexed connections is equal to the sum of the equivalent capacities of individual connections c_j, i.e., $C = \sum_{j=1}^{n} c_j$. This scheme assumes connections are not very bursty and have short average burst duration. It overestimates the bandwidth requirement whenever connections do not conform to that assumption.

The computed C overestimates the required bandwidth for the aggregate traffic since the interaction between the individual connections is not taken into consideration. To enable the effect of multiplexing, the Gaussian approximation is used together with the equivalent capacities. Hence, the total bandwidth C required for the aggregate traffic of n connections is now given by:

$$C = \min\left\{ m + \alpha' \sigma \quad , \quad \sum_{j=1}^{n} c_j \right\}, \tag{2.3}$$

where m and σ respectively denote the mean and standard deviation of the aggregate traffic. This scheme selects the best estimate of the two approximation methods. It requires traffic parameters such as the mean rate, peak rate, and burst duration of each connection to be declared.

Other effective bandwidth schemes using various inter-related techniques or source models can be found in references [CT95, CW95, DJM97, DLCRT95, EHLMW95, GG92, Gib96, GT99c, Lin91, Reg94, RMV96, VKW95]. Some effective bandwidth schemes may fail in specific situations highlighted by Elsayed et al. [EP97] and Choudhury et al. [CLW94, CLW96]. In particular, it fails when the probability that the traffic load exceeds the link capacity is assumed to be close to one for a bufferless system having the same input traffic.

2.4.3 Gaussian Approximation

Section 2.4.3.1 briefly describes the multiplexing gain relevant to Gaussian approximation. This is then followed by a review of some Gaussian admission control schemes in section 2.4.3.2.

2.4.3.1 Multiplexing Gain

Addie et al. [Add99, AMN99, AZ94a, AZ94b, AZN98] has shown that as a network carry increasing number of independent connections, the statistical multiplexing gain factor becomes increasingly significant to warrant re-consideration of the effective bandwidth scheme. In addition, they had illustrated that with the increased traffic aggregation, the unfinished work distribution weakly converge to a Gaussian model through the application of the central limit theorem onto the traffic of a network. In other words, the argument for the applicability of Gaussian model relies on the assumption that a very large number of sources are involved such that their superposition

follows a Gaussian process [LEWW95, PE95]. This is a realistic scenario given the penetration and the exponential growth in the demands for multimedia services. Here is an analysis of the Gaussian traffic model's multiplexing gain by Addie [Add98].

In [CLW96], the authors have argued that equivalent bandwidth computed base on the tail is not always effective. Many of the papers on dimensioning of ATM networks, especially those making use of large deviations theory, have proceeded on the assumption that precise estimation of the weight of an asymptotic tail of the buffer contents distribution is not necessary for dimensioning. This approach is brought into question by the results reported in [CLW96].

In fact, in many cases, it is the weight of the tail that matters and its rate is the irrelevant parameter. Dimensioning criteria based on the tail of the buffer contents distribution can be misleading if the weight of the tail is insignificant. The rate of the tail of the buffer contents distribution is not a continuous function of the traffic stochastic process as defined by weak convergence. However, there is a valid reason for using tails – if consideration is restricted to traffic models with certain regularity constraints, or properties, it may be valid to assume that the tail behavior is genuinely dominant over the entire stationary buffer contents distribution. This appears to be the case for a wide range of Gaussian input processes. However, when an argument requires consideration of a limit on traffic processes, for example as more and more traffic are aggregated together, there is a possibility that the limit could go into a region where tail behavior is no longer dominant, and therefore the use of tail behavior to characterize performance is no longer valid. In particular, this casts doubt on the use of the tail behavior of any non-Gaussian models under increasing aggregation because the queuing behavior in the limit tends to that of the corresponding Gaussian model [Add99], whereas the tail behavior is typically not the same as the Gaussian model. This situation is explained using the fact that the weight of the tail of the non-Gaussian model becomes

insignificant by comparison with the weight of the tail in the corresponding Gaussian model. It may well be that this is precisely what is happening in some of the examples depicted in [CLW96], where the asymptotic behavior appears to be inconsistent with the numerically computed stationary buffer contents distribution.

Even among Gaussian models, it is easy to identify cases where the tail behavior is misleading by simply mixing two traffics together, i.e., a very small amount of traffic with a very large autocovariance sum together with a large amount of traffic with a small autocovariance sum. The same argument can be applied to a mixture of two long range dependent Gaussian traffic processes. Hence, the tail behavior is likely to be misleading unless it is genuinely dominant. Furthermore, under aggregation, the tail behavior of non-Gaussian models is expected to become non-dominant, whereas in the corresponding Gaussian case, the tail behavior may remain dominant and a better model for the behavior of the non-Gaussian case will eventually be provided by the Gaussian model.

Assuming a sufficiently large network whose performance is determined solely by the carried traffic's first and second order characteristics, namely, mean, variance and autocovariance. With a suitable choice of server speed, a model handling k times as much traffic (assuming aggregation of independent sources of traffic) is similar to the original system in the sense that the buffer content distribution can be obtained by rescaling the original curve. For example, if 100 identical traffic streams are multiplexed together, then at the same time, the capacity of the server will be increased by a factor of 100.

Let $\{X_n\}_{n\in\mathbb{Z}}$ denotes an original traffic stream with mean μ, and let the original service rate be τ. Hence, the net mean is $m = \mu - \tau$, and suppose that $\mathrm{Var}\left(\sum_{j=1}^{n} X_j\right) = V(n)$, i.e., the variance-time curve is arbitrary. Now consider traffic $\left\{Y_n^{(k)}\right\}$, which is obtained by aggregating together k independent traffic

streams statistically identical to $\{X_n\}$. For this traffic, a faster service time is used, i.e., $\tau_k = \mu k - m\sqrt{k}$.

Needless to say, the performance margin for this aggregate model, which is the extra server capacity above the rate at which traffic arrives, is $|m|\sqrt{k}$, i.e., $\sqrt{k}$ times the performance margin for a single traffic stream.

The variance-time function for $\{Y_n^{(k)}\}$ is $V_Y(n) = kV(n)$. Now rescale the aggregate model by measuring work in units $\sqrt{k}$ times as large as the original units. The variance-time curve of $\{Y_n^{(k)}\}$ in these units is therefore $V(n)$, exactly the same as the original traffic. The mean net input into this system using the chosen units is also the same as the original, as follows:

$$\frac{kE(X) - \tau_k}{\sqrt{k}} = m .$$

In other words, when expressed in units proportional to $\sqrt{k}$, the stationary queuing distribution of the system is the same as the original system.

As shown, the performance margin steadily reduces as a proportion of the total system capacity. This is a concrete representation of the multiplexing gain that can be expected as networks become larger and traffic is aggregated.

2.4.3.2 Gaussian Traffic Models

Numerous papers have been written on Gaussian traffic modeling. The fractional brownian motion traffic model, first introduced by Norros [Nor94], captured the second order properties of self-similar traffic processes over multiple time-scales. Fonseca et al. [FMN99, FNM00] proposed a fractal brownian motion envelope process to characterize long range dependent traffic source. Analytical results for self-similar Gaussian queue based on large deviation theory are considered in [MV96], while [CM84, Rei84]

provided an analysis of correlated Gaussian queues under heavy traffic conditions.

As shown by [GT99a, GT99b], the application of a Gaussian model as an admission control algorithm is plausible. The advantage of using this model is the achievable increases in the statistical multiplexing gain factor, and hence the effect is the higher utilization of valuable network resources.

In the Gaussian approximation method, each connection is characterized by its average rate m_j and standard deviation σ_j. Let X be a random variable denoting the aggregate rate for n multiplexed connections. The problem is to determine the equivalent capacity c_g required by n connections such that the probability of the instantaneous aggregate rate exceeding c_g is less than a given value ε, as shown below:

$$\Pr\{X > c_g\} \leq \varepsilon . \tag{2.4}$$

In references [AS94, CS98, GG92, Sai92, SRLL95, SS91], c_g is estimated by assuming the aggregate rate distribution is Gaussian. For Guerin et al. [GAN91] the stationary bit rate approach using Gaussian approximation is computed as shown:

$$c_g \approx m + \alpha' \sigma , \tag{2.5}$$

where m and σ respectively denote the mean and standard deviation of the aggregate, i.e., superposed, traffic. The parameter α is the inverse of the Gaussian distribution with one possible value given by:

$$\alpha' = \sqrt{-2\ln(\varepsilon) - \ln(2\pi)} .$$

A new connection is accepted if $c_g \leq C$; otherwise it is rejected. Despite its simplicity, the Gaussian approximation scheme has disadvantages. Firstly, this approximation tracks the actual aggregate bandwidth requirement reasonably well only when the stationary distributions of individual

connections with similar parameters and long burst period are themselves Gaussian, and if a large number of these connections are multiplexed together. Secondly, the scheme treats all connections as if they share the same loss rate requirement. In reality, every connection's requirement may differ significantly. Thirdly, the scheme may overestimate the aggregate bandwidth requirement when connections have short bursts because these burst are normally smoothed out by the output buffer.

2.4.4 Measurement-based CAC

Admission control schemes discussed thus far are based on analytical modeling of the behavior of both the traffic sources and queuing structure. In reality, there are many types of traffic sources, each behaving quite differently. Hence, it is nearly impossible to model all of them accurately. Analytical models used to estimate the bandwidth required for one given class of traffic sources could over-estimate or under-estimate the bandwidth requirements for some other classes [Rob97]. Furthermore, some sources may not fully utilize their traffic descriptors. In light of this, admission control schemes based fully or partially on real-time resource measurements will estimate the resource usage more accurately. Basically, these schemes attempt to predict whether the QoS objectives can be achieved if a new connection is admitted, based on the real-time measurements of certain resources.

It is widely recognized that the maximum number of heavy-tailed flows that can be admitted into a network link whilst meeting QoS targets, can be much lower than in the case of markovian flows. Furthermore, the superposition of heavy-tailed flows shows long-range dependence (i.e., self-similarity, see [LTWW93, PF95, WTE96] and references therein), which has a detrimental impact on network performance. Bianchi et al. [BMN02] have shown through empirical studies that long-range dependence is significantly reduced when traffic is controlled by a measurement-based admission control algorithm. Their results appear to suggest that measurement-based admission control is a value added tool that improves performance in the presence of self-similar

traffic, rather than a mere approximation counterpart to the traditional parameter-based admission control schemes.

The authors of [GK97, GKK95] have derived a Chernoff bound measurement-based admissions control procedure that is based on the equivalent bandwidth model. The Chernoff bound gives a measurement-based admissions control procedure based on measurements of the aggregate arrival rate and on a single burstiness parameter for all of the admitted connections. Admission decisions are made based upon the current load being less than a pre-calculated threshold. The authors then developed a decision-theoretic scheme that does not require previous knowledge of the burstiness parameter value. Bayesian decision theory provides the framework for the choice of thresholds, and these are computed based on the assumptions that real-time sources can be accurately modeled through a set of finite source models. According to [JDSZ95], this approach is not applicable to a large and heterogeneous application base. On the other hand, the method of [GK97, GKK95] has the benefit that it provides a measurement-based scheme that is simple to implement. In reference [Flo96], the author has proposed the computation of the equivalent bandwidth based on the Hoeffding bound. This method gives a measurement-based admissions control procedure based on the policed peak rates of the admitted connections as well as measurements of the aggregate arrival rate.

In [DJM97], the required aggregate equivalent bandwidth is estimated using both the traffic descriptors and load measurements. The authors used a linear Kalman filter to optimize the estimates. In [CLLRTM97, CLMLRT97, Duf00, LRTMCL98, Rei01], large deviation theory is used to model the arrival traffic. This theory is also used by the authors of [GKT97] to model the performance of multiple time-scale traffic for use on Renegotiated Constant Bit Rate (RCBR) services. Hyman et al. [HLP93] have considered measurement-based admission decisions based on the assumption that all sources can be modeled by a finite set of source models.

Hiramatsu is one of the earliest researchers to propose using artificial neural network to solve ATM CAC problems, and it is based on the complex relationship between the offered traffic and the QoS requirements during stochastic multiplexing [Hir90]. Other neural network based algorithms are proposed in [KS93, LC99, YHS96], while [BLCT97, CC96, DD95, UH97] have proposed fuzzy-logic based algorithms instead. In [CCL99], Cheng et al. have proposed a neural fuzzy admission control algorithm that combines the best of both neural and fuzzy methods. Hah et al. [HTY97] have considered a neural network based CAC mechanism that estimates the cell delay and cell loss experienced by each class of traffic in a heterogeneous stream. In addition, Hah and Yuang have proposed an alternative CAC scheme using a delay-and-loss-based algorithm called quasi-linear dual class correlation, which conservatively estimates the cell delay and cell loss per traffic class using pre-computed vectors derived from the results of three dual arrival queuing models [HY97].

In [LCH95, LH93, LH97], the spectrum analysis method is used to characterize traffic that may include different frequencies. In [QK01, QK98], the authors have used measurement-based maximal rate envelopes of the aggregate traffic to capture its temporal correlation as well as the available statistical multiplexing gain. The authors of [LLD96] have considered a real-time computation algorithm based on the bufferless fluid flow model. Belenki [Bel02] have considered a heuristic-based per-hop admission algorithm that adapts the average rate of admission to the measured system performance. Siwko et al. [SR01] have applied admission control to satellite communication systems like the Low Earth Orbit Satellite (LEOS) systems, while Chou et al. [CS02] have applied it to wireless cellular networks.

Saito [Sai92, SS91] has proposed a dynamic scheme that uses the measured number of cells arrived during a fixed interval and the traffic descriptors to estimate the cell loss probability. Furthermore, Saito highlighted that an admission procedure based on real-time measurements of the arrival rates at

the gateway can tolerate possible policing errors at the network edge. Another cell loss approximation approach is a method by Zukerman and Tse [ZT97] for finite buffer queue. This approach relies on aggregate statistics and cell loss measurements for ongoing traffic. It assumes that new connections are transmitting at their peak rate during certain warm-up duration so as to be conservative in the admission decisions. Furthermore, this duration is adaptive to network conditions. This approximation approach is derived from the Reich's approach by Benes [Ben63], which is for infinite buffer queue. In one of our CAC frameworks, the [ZT97] method is used to approximate the cell loss rate. In section 3.3.1, this method is presented in greater details.

In addition to the above measurement-based admission control algorithms, several other approaches have been developed using different algorithms; and these can be found in references [CDS02, CL97, CLG95, KS00, MH02, RZKJ01, SCY98, SL02, WCKG94].

Currently, Internet traffic consists mainly of Transmission Control Protocol (TCP) flows. These are normally connection-oriented in nature, and have elastic/loose resource requirements. Lately, more streaming multimedia flows such as Real Time Protocol are transmitted in networks. Although these streaming protocols are often not explicitly connection-oriented in nature at the transport layer, they may be considered so at a higher, session layer [MPCC00]. In [MR99a], the authors have advocated the use of admission control to limit the number of TCP flows on a network link, so as to ensure that each has a minimal acceptable throughput. They have demonstrated that in the absence of such a control, the ineffective traffic due to the retransmission of lost packets constitutes a significant overhead and can even lead to congestion collapse in certain configurations.

The Internet Engineering Task Force (IETF) defines the Integrated Services (IntServ) [SPG97, Wro97] architecture for the Internet Protocol (IP) networks to provide QoS-oriented services. Flows must request service from the network and are accepted or rejected depending on the level of available

resources. For continuous-media applications, a new service is introduced to guarantee QoS on the Internet. Several papers in the area of delay bound calculations for queuing networks with regulated traffic have laid the basis for the guaranteed QoS service [Bou98, Cha00, Cru91a, Cru91b, FV90, GGPS94, PG94, RRR02, ZF94b]. However, as shown by Zhang et al. [ZF94a], when the traffic flows are bursty, the guaranteed QoS service will inevitably result in low utilization. Another type of QoS service is predictive QoS service. Various measurement-based admission control algorithms have been proposed to provide such relaxed real-time services [CSZ92, DKPS95, GTKT96, JDSZ97, JSD97]. Floyd [Flo96] and its improved variants [BS97, BS98], considered the Hoeffding bound.

While such architectures provide adequate QoS, they have significant scalability problems. Differentiated Services (DiffServ) [BBCDW98, NELC01] is another approach to provide QoS on an IP network. It requires no per-flow admission control or signaling, hence it does not suffer from scalability issue. Barry et al. [BCV01] have investigated DiffServ on a wireless packet networks.

Recently, several papers [BBCFP02, BCP00, BKSSZ00, CK00, CKK01, EKR00, GK99, JSSWZ02, KKZ00] have proposed a novel approach of using Endpoint Admission Control. It is an attempt to combine DiffServ's excellent scalability with IntServ's excellent QoS. These schemes aim to provide QoS, i.e., packet loss rate, to real-time flows within the IntServ networks. Typically, the end host will probe the network by sending probe packets at the data rate it would like to reserve, and then measuring the level of packet losses. If the loss level is below some threshold value, the host will admit the flow; otherwise the flow will be rejected.

In [RKT02], the authors have investigated the performance of cooperative congestion control approach. They presented techniques based on loss or delay observations at end hosts to infer if two flows are congested at the same

network resources. They argued that their techniques are also applicable to multicast flows.

We believe the fundamentals of our proposed connection admission control frameworks (details in chapter 3) are also applicable for use in IP related admission controller.

2.5 Conclusion

In this chapter, we have introduced the concept of admission control and how it is used as a network traffic control tool to ensure QoS requirements are met for established connections.

We have also presented a comprehensive literature survey of the state-of-the-art research in the areas of admission control, and highlighted research issues that are still unresolved.

3 Connection Admission Control

3.1 Introduction

We consider a store-and-forward (packet or cell switched) connection oriented multiservice network that is based on the concept of Asynchronous Transfer Mode (ATM). An ATM connection traverses a set of switching nodes in the network. Even within a switching node, a connection may traverse a number of queuing points. To set-up a connection on such a path, resources must be reserved at each queuing point to guarantee the contracted Quality of Service (QoS). The set of procedures that determine admissibility of a connection in a switch is commonly termed Connection Admission Control (CAC).

We believe that a CAC approach should be practical from the viewpoints of both network provider and users. In addition, the CAC algorithm should be simple to implement and only needs minimum traffic information as inputs from the user. Furthermore, the algorithm should be robust to drastic changes in traffic load, for example, sudden surge in arrival rates.

In this chapter, we present two CAC frameworks that are named Model and Histogram based frameworks. These CAC frameworks are formulated based on our philosophy on practical admission control methodologies. These frameworks contain CAC approaches that share common traits in their admission control algorithms. Some of the Measurement-based CAC

(MBCAC) approaches proposed here are customizable to the network provider's traffic control requirements.

Overall, the Model and Histogram based CAC frameworks can be used to test various traffic control strategies, and in particular, to focus on the simplicity versus efficiency tradeoff issues. That is, if significant complexity does not improve efficiency substantially, simpler admission control methods should be used instead.

The CAC and MBCAC approaches presented here aims to provide QoS for the aggregate flow, hence no QoS is offered at the per-flow level. For quality assurance, the QoS parameter – Cell Loss Ratio (CLR) is considered here.

The remainder of this chapter is as follows:

- Section 3.2 presents the Model-based CAC framework that comprises CAC and MBCAC approaches that use traffic models to aid in making admission decisions. In addition to making certain traffic behavior/modeling assumptions, these schemes use either a-priori or measured traffic parameters to help determine the amount of bandwidth required by the aggregate traffic stream. Some of the MBCAC approaches include an Adaptive Feedback Control Mechanism (AFCM) that adapts these approaches to varying traffic load conditions. By configuring certain AFCM parameters, the network providers can customize these approaches according to their traffic control requirements.

- Section 3.3 presents the Histogram-based framework that comprises MBCAC approaches that use traffic statistics derived from past traffic load to help determine the amount of bandwidth required by the aggregate traffic stream. At the heart of these MBCAC approaches are: (1) a unique procedure of evaluating 'Available bandwidth' values based on real-time measurements of the arrival traffic and the use of

three fundamental CAC algorithms; (2) updates of past traffic records whenever a connection departs from the network; (3) the use of a technique to increase link utilization through easing a constraint that is imposed on the algorithm that computes the service bandwidth values; and (4) the use of the AFCM to provide an additional layer of control to ensure maximum link utilization whilst meeting the QoS requirement.

- Section 3.4 describes in detail the adaptive feedback controller – AFCM, used by both CAC frameworks.

3.2 Model-based CAC Framework

Under ATM, when a user wants to establish a connection, a set of traffic parameters representing the statistical behavior of that source traffic is specified [ATM99, ITU00d]. In this thesis, we term a CAC approach based only on traffic parameters, plus certain traffic behavior (modeling) assumptions, a model-based CAC. This type of CAC may not be efficient because traffic is unpredictable, and users may not know how to describe their traffic sources optimally.

Figure 3-1 illustrates the admission decision process for three model-based CAC approaches and four measurement-based counterparts within the model-based framework. CAC approaches based on the traditional Gaussian and Effective Bandwidth models are denoted by GA and EB respectively. The enhanced Gaussian CAC approach is denoted by eGA, while the four measurement-based counterparts of the Gaussian and enhanced Gaussian models are denoted with a prefix ‘m-’ followed by their respective acronym.

The model-based approaches require a-priori statistics of the traffic source before the aggregate service bandwidth can be computed. Traffic parameters such as mean, variance and performance margin of a traffic source must be known before the connection set-up phase. The eGA approach requires an

additional traffic information in the form of a performance margin look-up table.

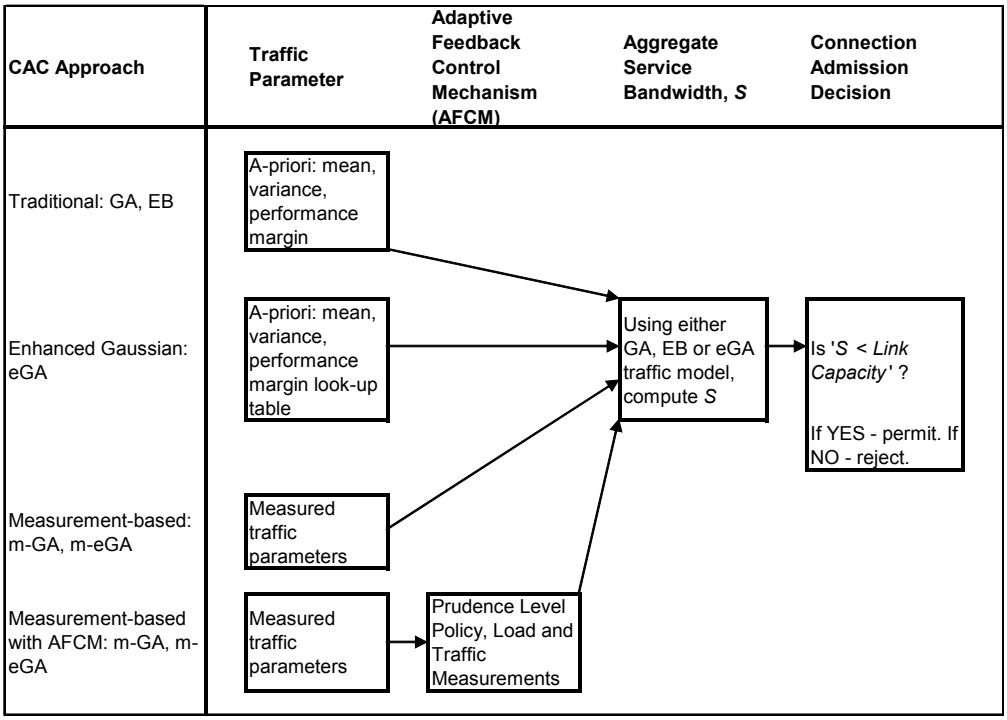


Figure 3-1. Model-based CAC Framework.

For the MBCAC approaches, a number of traffic parameters are measured in real-time. Hence, the network provider need not depend on the user providing accurate traffic information. However, because the arriving traffic load may exhibit non-stationary behavior and be unpredictable, additional control is needed on top of the measured quantities. As shown in the figure, an Adaptive Feedback Control Mechanism (AFCM) is used for this purpose.

The admission decision process is as follows. To compute an aggregate service bandwidth value, various a-priori or measured traffic parameters are used. Basically, the approaches estimate the amount of link bandwidth that may be consumed by both the new connection and the established connections. The new connection will be admitted if the aggregate service bandwidth value is smaller than the link capacity.

3.2.1 Traditional Approaches

In this section, we present two traditional model-based CAC approaches, i.e., Gaussian and Effective Bandwidth, which have a-priori traffic knowledge for all traffic sources that will be used. Using this traffic information together with a value equal to the present total number of established connections, the CAC schemes will calculate the equivalent bandwidth needed to service all the established connections (including the new connection) on the link whilst still meeting the QoS requirement. Typically, the computed equivalent bandwidth values will be used to aid in making connection admission decisions. A new connection will only be admitted if the amount of service bandwidth *S*, required by the total number of established connections including the new connection, is less than a link service rate *C*.

3.2.1.1 Gaussian Model-based CAC Approach

As mentioned earlier in section 2.4.2 of this thesis, the traditional effective bandwidth approach cannot benefit from the statistical multiplexing of other connections because of its independence property. Hence, the approach is unable to achieve optimal performance efficiency.

The Gaussian traffic model proposed by Addie [Add98] does consider the statistical multiplexing factor into its equivalent bandwidth calculation. The advantage of using this model is the achievable increases in the multiplexing gain factor. The effect of these increases is the higher utilization of valuable network resources. Furthermore, this traffic model has a sufficient degree of realism such that it can be used to gain valuable insight into how to efficiently design and operate a broadband network.

The Gaussian traffic model can be used as a model for the superposition of a variety of processes such as the autoregressive model of [MASKR88] and the Orenstein-Uhlenbeck process of [Sim91]. The main benefit of the Gaussian model is that 'deep' within the network where many connections are multiplexed together such that the central limit theorem applies, the

multiplexed traffic can be modeled by a Gaussian process. In other words, as a network becomes increasingly larger in size and carry traffic from more independent connections, the unfinished work distribution will weakly converge to a Gaussian model. This Gaussian model has three importance characteristics: (1) it allows for any short range dependent or long range dependent autocovariance function, (2) it is amenable to queuing analysis, and (3) it is closed under superposition, i.e., the sum of two or more Gaussian traffic processes is still a Gaussian process.

Let time be divided into fixed length intervals, the length of which may be chosen such that occasional traffic bursts can be captured. The following derivations and notations are for a particular *i*-th time interval. The type of traffic source used by each connection is denoted by *j*. Let b_j denotes a discrete random variable representing the number of established connections of type *j* traffic, and let u_j denotes the total number of new type *j* traffic connection requests. Let n_j denotes the total number of new (if applicable) and established connections of type *j* traffic, and it is obtained by:

$$n_j = b_j + y_j, \tag{3.1}$$

where y_j is used during bandwidth calculation as a 'counter' to represent the number of new connections that can be admitted, i.e.,

$$y_j = \begin{cases} 0, & u_j = 0; \\ 1,...,u_j, & u_j > 0. \end{cases}$$

For the Gaussian traffic model-based CAC (GA) approach, we compute a bandwidth S_{GA} required to service both the new and the established connections with a shared buffer of size *l*. Let μ_j denotes an a-priori constant equal to the mean amount of work by a single type *j* connection, and let $d_{j,l,QoS}(1)$ denotes the minimum bandwidth required to serve one connection of type *j* traffic in a link with buffer size of *l*, whilst meeting the *QoS* requirement

which is a certain desired CLR value. Assuming the traffic follows a Gaussian process, S_{GA} can be evaluated by:

$$S_{GA} = \sum_j S(y_j), \tag{3.2}$$

where,

$$S(y_j) = \mu_j n_j + m_{j,l}\sqrt{n_j}\ ; \tag{3.3}$$

and,

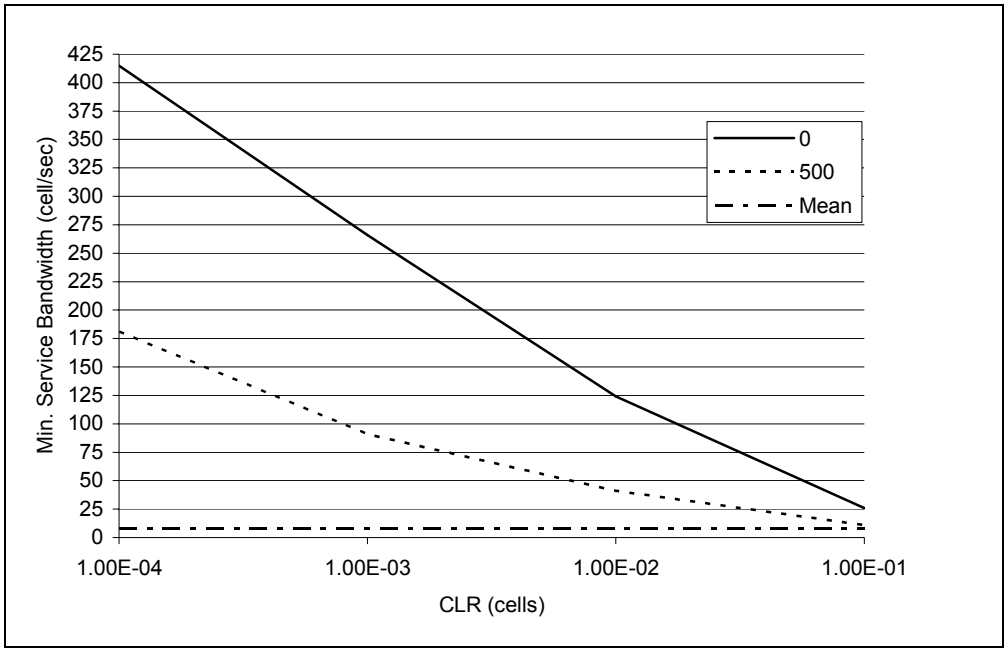
$$m_{j,l} = d_{j,l,QoS}(1) - \mu_j\ , \tag{3.4}$$

is referred to as a *performance margin* for a single connection of type *j* traffic. Basically, it stipulates the additional link capacity above the mean of a type *j* traffic flow. With this additional bandwidth allocated to service occasional bursts of arrival traffic from a connection, the CAC scheme will be able to provide the desired level of QoS for all connections. As shown in Figure 3-2, the $m_{j,l}$ value decreases with less stringent QoS requirements. The figure plots different values of $d_{j,l,QoS}(1)$ based on:

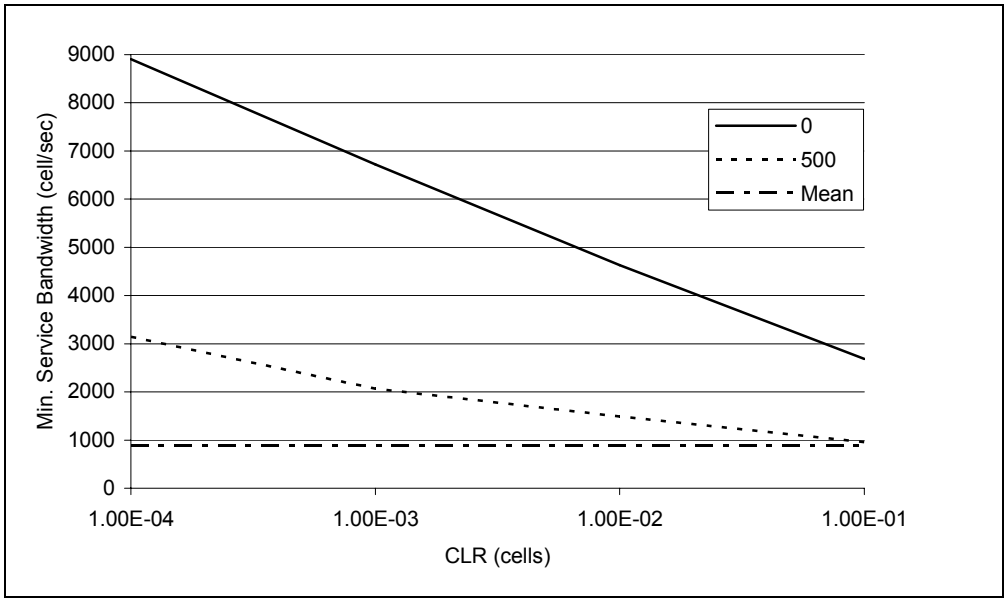
- *j* – Two types of traffic, i.e., Network Data (section 4.3.2.1) and Video (section 4.3.2.2). See the respective sections for more details about these traffic streams.
- *l* – Two SSQ buffer sizes, i.e., 0 and 500 cells.
- *QoS* – Four CLR requirements, i.e., 1e-4, 1e-3, 1e-2, and 1e-1 cells.

In (3.3), as *n* increases, the performance margin increases at a much lower rate than μn, and hence the equation shows that significant multiplexing gain can be achieved with large number of connections (see section 2.4.3.1 for more details). However, this formula is valid only in a link with a certain level of traffic aggregation such that the aggregate traffic stream exhibits Gaussian

behavior. To illustrate how this equation is used in the admission decision process, a flowchart is provided in Figure 3-3.



(a) Network Data traffic.



(b) Video traffic.

Figure 3-2. Minimum service bandwidth required by one active connection in order to meet the desired QoS requirements.

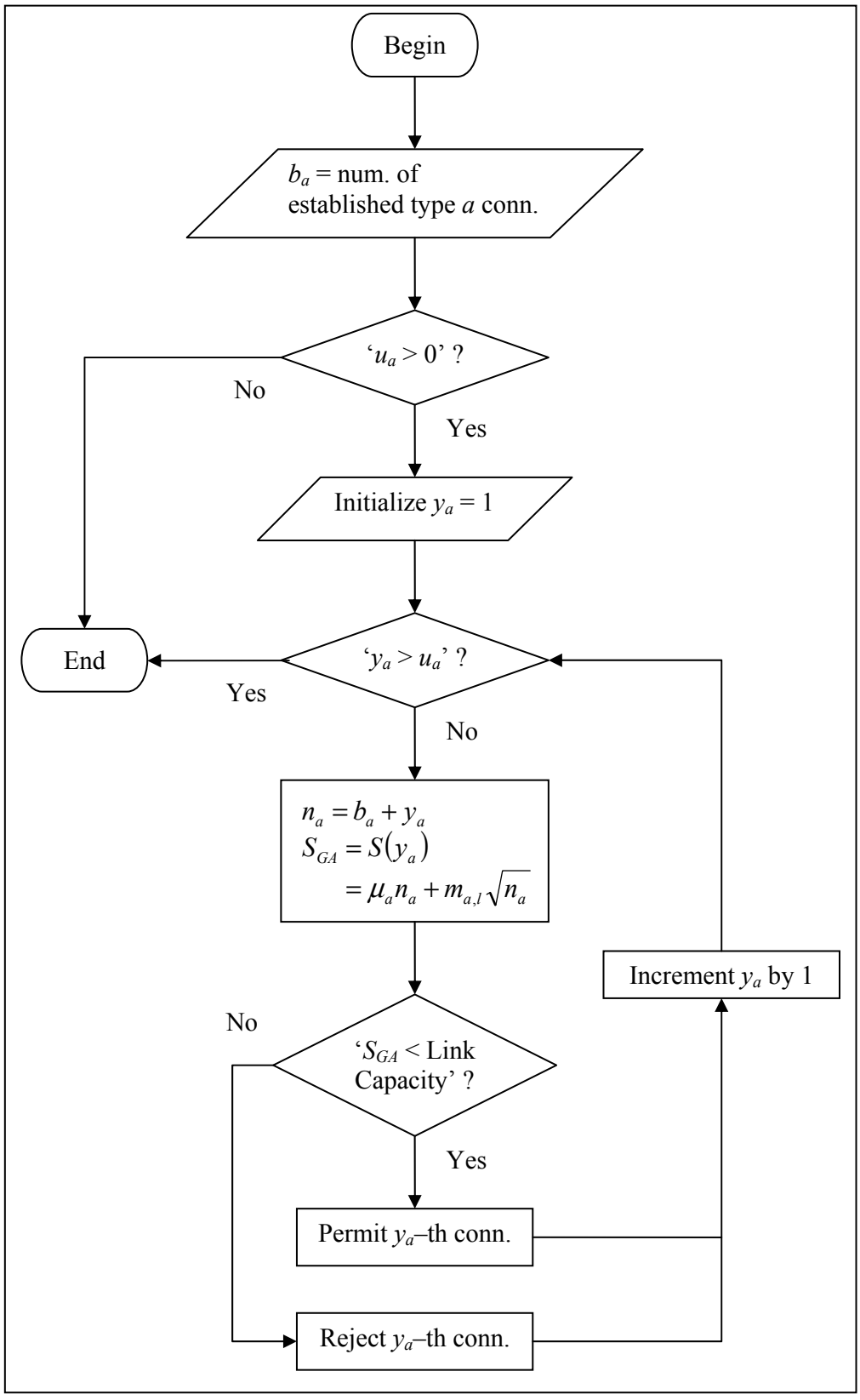


Figure 3-3. Admission decision process for the GA approach using homogeneous traffic streams scenario.

3.2.1.2 Effective Bandwidth Model-based CAC Approach

The Effective Bandwidth traffic model-based CAC (EB) approach is based on the concept that the effective bandwidth of the aggregate traffic stream is equal to the sum of the effective bandwidth of the individual traffic streams. Hence, the overall bandwidth required to service both the new and the established connections with a shared buffer of size *l* is:

$$S_{EB} = \sum_j S(y_j), \tag{3.5}$$

where,

$$S(y_j) = n_j(\mu_j + m_{j,l}). \tag{3.6}$$

Clearly, by not considering the presence of other neighboring connections in the link, the model does not take advantage of multiplexing connections together.

3.2.2 Enhanced Gaussian Model-based CAC Approach

As mentioned in section 3.2.1.1, when more and more traffic is aggregated together, the multiplexed traffic weakly converges to a Gaussian process. This convergence occurs in the heavy traffic case, where the system size is large and the traffic is highly aggregated. For such a scenario, the application of a Gaussian traffic model in a CAC approach is plausible. However, if the number of connections is not large and hence the traffic is not highly aggregated, is the Gaussian model-based CAC approach still applicable and effective in providing efficient admission control operations?

In this section, we aim to address to a certain limit, this issue in order to realize a realistic CAC approach that takes into consideration the statistical behavior of both lightly aggregated traffic, i.e., non-Gaussian, and highly aggregated traffic, i.e., Gaussian.

Here, we present a CAC approach that does consider the level of traffic aggregation in its service bandwidth calculation. This scheme, which is an enhanced version of the Gaussian model-based CAC approach, considers both the total number of established connections and its multiplexing gain effect, in order to better reflect the statistical behavior of the aggregate traffic stream. We call this scheme the enhanced Gaussian model-based CAC (eGA) approach, and it uses a threshold to approximate the aggregate traffic behavior, i.e., non-Gaussian and Gaussian regions. The threshold is basically an estimated value equal to the number of simultaneous connections required for an aggregate traffic stream to exhibit Gaussian behavior. With this knowledge, the enhanced Gaussian model-based CAC approach can be fine-tuned to approximate realistic traffic behavior.

An alternative representation of the a-priori performance margin expressed in (3.4) is:

$$m_{j,l} = \begin{cases} q_{j,l}(n_j)\sigma_j, & \text{for Gaussian related schemes;} \\ q_{j,l}(1)\sigma_j^2, & \text{for traditional Effective Bandwidth scheme;} \end{cases} \tag{3.7}$$

where $q_{j,l}(n_j)$ denotes a real number value unique to n_j connections in a link with shared buffer of size l, and σ_j^2 denotes the variance amount of work for a single type j connection. Basically, the multiple factor $q(n)$ qualifies the amount of extra bandwidth (in term of several standard deviations) needed to ensure the desired level of QoS.

For the GA approach, the aggregate traffic stream is assumed to be instantaneously Gaussian regardless of the level of aggregation. That is to say that $q(n) = q(1)$ for all n number of established connections. To derive the multiple factor for one connection, $q(1)$ is obtained as shown:

$$q_{j,l}(1) = \frac{d_{j,l,QoS}(1) - \mu_j}{\sigma_j}. \tag{3.8}$$

However, this 'one-rule-applies-to-all-scenarios' assumption is flawed whenever the aggregate traffic stream is not highly aggregated and n is not large, resulting in the aggregate traffic stream exhibiting non-Gaussian behavior. Therefore the GA approach is not applicable to this aggregate traffic stream since the approach will not estimate the service bandwidth S accurately. This implies that for '$n < r$' case, where r is the estimated number of simultaneous connections required for the aggregate traffic stream to exhibit Gaussian behavior, $q(n)$ cannot be equal to $q(1)$ for all '$1 < n < r$' number of established connections since the aggregate traffic stream is not Gaussian.

The eGA approach, which considers the level of traffic aggregation, rectifies this error by using a look-up table that contains the multiple factors, $q_{j,l}(n_j)$. For n number of type j connections, $q(n)$ is derived as follows:

$$q_{j,l}(n_j) = \frac{d_{j,l,QoS}(n_j) - \mu_j n_j}{\sigma_j \sqrt{n_j}}. \tag{3.9}$$

Whenever a new connection request arrives, the bandwidth required to service both the new connection and the established connections are computed as shown in equation (3.2). The new connection will only be admitted if the amount of service bandwidth S, is less than a link service rate C. To compute the service bandwidth S, a-priori traffic parameters – $q(n)$, μ and σ^2 are required.

3.2.3 Measurement-based Counterparts

The CAC procedures described until now are based on the analytical modeling of the traffic behavior. Typically, these approaches require a-prior traffic information in terms of traffic parameters based on a deterministic or stochastic model. However, there is a huge range of possible traffic sources in reality, and it is impossible to model all of them accurately. In addition, traffic models based solely on the user-declared traffic descriptor may compromise

efficiency when users over-estimate the required network resources so as to ensure the integrity of their data transmissions. The overall effect is the inefficient usage of the link.

Hence, CAC procedures based on online traffic measurements are in a better position of predicting the usage of the network resources more accurately. Measurement-based CAC (MBCAC) schemes avoid the aforesaid problem by re-allocating the task of specifying the traffic flows from the users to the network provider. It aims to ease users' responsibilities by relying less on user-declared traffic descriptor, and instead more on measured quantities relevant to making admission decisions. These measured quantities refer to the traffic statistics measured in real-time from the aggregate traffic stream. The QoS experienced by the multiplexed flows will depend on their aggregate behavior, since aggregate traffic statistics are easier to determine than the traffic statistics obtained from an individual flow. Moreover, since there is a lesser reliance on the user-declared traffic descriptor, an overly conservative user-declared traffic specification will not result in an over-allocation of valuable network resources for the entire duration of the connection.

We incorporate two measurement-based alternatives for the two model-based CAC approaches mentioned earlier, i.e., the traditional Gaussian (GA) and the enhanced Gaussian (eGA). We denote these approaches with a prefix 'm-', i.e., m-GA and m-eGA.

The first alternative scheme simply does online measurements for the mean amount of arrival work μ, and also estimates the performance margin m of (3.7). However for the second alternative scheme, it goes one step further by including an Adaptive Feedback Control Mechanism (AFCM) that adjusts the level of contributions by the performance margin m towards the computed service bandwidth S value, based on the measured traffic load conditions.

3.2.3.1 First Alternative Scheme

For the MBCAC approaches within this first alternative scheme, other than the connection's traffic type and the performance margin's multiple factor *q(n)* value(s), no other a-priori traffic information is required.

Hence, for the m-GA and m-eGA approaches, the service bandwidth $S(y_j)$ (per type *j* traffic) is given by:

$$S(y_j) = \hat{\mu} n_j + \hat{m}_{j,l} \sqrt{n_j} \; ; \tag{3.10}$$

and applying (3.7), $S(y_j)$ can now be expressed as:

$$S(y_j) = \begin{cases} \hat{\mu} n_j + q_{j,l}(1) \hat{\sigma} \sqrt{n_j}, & \text{for traditional Gaussian scheme;} \\ \hat{\mu} n_j + q_{j,l}(n_j) \hat{\sigma} \sqrt{n_j}, & \text{for enhanced Gaussian scheme;} \end{cases} \tag{3.11}$$

where $\hat{\mu}$ and $\hat{\sigma}^2$ are respectively the measured mean and variance work for a single connection. These traffic statistics are derived from the online measurements of the aggregate flow.

Taking an m-GA approach with aggregate flow made up of heterogeneous connections, i.e., traffic types *a* and *b*, as an example, the overall service bandwidth $S_{m\text{-}GA}$ is given by:

$$\begin{aligned} S_{m-GA} &= \sum\nolimits_j S(y_j) \\ &= \left(\hat{\mu} n_a + q_{a,l}(1) \hat{\sigma} \sqrt{n_a}\right) + \left(\hat{\mu} n_b + q_{b,l}(1) \hat{\sigma} \sqrt{n_b}\right). \end{aligned}$$

While for the m-eGA approach, the *q(n)* is obtained from a look-up table specific to the traffic type.

3.2.3.2 Second Alternative Scheme

For the MBCAC approaches within this second alternative scheme, an AFCM (see section 3.4 for more details) is used to adjust the computed service bandwidth *S* values based on the measured traffic load conditions. Preferably,

this control should be adaptive such that during periods of link congestion or cell loss, the adjusted $S(y_j)$ values will increase and be almost equal to the amount computed in (3.11). On the other hand, during periods of no congestion or cell loss, the adjusted $S(y_j)$ values will be almost equal to *only* the aggregate mean work submitted by both the new and established connections, as shown below:

$$S(y_j) = \hat{\mu} n_j . \tag{3.12}$$

To achieve this adaptive-ness, both MBCAC approaches use an adaptive prudence level factor p. Depending on the traffic load, p (a real number value) moves between 0 and 1; and it is applied as follows:

$$S(y_j) = \hat{\mu} n_j + \hat{m}_{j,l} \sqrt{n_j} (1 - p). \tag{3.13}$$

To gauge the severity of the link congestion or the cell loss, the arriving workload or the amount of lost cell is compared against a threshold. When the measured traffic load or cell loss is higher than a certain threshold, p will deviate towards a 0 value, which results in the conservative computation of the $S(y_j)$ values according to (3.11). However, when the traffic load or the amount of lost cell is lower than the same threshold value, p will deviate towards a 1 value, which results in the liberal computation of the $S(y_j)$ values according to (3.12).

In section 3.4, the AFCM is explained in greater detail.

3.3 Histogram-based CAC Framework

The previous section 3.2 described: (1) the model-based a-priori CAC approaches that do not require any online traffic measurement, and (2) their measurement-based counterparts.

In this section, a variety of MBCAC approaches within the histogram-based framework are presented. The admission decisions of these MBCAC approaches are based on the past traffic work submitted by all established connections into a Single Server Queue (SSQ). Basically, an MBCAC approach relies on real-time traffic measurements and these measurements are then used to aid in the admission decision process. In principle, if relevant traffic information is measured from the aggregate flow, then with such intimate traffic knowledge the MBCAC approach may be in a better position to predict near-future aggregate traffic flow behavior and thus achieve improved traffic control efficiency. Another advantage is that if the MBCAC approach requires user-declared traffic descriptor, then this requirement can be made easier by requiring only easy-to-measure traffic parameters, for example peak rate, to be provided.

This section describes a CAC framework which includes: traffic prediction model, measurement requirements, as well as an additional adaptive feedback controller, i.e., the Adaptive Feedback Control Mechanism (AFCM), that protects the network when an MBCAC approach fails to meet QoS requirement in term of a desired level of CLR, either because it tries to be too aggressive (to save bandwidth), or when the traffic exhibits unpredictable behavior, or both. This is because the behavior of the arrival traffic may be non-stationary and very much unpredictable; hence in addition to the traffic prediction mechanism that assumes stationary traffic, there may be a need for an additional control layer for QoS assurance. When things go wrong and the traffic prediction mechanism fails, hence resulting in an increased risk of not

meeting QoS requirement, the AFCM must be prompted to remedy this situation. In section 3.4, more details on AFCM are provided.

Another element of this CAC framework is the novel procedure of *available bandwidth* evaluation at all network bottlenecks along the connection's end-to-end route. In general, there are three basic CAC algorithms [ITU00a, RMV96]:

- Peak Rate Allocation (PRA) method, which considers neither sharing of the link bandwidth nor of the buffer. This method is suitable for CBR or mildly variable traffic streams. It is also suitable for VBR traffic streams that have strict QoS requirements such that for any QoS violations, large penalties will be incurred. In all these cases, the PRA method is efficient.

- Rate Envelope Multiplexing (REM) method, which is based on the zero buffer approximation and assumes the sharing of the link bandwidth but not of the buffer. This approach is suitable for real-time streaming traffic since it ensures the traffic is subjected to minimal jitter and delay.

- Rate Sharing (RS) method, which considers sharing of the link bandwidth and the buffer.

To evaluate the service bandwidth, the PRA method simply sums up all connections' user-declared peak rates. However, for the REM and RS methods, the bandwidth required to service the established connections whilst ensuring the desired QoS, i.e., CLR level, can be provided, is evaluated using past traffic arrival statistics to gauge the aggregate traffic stream behavior. The overall *available bandwidth* is then derived based on the choice of the AFCM techniques and the instantaneous traffic load condition.

To derive traffic statistics from the past records of the arrival traffic, a collection of multiple time-scale traffic histogram databases are used. Each

traffic histogram database stores records of the amount of work that arrives within consecutive window with each window having a fixed time-frame, i.e., work arriving during a particular time-scale. Hence, the framework maintains traffic load records across multiple time-scales. Preferable, the choice of the time-scales is adequate to capture occasional traffic bursts in order to accurately derive traffic statistics on the past arrival traffic.

Another element of this CAC framework is the three traffic histogram update techniques, in varying complexities and storage requirements – from no update to complete updates, that will be used to alter, if applicable, all traffic histogram records whenever a connection departs from the network.

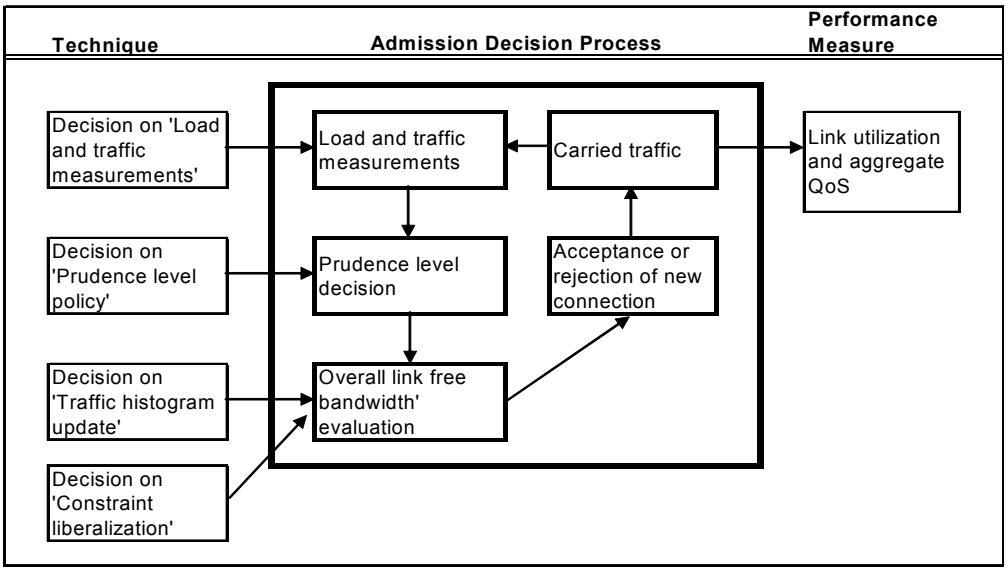


Figure 3-4. Histogram-based CAC Framework.

Figure 3-4 illustrates the admission decision process for the MBCAC approaches within the histogram-based framework. Each module contains a variety of techniques that can be used as part of an MBCAC approach's admission decision process. The CAC framework allows the network provider to 'mix and match' different techniques so as to create a custom-made MBCAC approach that will meet the desired traffic control requirements. Hence, many different MBCAC schemes can be formulated from this CAC framework.

From the figure, a chosen 'load and traffic measurements' technique together with a chosen 'prudence level policy' technique will adjust a weight factor within the 'prudence level decision' component. This factor is a real number between 0 and 1, and it moves between the conservative (PRA) and the bold (REM/RS) service bandwidth computation methods. When this adaptive factor is used in the 'overall link free bandwidth evaluation' component, an overall available bandwidth value is then derived. To further improve on this derived value, the 'traffic histogram update' and/or 'constraint liberalization' techniques are used. Next, using this overall available bandwidth value, connection admission decisions are then made. These decisions will affect the amount of traffic that is carried in the network, as reflected in the 'carried traffic' component. Moreover, it is this volume of carried traffic that will decide the measured performance quantities in the 'link utilization and aggregate QoS' component.

The following sections will define in detail:

- Section 3.3.1 – A model to approximate the cell loss that may be experienced by the aggregate flow.
- Section 3.3.2 – The 'Available bandwidth' evaluation technique using the three basic CAC algorithms.
- Section 3.3.3 – Different techniques to update the traffic histogram databases whenever a connection departs from the network.
- Section 3.3.4 – A technique to reduce/remove a CAC constraint.

Detail discussions on the 'prudence level policy' and the 'load and traffic measurements' techniques are in section 3.4. These techniques collectively form the adaptive feedback controller – AFCM, which is used by both model and histogram based CAC frameworks.

3.3.1 Cell Loss Approximation

Congestion theory is used to help estimate the amount of cells that may be lost in a network. Most of the theories rely on making particular statistical assumptions about the arrival traffic, such as negative exponential distribution or independent random variable characteristic. However, it is difficult to model all types of traffic that may be transmitted on a network, and hence it will be equally difficult to make statistical assumptions about them.

In this CAC framework, a virtual delay analysis by Reich [Rei58] to approximate the cell loss is used. The advantage of the Reich's approach, and its derivatives, is that it provides the cell loss estimates without any restrictive assumption on the statistical behavior of the input traffic (except stationary). Hence it can handle many types of traffic that classical traffic and queuing models cannot, for example, B-ISDN traffic, which is self-similar in nature and exhibits long range dependence [LTWW93].

In this section, we present: (1) the Reich's approach by Benes [Ben63] for infinite buffer queue, and (2) an approximation approach (derived from the Reich's approach) by Zukerman et al. [ZT97] for finite buffer queue. The former gives the exact waiting-time of the queued cells, while the latter approximates the cell loss probability.

Consider a multiplexer SSQ model with a buffer that is shared amongst all connections, and the departure process uses the First-In First-Out (FIFO) scheduling discipline. The service time for any cell is assumed to be the same; hence the duration to service one cell is chosen as the unit of time.

Let $W(t)$ denotes a virtual waiting-time function, defined as the time an ATM cell would have to wait for service if it arrives at time t. The stochastic process $W(\cdot)$ is expressed in terms of the instant of arrival t_w and the service time S_w of the w-th arriving cell. At t_w, $W(\cdot)$ increases discontinuously by an amount equal to S_w; otherwise $W(\cdot)$ decreases continuously at a rate equal to

negative one. If the value of *W(·)* is reduced to zero, it remains at zero until another instant of cell arrival.

In [Ben63], the author provides an elegant function *K(·)*, which describes the instant of cell arrival and service time simultaneously. The following paragraphs will explain how *W(·)* is formally defined in term of the *K(·)* function.

For $t \geq 0$ and between successive increases, the function is non-decreasing and it remains constant. The locations of the increases are at t_w, and its magnitudes are S_w. If *K(t)* is the work offered to the server in the interval [0,*t*), then *W(t)* can be thought of as the amount of work remaining in the queue and waiting to be served at time *t*. For simplicity, it is assumed that *W(0) = K(0)*. Hence, *W(·)* is formally defined in term of *K(·)*, by the following integral equation:

$$W(t) = K(t) - t + \int_0^t U[-W(u)]du, \qquad t \geq 0, \tag{3.14}$$

where *U(t)* is the unit step function, i.e.,

$$U(x) = \begin{cases} 1, & x \geq 0; \\ 0, & \text{otherwise.} \end{cases}$$

An explicit solution of (3.14) in terms of *K(·)* and the supremum function is as shown in the lemma below:

- If *K(x) – x* has a zero in (0, *t*),

$$W(t) = \sup_{o<x<t} \{K(t) - K(x) - t + x\}. \tag{3.15}$$

- If *K(x) – x* > 0 for x ∈ (0, *t*),

$$W(t) = K(t) - t. \tag{3.16}$$

From (3.15), the Reich's formula basically states that:

$$\text{unfinished work} = \sup_{o<x<t}\{\text{overload in } [x,t)\},$$

where $\{K(t)-K(x)-t+x\}$ denotes the arrival traffic load during the interval [x, t) that is in *excess* of the service rate.

An alternative expression for *W(t)*, the unfinished work in the queue waiting to be served at time t, during the period $t < t_0$ (i.e., the server has yet to be idle and the *K(t) – t* value has yet to become negative), is:

$$W(t) = K(t) - t - \min\{0 \quad , \quad \inf_{0<x<t}[K(x)-x]\}. \tag{3.17}$$

Even though (3.14) and (3.17) have expressed the relationship of *W(·)* in term of the load *K(·)*, another QoS measure – cell loss probability, is of more relevance to the CAC approaches considered here. In [ZT97], the authors have formulated this QoS measure into the *W(·)* function. This will be explained in the following paragraphs.

Consider a discrete-time SSQ with infinite buffer. Let time be divided into fix length intervals, the length of which may be chosen arbitrarily. Suppose A_i is the amount of work that arrived during the i-th interval, and s be the amount of work that can be served during an interval. Hence, between the period (0, x), the offered traffic load is $K(x) = \sum_{i=0}^{x} A_i$, while the total amount of traffic served is equal to $\sum_{i=0}^{x} s$.

Assuming the server utilization is not very high, then $\inf_{0<x<t}[K(x)-x]$ can be re-expressed to be:

$$\inf_{0<x<t}\left[\sum_{i=0}^{x}(A_i - s)\right] \le 0,$$

and (3.17) can now be expressed as:

$$W(t) = [K(t) - ts] - \min\left\{0 \quad , \quad \inf_{0<x<t}[K(x) - xs]\right\}$$
$$= \left[\sum_{i=0}^{t}(A_i - s)\right] - \inf_{0<x<t}\left[\sum_{i=0}^{x}(A_i - s)\right].$$

Hence, *W(t)* is given simply by:

$$W(t) = \max_{0<x<t}\left\{\sum_{i=x}^{t}(A_i - s)\right\}. \tag{3.18}$$

Let Z be the unfinished work distribution, and T be a pre-defined threshold. The estimated probability is:

$$P(Z > T) = P\{W(t) > T\}, \qquad t \in (0, \tau). \tag{3.19}$$

By substituting (3.18) into (3.19),

$$P(Z > T) = P\left\{\max_{0<x<t}\left[\sum_{i=x}^{t}(A_i - s)\right] > T\right\}, \qquad t \in (0, \tau).$$

Letting $w = \tau - x$, and $w > 1$,

$$P(Z > T) = \max_{\substack{0<x<\tau-w, \\ w\geq 1}}\left\{P\left[\sum_{i=x}^{x+w}(A_i - s) > T\right]\right\}.$$

Hence, the exact unfinished work (virtual delay) distribution is:

$$P(Z > T) = \max_{w\geq 1} P\left\{\sum_{i=1}^{w}(A_i - s) > T\right\}. \tag{3.20}$$

Reich's approach is exact for a SSQ with infinite buffer space and it requires optimization over all possible window size. Note that a window size is made up of w consecutive intervals, and it is exclusive. Hence, no two windows use the same w size.

In reality, buffer space is finite and the computing resources can only optimize limited number of windows. Hence, [ZT97] proposed an alternative (inequality) cell loss approximation based on the Reich's approach:

$$\max_{w\geq 1}\left\{\frac{\int_{sw+l}^{\infty} P\left(\sum_{i=1}^{w} A_i = x\right)\cdot(x-sw-l)dx}{E[A]w}\right\}\leq L, \qquad (3.21)$$

where l is the finite buffer space in an SSQ, and L is the desired CLR. The left-hand side of (3.21) represents a ratio between the average amount of work that must be lost and the total amount of work arrived, during a time interval in which the loss is maximized. Note that the average amount of work lost is calculated by considering the events when the total amount of work arrived is higher than the total amount of work that can be served sw, plus that can be buffered l. Obviously, this does not represent the total work lost and therefore it provides only a lower bound loss probability. The authors of [ZT97] treat the inequality in (3.21) as equality.

3.3.2 Available Bandwidth

In this histogram-based CAC framework, the smallest available bandwidth on an end-to-end connection is denoted by the Overall Link Free Bandwidth (OLFB) value. An end-to-end connection is composed of n links that make up a chosen end-to-end path. For all MBCAC schemes within this CAC framework, a new connection is admitted if its peak rate is less than the OLFB; otherwise, the new connection request is rejected, or another path is chosen and the admission decision process is repeated. By considering only the user-declared Peak Cell Rate (PCR) traffic parameter during the admission decision process, we adopt the conservative assumption that the new connection will transmit at its peak rate during its entire holding-time duration.

An advantage of this CAC framework is the non-centralized approach to making admission decisions. Control signals, including the PCR of a new

connection, are sent from the originating node to all relevant nodes along the end-to-end route. In their replies, these nodes will indicate if the new connection request can be accepted or rejected. All nodes must indicate "accept" before the new connection can be admitted.

As mentioned earlier, the three basic CAC algorithms, i.e., PRA, REM, and RS, are used to obtain the available bandwidth values. We will start by explaining the concept of available bandwidth calculated under the PRA method. It is called the PRA Free Bandwidth (PRAFB) for an end-to-end connection with n links and each link may have m established connections. Let P_k be the computed available bandwidth for the k-th link, and let C_k denotes the k-th link capacity, while $q_{k,u}$ denotes the u-th user's declared peak rate. Hence,

$$P_k = C_k - \sum_{u=1}^{m} q_{k,u} ,$$

$$PRAFB = \min_k \{P_k\}, \qquad k = 1, \ldots, n . \tag{3.22}$$

This approach is very conservative, and hence it can guarantees the requested QoS but at the cost of inefficient utilization of the network capacity.

Next, the REM and RS methods are considered. The former method does not consider the use of a shared buffer amongst the established connections of a link. However, the same formula is used for both methods, and the only difference is that the shared buffer parameter is set to a zero value whenever the REM method is considered. The estimated available bandwidth computed using the REM/RS method is called the RS Free Bandwidth (RSFB) for an end-to-end connection with n links. This method will achieve a higher utilization of the network resources because it takes advantage of multiplexing gain, although it may occasionally not meet the requested QoS.

The occasional failure to meet the QoS requirement can be attributed to the failure to predict accurately the near-future aggregate traffic behavior, based

on past traffic information contained in the traffic histograms. Notice that the real traffic processes may not be stationary. Even if these are stationary, the MBCAC approach may not have sufficient traffic information on the recently accepted connections, simply because these connections may have not been running long enough for their statistical characteristics to be known.

Assuming the traffic process is stationary and all statistical characteristics of the aggregate traffic stream are known, then the estimated service bandwidth required for all connections will be lower bounded by the RS method and upper bounded by the REM method. In other words, the aggressive RS method will estimate the service rate that is less than or equal to what is really needed, while the REM method computes the service bandwidth that is equal to or more than the actual bandwidth consumption.

In reality, the traffic process is likely to be non-stationary, and therefore the RS and the REM methods cannot guarantee to provide bounds for the estimated bandwidth needed; rather the estimation procedure is somewhat like a game of guesses. In this case, the RS method still provides a lower bounded guess, or an optimistic estimate of the bandwidth required, but the upper bounded guess is given by the most conservative PRA method and not by the REM method.

To calculate the RSFB, statistical traffic information is collected for the aggregate traffic stream on every link, i.e., locally at each network bottleneck. This is done by measuring the total amount of work that arrives within a window, and recording it in a database called *Traffic histogram*. A collection of such traffic histograms are used for several different window sizes. With the use of different window-size based traffic measurements, the MBCAC approach will have on record the past traffic intensities for a range of time-scales [MV96].

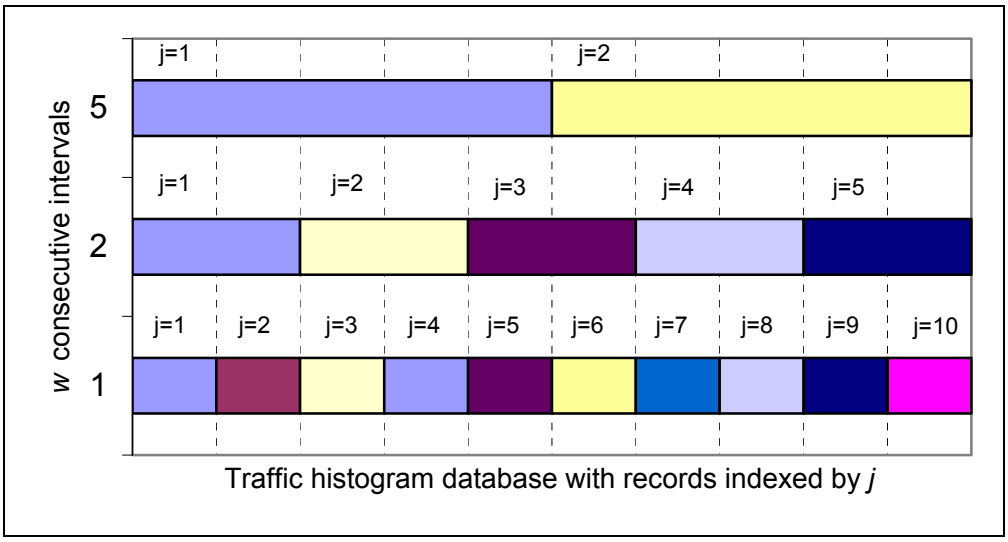


Figure 3-5. Time-scale diagram of different traffic histogram database for w = 1, 2 and 5.

Let time be divided into fixed length intervals, the length of which may be chosen such that occasional traffic bursts can be captured. Let j_w denotes the record counter for a block of w consecutive intervals. Each traffic histogram will then record the total traffic measured during a span of w consecutive intervals. This is illustrated in Figure 3-5 for cases where w = 1, 2 and 5, given 10 samples of aggregated traffic measured during the first 10 fixed length intervals. For w = 1 interval length case, the traffic histogram will record 'j_1 = 1,…,10' samples of aggregated traffic measurement; while for the w = 2 consecutive intervals case, the 'j_2 = 1' record of the traffic histogram is the total traffic measured during the 1st and the 2nd fixed length intervals. The 'j_2 = 2' record is the total traffic measured during the 3rd and the 4th intervals etc. Likewise, for the w = 5 consecutive intervals case, the 'j_5 = 1' record of the traffic histogram is the total traffic measured during the first five non-overlapping consecutive intervals, namely during intervals 1st to the 5th (inclusive), and the 'j_5 = 2' record is the total traffic measured during the 6th to the 10th intervals.

Let the random variable $X_k(w)$ represent the total amount of aggregated work that arrives during w non-overlapping consecutive intervals in the k-th link.

Let A_k denotes a random variable representing the amount of aggregated work arriving from all established connections in the k-th link during one fixed length time interval, and let $A_{k,r}$ denotes a random variable representing the amount of aggregated work arriving from all established connections in the k-th link during the r-th time interval. Hence,

$$X_k(w) = \sum_{r}^{r+w-1} A_{k,r} \,. \tag{3.23}$$

Let L denotes the required QoS, i.e., desired CLR, and l denotes the shared buffer capacity. Using the past traffic load records stored in all traffic histograms, the minimum service bandwidth V_k, required by all established connections on the k-th link whilst meeting the desired QoS requirement, is derived as follows:

$$V_k = \max_{w}\{S_{k,w}\}, \tag{3.24}$$

where $S_{k,w}$ is the service bandwidth for a window size of w consecutive intervals, and it is obtained by:

$$S_{k,w} = \min\left\{ s : \ \frac{E\left[\{X_k(w) - sw - l\}^+\right]}{E[A_k]w} \le L \right\}, \tag{3.25}$$

and where the superscript '+' is defined by:

$$\{x\}^+ = \begin{cases} x, & \{x\} > 0; \\ 0, & \{x\} \le 0. \end{cases}$$

Equation (3.25) is explained as follows. The work that arrives during w consecutive time intervals that could not be served or stored (assuming the buffer is empty at the beginning of each window of w time intervals) must be lost. Furthermore,

$$\frac{E\left[\left\{X_k(w) - sw - l\right\}^+\right]}{E[A_k]w} = L_w, \tag{3.26}$$

is an estimate of loss which will be accurate to a satisfactory degree if sampling at intervals of length w, which represents the dominant time-scale for the loss process. In other words, it is an estimate of loss based only on time-scale of w. Equation (3.25) then selects the dominant time-scale as the one giving the worst loss.

Next, using V_k and C_k, which denotes the k-th link capacity, the free bandwidth is computed as shown: $R_k = C_k - V_k$,. Hence, the available bandwidth RSFB for an end-to-end connection of n links is given by:

$$RSFB = \min_k \{R_k\}, \qquad k = 1, \ldots, n. \tag{3.27}$$

This procedure of 'available bandwidth' evaluation is different from the effective bandwidth concept because the statistical multiplexing of all connections is considered. With both PRAFB and RSFB values calculated, the next step is to derive the overall maximum available bandwidth OLFB value for an end-to-end connection of n links. The OLFB value basically states the maximum spare bandwidth that is available for the newly admitted connection(s), whilst ensuring the QoS required by the established connections is still provided. For admittance of a new connection, its declared PCR must be smaller or equal to the derived OLFB value.

Section 3.4 provides more details on how the OLFB value is derived based on the choice of AFCM techniques and the instantaneous traffic load condition.

3.3.3 Traffic Histogram Update Issues

As explained in the previous section 3.3.2, the accuracy of the available bandwidth calculations is heavily dependent on the past traffic records stored in the traffic histograms. These traffic histograms are only useful if they hold records that can be used to generate accurate traffic statistics of the past

aggregate flow. This poses the question of how 'up-to-date' the traffic histograms are required to be in order to make efficient admission decisions in the presence of random connection departures.

Three traffic histogram update approaches, in varying complexities and storage requirements – from no update to complete updates, will be used to alter, if applicable, all traffic histogram records. The first two approaches will remove a departing connection's previous traffic contributions from the records of all traffic histograms. The difference between these two approaches is the amount of traffic contributions that are removed. And the last approach makes no attempt to update any records stored in the traffic histograms.

3.3.3.1 Exact Histogram Update

This is the most complex approach and it is called the *Exact histogram update*. In this approach, every amount of traffic that was transmitted by the departing connection is removed from all traffic histograms. This is achieved by recording the number of cells (work) each established connection transmits into a network bottleneck at every sampling interval. After the removals, all traffic histograms will now only contain the traffic contributions from connections that are still in progress. The requirement to record the number of cells from each connection is impractical and unrealistic.

Nevertheless, this approach is used as a performance benchmark for the study on the maximum level of efficiency that an MBCAC approach can achieve, when given accurate traffic statistics derived from the 'up-to-date' records of all traffic histograms.

3.3.3.2 ZT Histogram Update

This approach is less complex and it is called the *ZT histogram update* [ZT97]. In this approach, an approximate value is computed using the most recently derived minimum service bandwidth V_k, and the number of connections presently in the link. This computed value is then subtracted from all traffic

histograms' records. This simpler method is conservative in the fixed amount of traffic that is subtracted from the traffic histograms, and hence it is less efficient than the exact histogram update approach.

Under the ZT histogram update approach, when a connection departs from the network, all traffic histograms are modified by an approximate value U_k as if a CBR traffic stream (at that rate) had departed from the k-th link. Let m denotes the number of established connections at the k-th link. Using the estimated minimum service bandwidth V_k from (3.24), U_k is computed as shown:

$$U_k = \left(1 - \sqrt{1 - \frac{1}{m}}\right) V_k . \tag{3.28}$$

3.3.3.3 No Histogram Update

This is the simplest approach amongst the three traffic histogram update approaches. It is called the *No histogram update* since no update is performed when a connection departs from the network. In other words, no attempt is made to remove the traffic contributed by the departing connection from all traffic histograms. Hence, the available bandwidth values are computed based upon the conservative viewpoint that recently departed connections *still contribute* to the overall statistical behavior of the present established connections. This would imply an inefficient utilization of valuable network capacity. Nevertheless, this approach demands minimal computing and storage requirements.

Notice that in this approach, no exponential weighting is given to recent traffic histogram data; hence, to maintain the adaptability of the MBCAC approach to changing traffic load conditions, the AFCM mentioned in section 3.4 is used for this purpose.

3.3.4 Constraint Liberalization Issues

From (3.25), an estimated bandwidth (for time-scale w) required to service the established connections is computed. However, by setting the L to be equal to the required CLR, it will ensure that the V_k always has a high minimum value. Nevertheless, if this constraint is slowly removed, for example, by increasing L such that it results in a very low or zero V_k value, the subsequent RSFB value will likewise increase, leading to a possible increase in OLFB value. With a higher OLFB value, more new connections will be accepted and the link utilization will increase, albeit at the cost of not meeting the required QoS.

3.4 Adaptive Feedback Control Mechanism

From Figure 3-1 (model-based CAC framework) and Figure 3-4 (histogram-based CAC framework), the MBCAC approaches use an additional method to ensure the desired QoS can be provided. Ideally, this method should continuously gauge the arrival traffic load and then actively adjusts the admission decision *mood* from conservative to daring, and vice-versa.

We term this method the Adaptive Feedback Control Mechanism (AFCM). It is basically a collection of two inter-dependent modules:

- Prudence level policy module, which actively adapts the MBCAC approach to changing traffic load conditions.
- Load and traffic measurements module, which compares the traffic load against a choice of different thresholds.

To use the AFCM on an MBCAC approach, a user will first select a technique from each module. Next, the user will decide on a threshold value for the chosen load and traffic measurements technique. Preferably, this threshold should be conservative so as to allow the chosen prudence level policy technique to be activated earlier; and hence be more effective in performing its duty to control the admission decision process between conservative and

daring decision behaviors, whilst still meeting the QoS requirement. However, the threshold should not be overly conservative such as to impede the admission decision process from achieving its primary aim of efficient link utilization performance.

3.4.1 Prudence Level Policy Module

In this section, two different preventive-action approaches are presented. The first approach provides a marker between two extreme CAC methods, while the second approach derives a '*Warming-up period*' for the newly accepted connections. Both approaches use an adaptive prudence level factor p (a real number value between 0 and 1) to adapt the admission decision process to changing traffic load conditions.

The following sections will describe the two approaches in detail:

- Section 3.4.1.1 – Adaptive weight feedback method: The adaptive prudence level factor is used to continuously adjust the derived overall available bandwidth value in order to indirectly calibrate the admission decision process between conservative and daring behaviors.

- Section 3.4.1.2 – Adaptive warming-up period method: The adaptive prudence level factor is used to compute a value equal to the initial duration of a connection's holding time. During this duration, the admission decision process considers the newly permitted connection to be transmitting at its declared peak rate. After this duration, the connection's arrival traffic load will then be recorded in all traffic histograms.

3.4.1.1 Adaptive Weight Feedback (AWF) Method

Excluding the few special cases of CBR connections, very strict QoS requirements, and/or very bandwidth-hungry applications, the PRA method is clearly too wasteful. In a real practical sense, it is highly unlikely that a connection will transmit at its declared peak rate continuously. Therefore,

statistical multiplexing can be exploited to achieve higher link utilization, and hence this would imply the frequent use of the REM/RS method to achieve this aim. However, this method is daring and may lead to QoS violation. The QoS breach may occur because the traffic process is non-stationary, and/or its variance is too high.

Hence, the optimal point of an MBCAC operation may not be solely based on either the PRA method, or the combined REM/RS method; rather, it could be a linear combination of the two, i.e., anywhere in between, and the movement of the optimal point that defines the relative contribution of each method should be based on the present traffic load conditions.

In particular, a control parameter called the adaptive prudence level factor p is used. Preferably, this parameter should be adaptive so that during periods of network congestion, the OLFB at any congestion points will be reduced and approach a value almost equal to the PRAFB. On the other hand, during periods of no congestion, the OLFB should approach the RSFB value so as to admit more new connections and hence maximize link utilization.

Therefore, the derived OLFB value expressed in (3.29) will depend on the instantaneous traffic load condition illustrated in Figure 3-6. The benefit of this concept is that it allows for inaccurate service bandwidth predictions.

The OLFB value is computed by:

$$OLFB = (1-p)PRAFB + pRSFB\,, \tag{3.29}$$

where p ranges between 0 and 1.

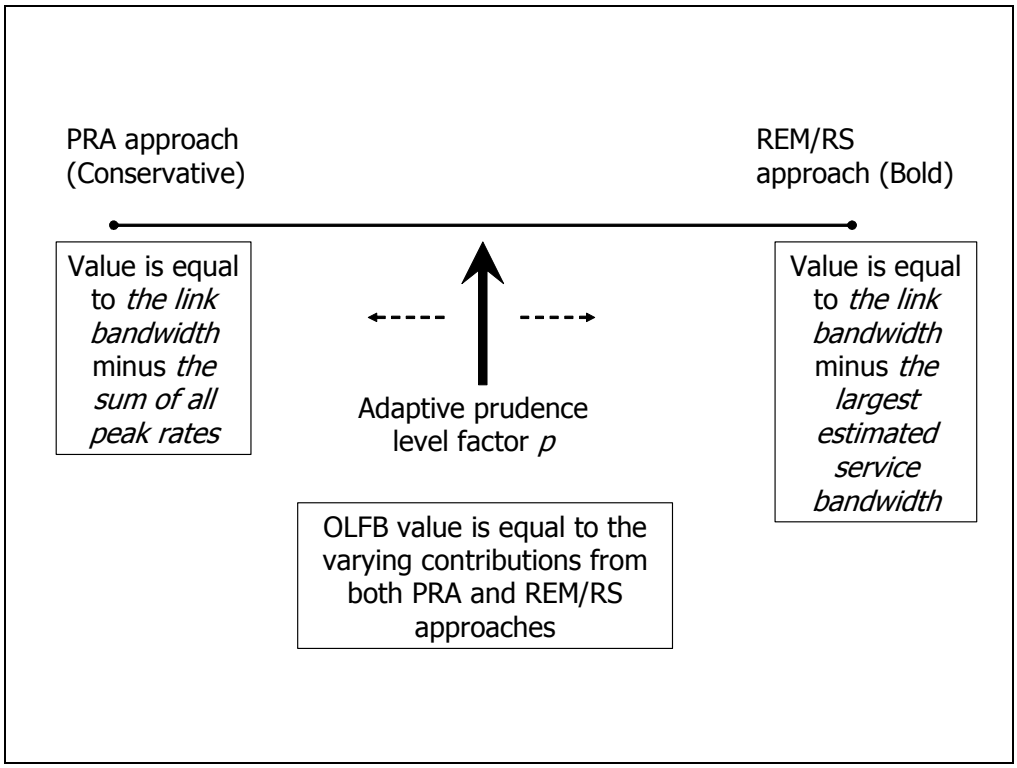


Figure 3-6. Overall Link Free Bandwidth (OLFB) evaluation process.

We will now explain how to derive the adaptive prudence level factor on the *k*-th link. Let C_t denotes the link capacity during one fixed length interval (at the *t*-th sampling interval), and let A_t denotes the arrival rate from all established connections during the *t*-th interval. Suppose *l* is the shared buffer capacity and *H* denotes a threshold whose value depends on which of the approaches mentioned in section 3.4.2 is used. Typically, these approaches, i.e., Link occupancy, Buffer occupancy, and Cell loss conservative period (see section 3.4.2 for further details), determine how the arriving traffic will be measured against the threshold.

When the amount of arriving work is more than the threshold, *p* is decreased. For the Link occupancy approach, it is computed for the (*t*+*1*)-th interval as shown:

$$p_{t+1} = \max\left\{0 \quad , \quad p_t - \frac{A_t - H}{C_t + l - H}\right\}. \tag{3.30}$$

For the Buffer occupancy approach, we measure the amount of work that overflows into the shared buffer during the *t*-th interval, and we call this Y_t. Therefore, *p* is computed by:

$$p_{t+1} = \max\left\{0 \quad , \quad p_t - \frac{Y_t - H}{l - H}\right\}. \tag{3.31}$$

However, when the work is less than the threshold, *p* is increased as follows for both approaches:

$$p_{t+1} = \min\{1 \quad , \quad p_t + F\}, \tag{3.32}$$

where *F* is a real number used to increment the adaptive prudence level factor. A low real number value for this *F* parameter should be chosen such that an MBCAC approach will slowly (in small increment steps) change from conservative to daring admission decision behavior.

For the Cell loss conservative period approach, *p* is set to 0 for a certain duration whenever cells are lost. When there is no loss, *p* will be increased as shown in (3.32).

3.4.1.2 Adaptive Warming-up Period (AWP) Method

Typically for MBCAC approaches, when a new connection is admitted, minimal traffic information is known initially about this connection's traffic behavior. Hence, the CAC algorithms will not be able to compute accurately the aggregate service bandwidth required by all the established connections. To alleviate this error, a new connection is assumed to be transmitting at its declared peak rate for a certain initial duration called the *Warming-up period* (WP). Only for duration exceeding this WP, will the traffic transmitted by any connection be considered in all traffic histograms [ZT97].

Using the calculated adaptive prudence level factor *p* of section 3.4.1.1, and having WP_{t+1} denotes the warming-up duration that all new connections in the *(t+1)-th* sampling interval will be subjected to, the WP is calculated as shown:

$$WP_{t+1} = (1 - p_t)D\,, \tag{3.33}$$

where D is a fixed real number value representing the maximum allowable warming-up period. Hence, the OLFB is computed by:

$$OLFB = RSFB - PW\,, \tag{3.34}$$

where PW denotes the sum of peak rates for connections whose holding times are still within their individually computed WP duration.

3.4.2 Load and Traffic Measurements Module

The load and traffic measurements module contains different approaches that will determine: (1) the type of threshold, and (2) the rule of comparison, i.e., comparing the arrival traffic volume or the amount of cell loss against a criterion that is unique to each approach. Basically, the choice of approach will affect the adaptive prudence level factor value, which in turn affects how conservative or daring the admission decision process is towards the admittance of new connections.

The approach will decide at which *point-in-time* preventive actions should be taken by the prudence level policy technique, with advance knowledge that there may be a possible QoS breach by the present established connections. For all approaches, the traffic comparisons are made per sampling interval. The load and traffic measurements approaches are:

- Link Occupancy (LO).
- Buffer Occupancy (BO).
- Cell Loss Conservative Period (CLCP).

The LO method measures the amount of traffic load against a threshold whose value is some percentage of the SSQ server rate. The BO method measures the amount of buffer load against a threshold whose value is some percentage

of the buffer size. Basically, for the occupancy-based methods, the threshold *H* used in section 3.4.1.1 is set as a percentage of the link or buffer capacity.

The CLCP method measures the amount of traffic loss and then computes a threshold value, in term of the time-periods that a CAC approach will be conservative. For example, if a cell loss event occurs, and the CLCP estimates a QoS-recovery duration of 5 seconds, then during the immediate following 5 seconds, the adaptive prudence level factor is fixed at 0 and hence forcing the admission decision process to be conservative. To compute this time-period value, the amount of cell loss is multiplied with a maximum allowable conservative period (max_CP) value, as shown:

$$\text{Conservative Period} = \text{Lost Cells} * \text{max\_CP} .$$

The use of the CLCP is the most daring approach because we wait until damage is done before any actions are taken. On the other hand, the LO approach is the most conservative because preventive actions are initiated whilst there is still unused bandwidth left in the link.

3.5 Conclusion

In this chapter, we have presented two CAC frameworks for a multiservice network. These frameworks are formulated based on our philosophy of practical admission control methodologies. Consequently, a CAC approach should:

- Be practical and simple to use from the viewpoints of both network provider and users.
- Require minimum number of a-priori traffic information at connection set-up phase.
- Be robust to occasional surge in bandwidth demands.

The first CAC framework consists of the traditional Gaussian model-based CAC scheme. This model is used to compute the equivalent bandwidth of the aggregate flow, and it is exact only when large number of established connections is present such that the aggregate traffic stream is highly aggregated. In light of this observation, we have introduced an enhanced Gaussian CAC scheme that is efficient for aggregate traffic stream that exhibits either Gaussian or non-Gaussian behavior. Because these a-priori traffic models require accurate traffic descriptor to be provided at connection set-up phase, alternative CAC schemes that lessen this stringent pre-requisite are introduced. These alternative schemes use real-time measurements of the aggregate traffic load to compute relevant traffic statistics such as mean and variance required by the a-priori traffic models. Some of these MBCAC schemes use the AFCM to provide an additional control layer for QoS assurance.

The second CAC framework is a collection of different modules that makes up the complete admission decision process that is purely measurement-based. Each module is made up of different techniques, hence the network provider can 'mix and match' each unique technique from all the modules to create a custom-made MBCAC scheme that will meet the desired traffic control requirements. In other words, many possible MBCAC schemes can be formulated from this framework. These modules are: Available bandwidth evaluation methods, Traffic histogram update methods, Constraint liberalization methods, and AFCM methods.

The AFCM is basically an adaptive feedback controller that aims to ensure the desired QoS is provided to the established connections. It is made up of two inter-dependent modules, i.e., 'Prudence level policy', and 'Load and traffic measurements'. For the MBCAC schemes that use the AFCM, different combinations of technique chosen from the two modules will produce different admission decision patterns. Hence, depending on the chosen AFCM

settings, an MBCAC scheme can be customized to be either very conservative or very daring in its admission decisions.

The two CAC frameworks can be used to test various traffic control strategies, and in particular, to focus on the simplicity versus efficiency tradeoff issues. In other words, if significant additional complexity does not improve efficiency substantially, simpler methods may be used.

Although this thesis is built upon ATM technology, many of the connection admission control ideas proposed here are general, and hence versatile enough to be applied to the future Internet evolution.

4 Simulation Methodology

4.1 Introduction

The performance of Connection Admission Control (CAC) schemes based on cell loss models using only a-priori traffic descriptor can be verified through formal proof. However, Measurement-based CAC (MBCAC) schemes can only be verified through empirical studies on either real networks or a simulator. In this thesis, all CAC and MBCAC schemes are tested through a simulation system driven by various traffic sources made up of real traces.

The remainder of this chapter is as follows:

- Section 4.2 presents the common simulation system used to test the CAC and MBCAC schemes.
- Section 4.3 presents a connection arrival and holding-time model, and various real traces.
- Section 4.4 presents the different parameter values chosen to test the performance of the CAC and MBCAC schemes.

4.2 Simulation System

In this thesis, the performance studies of all CAC and MBCAC schemes are carried out using software simulation of a multiplexer Single Server Queue (SSQ) with a buffer that is shared amongst all connections, and with First-In

First-Out (FIFO) queuing discipline on an end-to-end link as depicted in Figure 4-1.

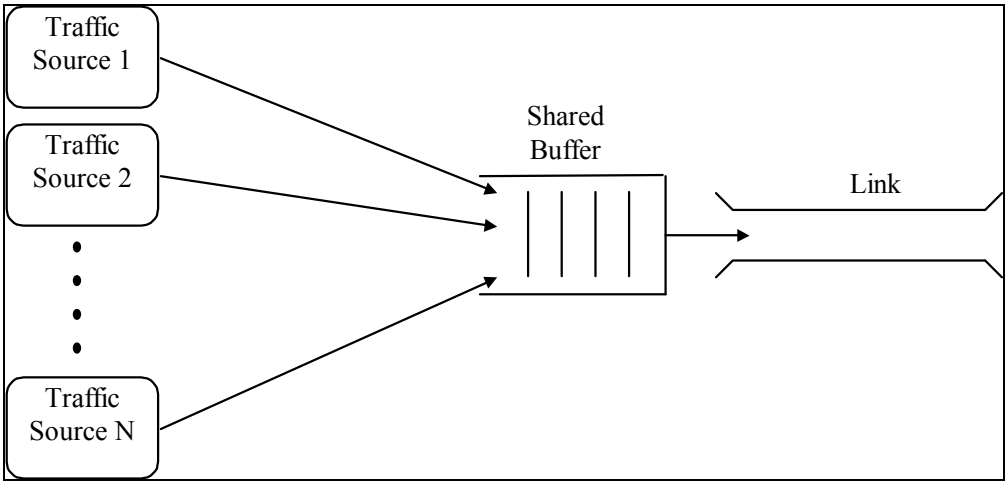


Figure 4-1. Multiplexer queue model.

This SSQ will service all traffic connections wanting to transmit work through the link. However due to its limited link capacity, as increasing number of new connections are accepted readily without any admission controller, the demand on the SSQ may be greater than what it can service and buffer. Hence, when such a situation arises, the overflowing traffic will be discarded. Consequently, the users will perceive this as a degradation of the service rendered by the network provider.

To provide top quality service is the Holy Grail for all network and service providers. Such quality service comes at a price, with the best service plan commanding a premium price payable by the users. Hence, this thesis studies the effectiveness of using a variety of admission controllers to provide Quality of Service (QoS) to the established connections transmitting work through the SSQ. Only QoS for the aggregate flow is considered in this thesis because it is a commercially viable and realistic service model. QoS for an individual VBR traffic stream is not considered because in reality, such QoS assurance places huge computing and storage demands on the network switches.

4.3 Traffic Sources

To simulate realistic VBR traffic sources that may be relevant to CAC operations for the purpose of performance evaluation and dimensioning, combinations of a connection arrival and holding-time model, and four real traces, are used in this thesis. These traffic sources exhibit different statistical characteristics of real traffic in terms of correlation and burstiness factors. Through these sources, the performance of various CAC and MBCAC approaches are studied under homogeneous and heterogeneous traffic scenarios.

For the connection arrival and holding-time model, the *M/Pareto Model* is used. While for the real traces, these are recorded from two types of traffic, namely, network data and video. These are described further in the following sections 4.3.1 and 4.3.2 respectively.

4.3.1 M/Pareto Model

Studies have recently shown that network traffic often exhibits Long Range Dependence (LRD) characteristic, resulting in long congested periods leading to possibly large increases in the cell loss rate [BSTW95, DMRW94, ENW96, GW00, LTWW93]. Currently, LRD traffic forms a significant part of the broadband networks' payload. Typically, regardless of its traffic source, LRD traffic is characterized by substantive long bursts. Examples of activities that may cause such bursts are the transfer of large files or the transmission of large amount of data to update remote databases.

The traffic model used to simulate such long bursts is the M/Pareto (MP) model considered in references [ANZ02, NZA02, NZA99]. This model is closely related to [LTG95], and is one of a family of such processes that form a sub-group of the more general M/G/∞ model [KM98].

Figure 4-2 illustrates an example of the M/Pareto traffic. In this example, a total of five connections are admitted into a link with independent start-times

and holding-times. The total work from all these established connections forms the aggregate flow's traffic load as 'seen' by the SSQ.

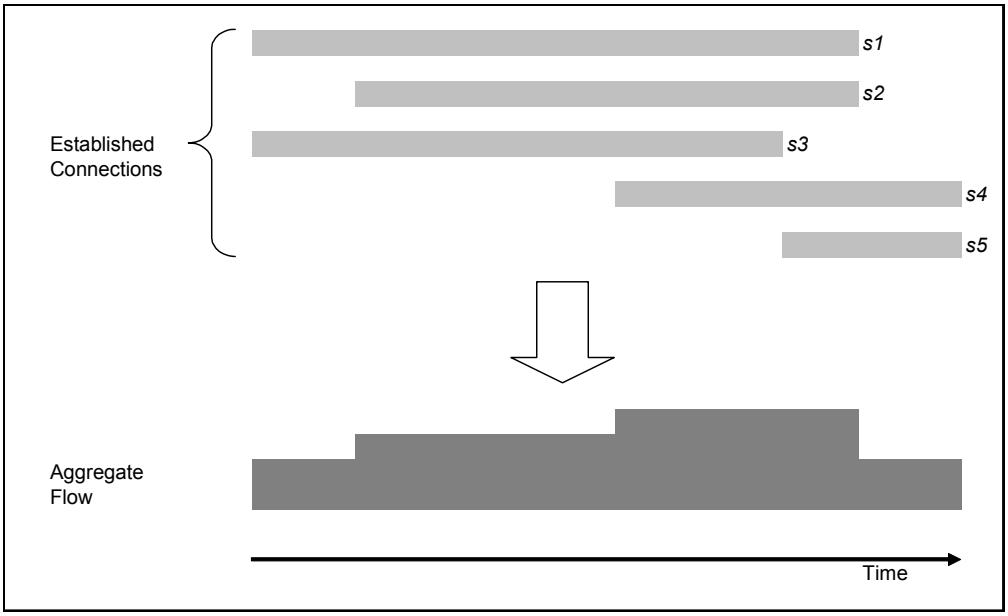


Figure 4-2. Traffic source – M/Pareto model.

In our performance studies, the arrivals of new connections are basically modeled by a Poisson process with rate λ of Pareto distributed overlapping bursts. For each new connection, its holding-time is equivalent to the computed burst time. In other words, λ controls the frequency with which new bursts (or connections) commence; and the burst time of a connection Y, is random and derived from the Pareto distribution. Hence, the complementary distribution function for a Pareto distributed random variable is given by (4.1), where $1 < \gamma < 2$ and $\delta > 0$.

$$P(Y > y) = \begin{cases} \left(\frac{y}{\delta}\right)^{-\gamma}, & y \geq \delta; \\ 1, & \text{otherwise}. \end{cases} \tag{4.1}$$

In addition, the mean burst time is:

$$E[Y] = \frac{\delta\gamma}{\gamma - 1}; \tag{4.2}$$

and the variance of Y is infinite. During a burst, the connection will transmit work taken from one of the real traces.

This model is asymptotically self-similar with Hurst parameter given by:

$$H = \frac{3-\gamma}{2}. \tag{4.3}$$

The superposition of two independent M/Pareto processes with identical burst length distributions will itself be an M/Pareto process with Poisson arrival rate equal to the sum of the two processes' arrival rate. Therefore, increasing λ will increase the number of connections that make up the aggregate M/Pareto traffic stream. Take for example an aggregate stream with $\lambda = 10$; this would be equivalent to multiplexing 10 independent streams each with $\lambda = 1$.

4.3.2 Real Traces

To measure the performance of the CAC and MBCAC schemes under realistic traffic scenarios, these schemes are subjected to real traffic traces collected from four sources.

4.3.2.1 Network Data Traffic

For the network data traffic trace, it was collected along an Ethernet backbone within the University of Melbourne. The traffic load was measured every one second in term of the number of 53-byte cells handled at the bottleneck. A total of 479,800 per-second-samples were recorded across 5.6 days, i.e., from Monday 10 pm to Sunday 11 am. Thus, the Ethernet trace captures the typical traffic volume experienced by a network during weekdays and weekends. As shown by Leland et al. [LTWW93], Ethernet traffic exhibits LRD characteristic.

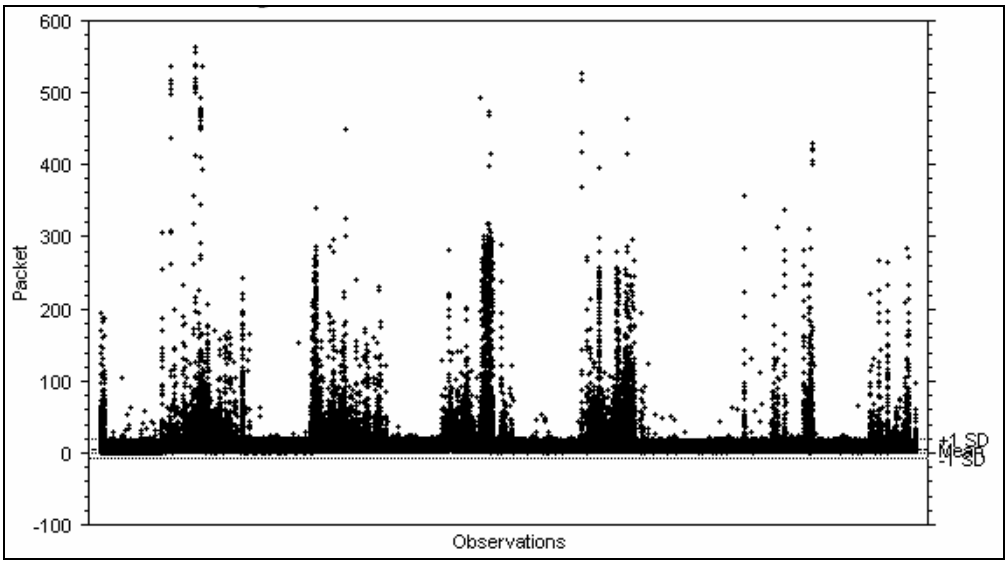


Figure 4-3. Real trace – Network data traffic, 5.6 days: Monday 10 pm to Sunday 11 am.

Figure 4-3 shows the complete Ethernet trace with unit-of-time equal to one second and time (x-axis) increasing from left-to-right. On the right side of the graph, the mean (μ = 5.5 cell/sec) and +/- 1 standard deviation (SD) are outlined. Variance σ^2 is 168.3 cell/sec, and the peak rate *pr* is 564 cell/sec.

From this 5.6 days trace, we truncate it down to a 12 hours long weekday sample, i.e., 43,200 per-second-samples from Wednesday 10 am to 10 pm; as shown in Figure 4-4. The measured parameter values (in cell/sec) are: μ = 8; σ^2 = 120.8; and *pr* = 449. Burstiness (the ratio of its peak to average rates, pr/μ) is 56.1.

In our performance studies, we use this 12 hours Ethernet trace as the default network data traffic trace, and it is denoted by NT.

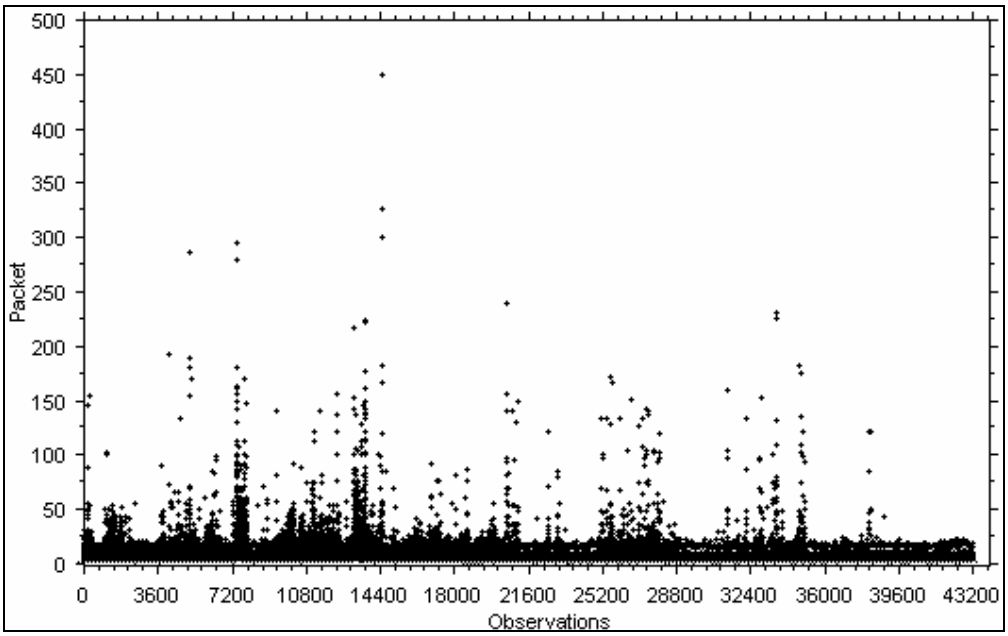


Figure 4-4. Real trace – Network data traffic, 12 hours: Wednesday 10 am to 10 pm.

To ensure the results of the CAC performance studies are valid, additional studies to verify these results are performed using different real trace (but same traffic type), and this trace is denoted by NT2. More details on how this trace is used for verification purpose are provided in chapter 5.

NT2 is a network data trace containing 12 hours long traffic collected from the same Ethernet backbone within the University of Melbourne during Tuesday 10 am to 10 pm. The measured parameter values (in cell/sec) are: $\mu = 8.5$; $\sigma^2 = 385.2$; and $pr = 564$. Burstiness is 66.3. Note that this NT2 trace is more bursty than the default NT trace.

4.3.2.2 Video Traffic

The authors of [GW94, Ros95] have shown that LRD exists in VBR video traces. In our performance studies, we use a 121 minutes trace produced by Garrett et al. [GF94] of a MPEG-1 [ISO93] encoded 'Star Wars' movie that exhibits statistical properties characteristic of LRD traffic.

The original full-length video was captured in 408 lines by 508 picture elements (pels), and then interpolated and filtered to the Common Interchange Format (CIF) [ITU00b] frame size, which is 240x352 pixels for intensity (Luminance - Y), and 120x176 pixels for two color components (Chrominance - U and V).

The video was compressed at 24 frames per second. The sequence of MPEG Intra (I), Predictive (P), and Bidirectional (B) frames used is IBBPBBPBBPBB IBB...; which gives 12 frames in a Group of Pictures (GOP). The trace records the number of bits per video frame in a sequence of 174,136 consecutive frames.

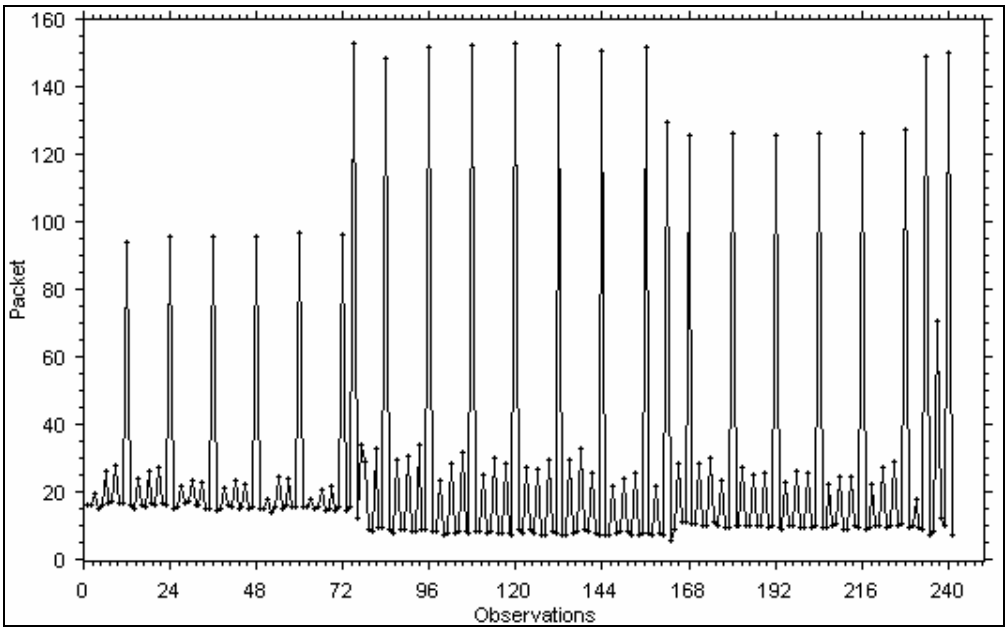


Figure 4-5. Real trace – Video traffic. Only 10 seconds worth of video trace is shown here.

Figure 4-5 shows a line graph for a randomly chosen 10 seconds window. The y-axis denotes the number of recorded 53-byte cells, while the x-axis is in unit-of-time equal to one frame-time, i.e., $\frac{1}{24}$ th of a second. The consistently large frames observed every 0.5 seconds are the I-frames.

The measured parameter values for the complete video trace (in cell/frame-time) are: $\mu = 36.8$; $\sigma^2 = 1835.4$; and $pr = 437$. Burstiness is 11.9.

In our performance studies, we use this 'Star Wars' movie trace as the default video traffic trace, and it is denoted by VT.

In addition, we also use another video trace, which is denoted by VT2, for CAC performance verification purpose. The VT2 trace is a 'Terminator 2' MPEG-1 encoded half-hour video trace produced by Rose [Ros95]. The video was compressed at 24 frames per second. The measured parameter values (in cell/frame-time) are: $\mu = 25.7$; $\sigma^2 = 574$; and $pr = 187.6$. Burstiness is 7.3. Note that this VT2 trace is less bursty than the default VT trace.

4.4 Parameter Settings

In this thesis, the CAC and MBCAC schemes are subjected to different traffic scenarios. In the following sections, the various parameter values used in the performance studies are provided.

4.4.1 For All CAC Approaches

The following list shows the common simulation settings used in the performance studies:

- For the discrete-time simulations, the simulation time is divided into fixed length sampling intervals; the length of which may be chosen arbitrarily such that occasional traffic bursts can be captured. In our performance studies, the length of this sampling interval is dependent on the traffic sources used by the connections in the simulated SSQ.

- Simulations are run up to 200,000 sampling intervals with a simulation start-up period equal to the initial 10,000 intervals. During such a period, the CAC is very conservative because admission decisions are made based upon the PRA approach of available service bandwidth

computation. Consequently, performance-related quantities are only measured after this start-up period.

- Performance-related quantities are measured from the aggregate traffic stream, rather than from the per-flow traffic.

- Required QoS for the aggregate traffic stream is fixed at CLR = 10^{-4} cells.

- Two SSQ buffer sizes are used: $l = 0$ and 500 cells.

- Connection inter-arrival times are exponentially distributed with rate, λ = 0.6 connection/second.

- Connection holding times are derived from the Pareto distribution with mean burst time, $E[Y]$ = 300 seconds. Throughout the holding time of a connection, work is sent to the SSQ. Typically, each traffic flow is a sub-sequence of a real traffic trace with randomly chosen starting point. If need be, the trace is wrapped around to the beginning when the end is reached. A traffic flow can only transmit work derived from a real trace, either network data or video.

- Hurst parameter, $H = 0.8$ [Ros95].

- The number of established connections in a link changes dynamically as random number of connections arrive or depart.

- The CAC and MBCAC approaches are subjected to homogeneous and heterogeneous traffic scenarios. A homogeneous aggregate traffic stream contains established connections using only one real trace, either network data or video. For the heterogeneous case, the aggregate traffic stream is made up of two groups of connections, i.e., one group using network data trace and the other group using video trace. A new connection will choose either trace with a 50%

probability. Hence, at the SSQ bottleneck, the aggregate flow is made up of two different types of traffic.

Real trace	SSQ server rate	
	bit/sec	cell/sec
NT	650 K	1533
VT	60 M	141.5 K
NT and VT	30 M	70.75 K
NT2	850 K	2004.7
VT2	40 M	94.34 K
NT2 and VT2	20 M	47.2 K

Table 4-1. SSQ server rates for real traces.

Table 4-1 shows the SSQ server rate for network data (NT and NT2) and video (VT and VT2) traffic traces. These rates are used in the performance studies for both SSQ buffer sizes of 0 and 500 cells. SSQ server rates for NT2 and/or VT2 connections are chosen so as to give connection-blocking rates slightly on par with those obtained when using the default NT and/or VT connections. This is to allow for fair comparisons of measured CAC performance.

We assume that CAC is responsible in maximizing link utilization, subject to meeting QoS required by the admitted connections. Thus, reducing blocking probability is not the main CAC's responsibility but rather it is a separate routing issue. We intentionally load the CAC heavily in order to study its traffic control efficiency. Henceforth, we are not concern with how high the blocking probability may be.

4.4.2 Model-based CAC Framework: *q(n)* Parameter

As mentioned in section 3.2, all CAC and MBCAC schemes within the model-based framework rely on the performance margin parameter $m_{j,l}$ of (3.7) to compute the service bandwidth required by the new and established connections. For easy reference, this parameter is shown below:

$$m_{j,l} = \begin{cases} q_{j,l}(n_j)\sigma_j, & \text{for Gaussian related schemes;} \\ q_{j,l}(1)\sigma_j^2, & \text{for traditional Effective Bandwidth scheme.} \end{cases}$$

4.4.2.1 Aggregate Traffic – Gaussian Assumption

For the traditional Gaussian CAC and MBCAC schemes, the aggregate traffic stream is assumed to be Gaussian. That is to say that the minimum number of established connections required before the aggregate traffic stream reach a critical size that is large and highly aggregated, is not considered into the algorithm.

Hence, assuming the aggregate traffic stream exhibits Gaussian behavior *immediately* with the commencement of traffic flows in the link, the multiple factor *q(n)* is equal to *q(1)* for all *n* number of established connections.

The a-priori traffic parameter *q(1)* is obtained as follows (For completeness, we include the traditional Effective Bandwidth CAC scheme.):

$$q_{j,l}(1) = \begin{cases} \dfrac{d_{j,l,QoS}(1) - \mu_j}{\sigma_j}, & \text{for traditional Gaussian scheme;} \\ \dfrac{d_{j,l,QoS}(1) - \mu_j}{\sigma_j^2}, & \text{for traditional Effective Bandwidth scheme;} \end{cases} \tag{4.4}$$

where $d_{j,l,QoS}(1)$ denotes the minimum service bandwidth required to serve one connection of type *j* traffic in a link with buffer size *l*, whilst meeting a *QoS* guarantee of CLR = 10^{-4} cells; μ_j and σ_j^2 denote the mean and variance amount of work by a single type *j* connection respectively.

Table 4-2 lists the computed *q(1)* values for the network data (NT) and video (VT) traffic traces.

Traffic type, *j*	Buffer size, *l* (cell)	For Gaussian schemes	For Effective Bandwidth scheme
NT	0	37.04	3.37
	500	15.75	1.43
VT	0	7.80	0.182
	500	2.20	0.051

Table 4-2. A-priori traffic parameter *q(1)* values.

4.4.2.2 Aggregate Traffic – Gaussian Boundaries

Recall that when the number of established connections is small, the resulting aggregate traffic stream will be lightly aggregated and hence it may not exhibit Gaussian behavior. Therefore, the application of a traditional Gaussian CAC approach may not be appropriate.

For such a traffic scenario, the use of the enhanced Gaussian CAC approach will be more suitable because it considers the size of the aggregate traffic stream into its algorithm. As the number of established connections denoted by *n* slowly increases, the aggregate traffic stream will likewise slowly converge to Gaussian, leading the multiple factor *q(n)* to slowly converge to a stable value when the aggregate traffic stream actually exhibits Gaussian behavior.

For the enhanced Gaussian CAC approach, the multiple factor *q(n)* is not fixed, rather it is dependent on *n*. Through empirical studies, the computed *q(n)* values are plotted in Figure 4-6 and Figure 4-7 for the NT and VT homogeneous streams respectively. In both figures, it is observed that *q(n)* converges to a stable value with increasing *n* number of homogeneous connections.

Using these figures, the threshold boundaries that distinguish between non-Gaussian and Gaussian regions, are listed in Table 4-3. These boundaries are expressed in term of the minimum *n* number of homogeneous connections required for the *q(n)* to converge to a stable value.

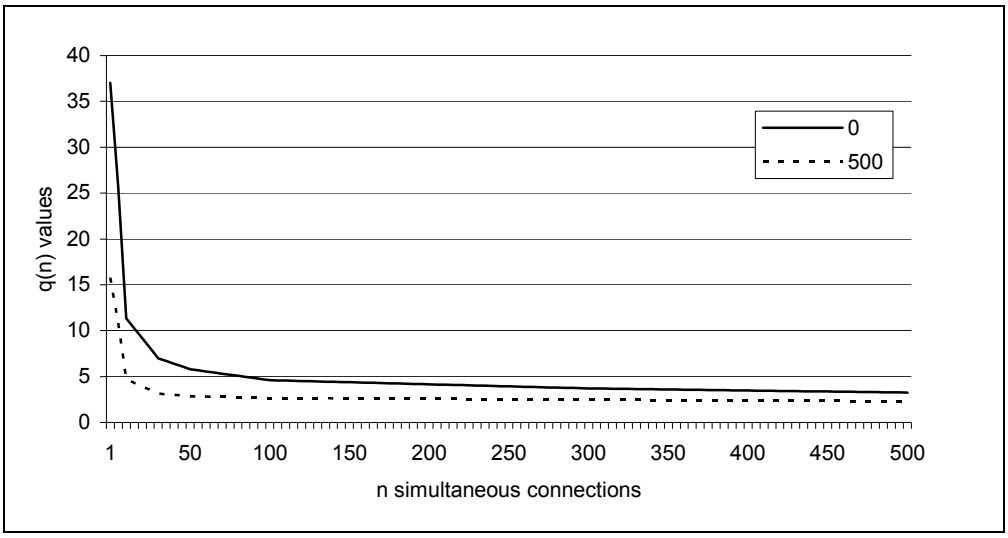


Figure 4-6. Convergence of *q(n)* for NT streams.

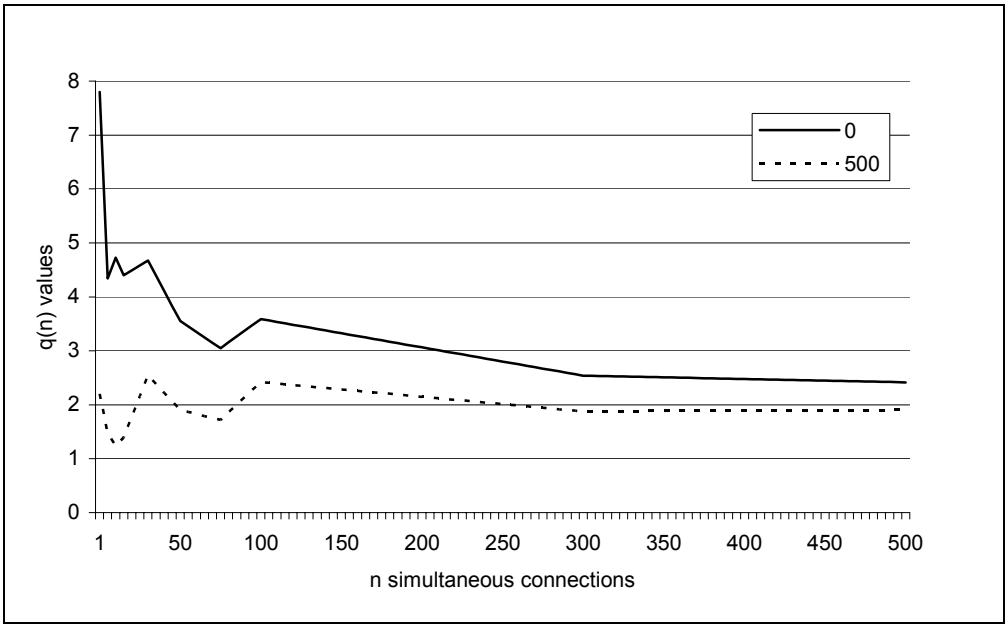


Figure 4-7. Convergence of *q(n)* for VT streams.

Traffic type, *j*	**Buffer size, *l* (cell)**	**Number of connections, *n***
NT	0	300
	500	200
VT	0	300
	500	300

Table 4-3. Estimated Gaussian boundaries – Distinguish aggregate traffic stream between non-Gaussian and Gaussian regions.

For the enhanced Gaussian CAC approach, the admission decision process includes looking up a $q(n)$ value unique to the present n number of type j connections from a $q_{j,l}(n_j)$ table for buffer size of l. However, if the $q_{j,l}(n_j)$ table contains $q(n)$ samples for every possible n values, the resulting table size will be too big and hence impractical.

Thus, for use in our CAC performance studies, this table instead contains a $q(n)$ value for every 5 connections, up to a maximum of 500 connections. Take for example a 10 connections case, the look-up table will have 3 samples, i.e., $q_{j,l}(1)$, $q_{j,l}(5)$, and $q_{j,l}(10)$; and for n = 5,…,9 cases, the $q_{j,l}(5)$ value will be used by the admission decision algorithm. Notice that the second $q(n)$ sample is used even for n = 8 and n = 9 cases. For low n value, such choice results in the CAC being slightly more conservative because the service bandwidth computation is made on the assumption that there is a slightly lower level of traffic aggregation.

4.4.3 Histogram-based CAC Framework: Parameters

Within the histogram-based framework, one parameter used by the Adaptive Feedback Control Mechanism (AFCM) is fixed as follows:

- F = 0.01, where F is the factor used in (3.32) to increase the adaptive prudence level factor p, whenever the arrival traffic volume or the amount of cell loss is less than a threshold.

A traffic histogram records the total amount of work that arrives within consecutive windows with each window having a fixed time-frame. Hence, to store the total amount of work across multiple time-scales, five traffic histograms for five different time-scales are used, and they are listed in Table 4-4.

Window size (sampling interval)	Number of samples
1	1000
2	500
5	200
10	142
100	50

Table 4-4. Simulation settings for five traffic histograms.

4.4.4 Adaptive Feedback Control Mechanism

The AFCM is used by some MBCAC schemes to provide an additional control layer for QoS assurance. It is basically a collection of two inter-dependent modules, i.e., Prudence level policy, and Load and traffic measurements.

The prudence level policy module contains two methods: Adaptive Weight Feedback (AWF), and Adaptive Warming-up Period (AWP).

The load and traffic measurements module contains three methods: Link Occupancy (LO), Buffer Occupancy (BO), and Cell Loss Conservative Period (CLCP).

In the performance studies, the mean prudence level factor is recorded and denoted by 'MWei' in the table of results. The mean warming-up period is denoted by 'MWP'.

4.4.4.1 Model-based CAC Framework: MBCAC Approaches

AFCM techniques used in the performance studies:

- For the prudence level policy module, only the AWF method is applicable. The AWP method is meant for MBCAC approaches within the histogram-based framework. It basically defines the time from which an active connection's traffic will be recorded into all traffic histograms.

- For the load and traffic measurements module, all methods are applicable.

To measure the effectiveness of the AFCM, the AWF method is used in combination with the common threshold values listed below:

- LO – 80% and 90% of SSQ server rate.
- BO – 10% and 20% of buffer size.
- CLCP – Maximum conservative periods of 5 and 30 seconds.

Each technique has a conservative threshold, e.g., LO=80%, and BO=10%. If the performance of the MBCAC schemes using these thresholds are poor, then we do not report on results using liberal threshold values, i.e., LO=90%, and BO=20%. This is because if a conservative threshold cannot guarantee QoS, then the liberal threshold will definitely not be able to meet the desired QoS.

On the other hand, for CLCP, we report on results using the liberal threshold, i.e., CLCP=5. If the performance is poor, we will then use the conservative CLCP=30 threshold.

4.4.4.2 Histogram-based CAC Framework: MBCAC Approaches

AFCM techniques used in the performance studies:

- For the prudence level policy module, all methods are applicable.
- For the load and traffic measurements module, all methods are applicable.

The maximum allowable warming-up period is set at 15 and 30 seconds. Below is a list of the common threshold values used in the performance studies:

- LO – 90% of SSQ server rate.

- BO – 20% of buffer size.
- CLCP – Maximum conservative periods of 5 and 30 seconds.

Note that the AWP method and the CLCP technique are not used together because both methods serve nearly the same purpose.

4.5 Conclusion

In this chapter, the simulation methodology we adopt to compare the performance of various CAC and MBCAC schemes is introduced.

The simulation system used in this thesis is driven by various traffic sources made up of real traces. To understand the behavior of these traces, a variety of traffic statistics and graphs are provided in section 4.3.2.

In section 4.4.2, the performance margin's multiple factor *q(n)* values used in the performance studies are given. In addition, Gaussian boundaries derived based on empirical studies are provided for both network data and video traffic.

We end this chapter by providing the numerical values for some of the common traffic and algorithm parameters used in the CAC performance studies.

5 Comparative Performance Studies

5.1 Introduction

The aims of this thesis are:

- To provide two CAC frameworks that contain CAC schemes grouped base on their admission control algorithm.
- To investigate how complex a CAC scheme needs to be in order to achieve a certain network efficiency level.

Accordingly, this thesis investigates simplicity versus efficiency tradeoffs for various CAC schemes and provides practical recommendations.

In chapter 3, we fulfilled the first aim by formulating two novel CAC frameworks. In addition, we provided details of each CAC scheme contained within these frameworks.

In this chapter, we aim to fulfill the second aim by conducting an intensive comparative investigation of the performance of the two CAC frameworks under both homogeneous and heterogeneous traffic scenarios. Different realistic traffic traces recorded from network data and video sources are used in these studies.

The fundamental role of a CAC is to ensure that established connections can experience an aggregate level of QoS that was agreed upon prior to connection set-up. In addition, the CAC must ensure that network providers can achieve good returns on their infrastructure investments through the admittance of as many connections as the network can cater within the QoS limits. Finally, the CAC should not demand unrealistic storage and computing resources from the network switches.

In other words, even if a CAC is effective in increasing network utilization whilst still meeting the required level of QoS, the CAC is not considered useful if it cannot be implemented in a cost effective manner. All CAC and MBCAC schemes considered in this thesis use admission controls that are based on session-level, and not packet-level. Hence, each scheme's operational cost is not expected to be exorbitant.

All results of the CAC performance studies are presented in tables. These tables list the measured performance quantities (PQ) such as: aggregate link utilization (Util.), aggregate cell loss ratio (CLR), mean simultaneous connections (MSC), mean blocking rate (MBR), and mean aggregate service bandwidth (MSB). In all studies, the simulation methods mentioned in chapter 4 are used.

For easy reference, a list of common acronyms and abbreviations is provided below:

- BO – Buffer occupancy technique.
- CLCP – Cell loss conservative period technique.
- EU – Exact histogram update technique.
- LO – Link occupancy technique.
- MBR – Mean blocking rate.

- MSB – Mean aggregate service bandwidth.
- MSC – Mean simultaneous connection.
- MWei – Mean prudence level factor.
- MWP – Mean warming-up period.
- NU – No histogram update technique.
- PQ – Performance quantity.
- TS – Traffic stream of type j.
- Util. – Link utilization.
- ZU – ZT histogram update technique.
- l – Buffer size in cell unit.
- L – Cell loss rate parameter.
- S – Aggregate service bandwidth.

The remainder of this chapter is as follows:

- Section 5.2 presents the performance results for CAC and MBCAC schemes within the model-based framework.
- Section 5.3 presents the performance results for MBCAC schemes within the histogram-based framework.
- Section 5.4 summarizes and lists the MBCAC schemes from both frameworks that consistently produce promising performance results.

5.2 Model-based CAC Framework: Performance Issues

The model-based framework contains three model-based CAC approaches and four measurement-based counterparts. Empirical-based performance studies of these CAC and MBCAC approaches are reported in the following sections:

- Section 5.2.1 – This section reports on the accuracy of the three model-based CAC approaches, i.e., the traditional Gaussian and Effective Bandwidth, and the enhanced Gaussian, in computing the amount of bandwidth required to service fixed number of established connections whilst meeting the desired aggregate QoS.

- Section 5.2.2 – The model-based CAC approaches require a-priori statistical knowledge of the traffic sources before service bandwidth can be computed for the aggregate flow. Hence, traffic parameters such as mean, variance and performance margin must be known prior to connection set-up phase. For the enhanced Gaussian CAC approach, it requires additional a-priori traffic information in the form of a multiple factor *q(n)* look-up table.

- Section 5.2.3 – The MBCAC approaches, i.e., measurement-based counterparts of the traditional Gaussian and the enhanced Gaussian, require minimal a-priori traffic information because a number of traffic parameters required by the approaches are measured real-time. However, because the arriving traffic load may exhibit non-stationary behavior and be unpredictable, an additional control layer is needed for QoS assurance. The Adaptive Feedback Control Mechanism (AFCM) is used to provide this additional control layer. Hence, the performance of each MBCAC approach with and without the AFCM are studied and reported here.

- Section 5.2.4 – All CAC and MBCAC approaches within the model-based framework use the performance margin's multiple factor *q(n)* values to compute the service bandwidth required by the established

connections to meet the desired aggregate QoS. Hence, this traffic parameter *q(n)* is central to the performance of these approaches. However, the *q(n)* needs to be measured for every traffic sources before a connection using one of these sources is admitted. This requirement is both impractical and operationally expensive. A realistic approach would be to use a standard set of *q(n)* values to help compute bandwidth consumption by connections belonging to the same type of traffic. For example, to compute the service bandwidth for connections with video type traffic, a standard set of *q(n)* values unique to video type traffic is used. Through empirical studies, we present and report on the performance of CAC and MBCAC approaches using these generic sets of *q(n)* values. Basically, we are interested in how accurate and closely-match the *q(n)* values must be to the actual traffic submitted by a connection in order to make efficient admission decisions.

- Section 5.2.5 – Here, we list the CAC and MBCAC schemes that consistently produce promising performance results.

5.2.1 Service Bandwidth *S*: Accuracy Issue

In this section, we investigate the accuracy of the three model-based CAC approaches, i.e., the traditional Gaussian (GA) and Effective Bandwidth (EB), and the enhanced Gaussian (eGA), in estimating the aggregate bandwidth *S* required to service static number of active connections whilst ensuring the desired QoS can be achieved.

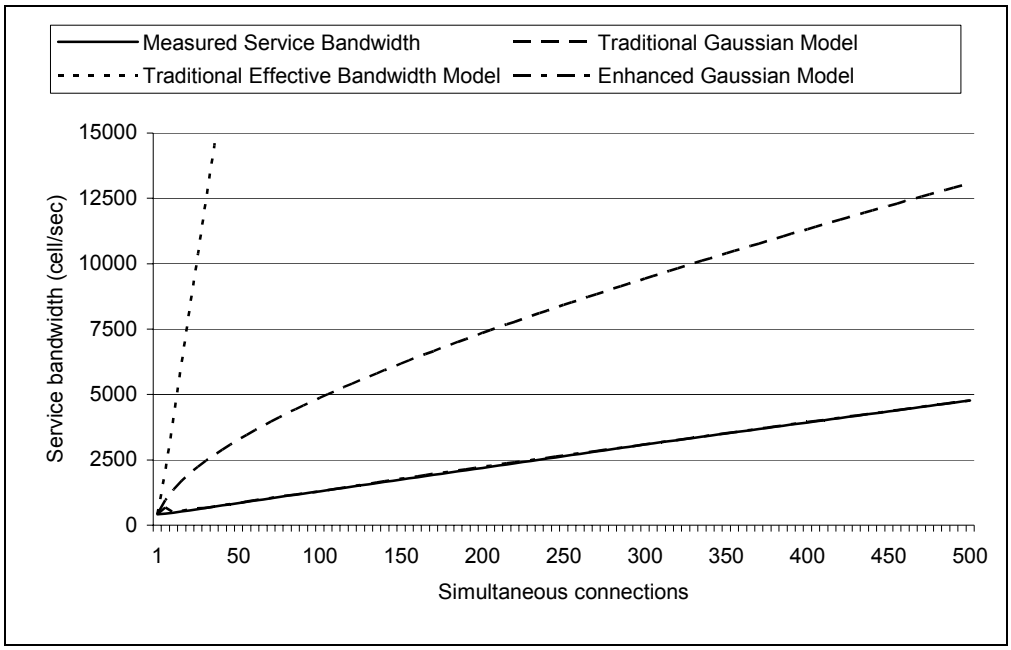
All figures in this section show the computed *S* value plotted against fixed number of simultaneous connections for buffer sizes of 0 and 500 cells. This trace is used with random starting point, and wrapped around to the beginning when the end-of-the-trace is reached. All connections are active from the start of the simulation run, and their holding times are equal to the complete trace's duration.

5.2.1.1 Homogeneous Traffic Streams

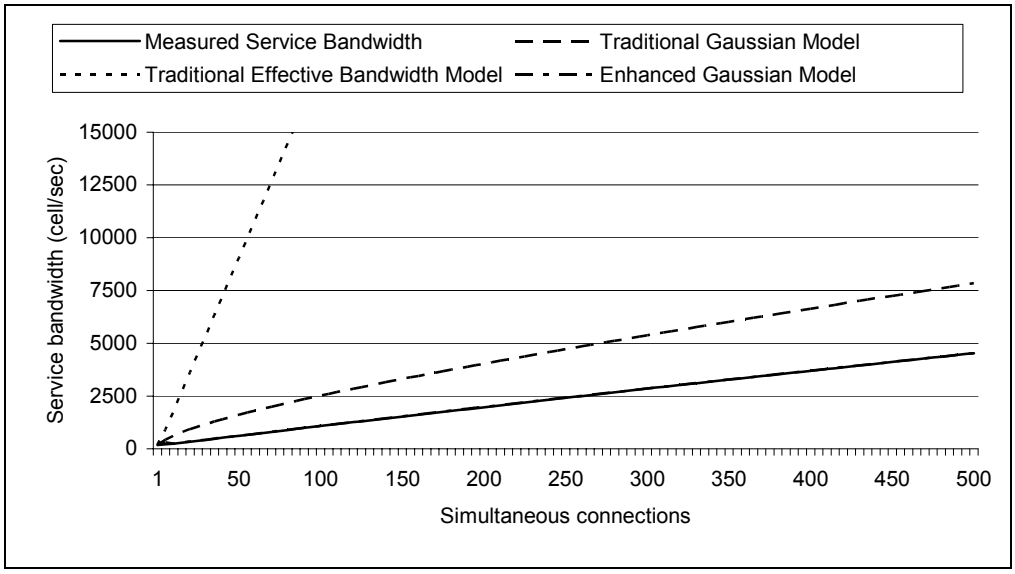
Figure 5-1(a) and (b) show the computed *S* against a fixed number of simultaneous connections transmitting work taken from the NT trace for buffer sizes of 0 and 500 cells respectively.

For connections using the VT trace, the results are plotted in Figure 5-2(a) and (b) for buffer sizes of 0 and 500 cells respectively.

From these figures, the eGA approach is observed to match the measured service bandwidth closely. This is expected since the approach uses a *q(n)* look-up table whose values are derived from actual measurements of the traffic source. However, the EB approach over-estimates the *S* values by a big margin. Hence, this approach is highly inefficient because it does not consider the level of traffic aggregation. For the GA approach, it only performs well for $l = 500$ case. The reason being the fixed a-priori multiple factor *q(1)* value is lower than the multiple factor *q(1)* value for the $l = 0$ case. Hence, this results in lower *S* values.

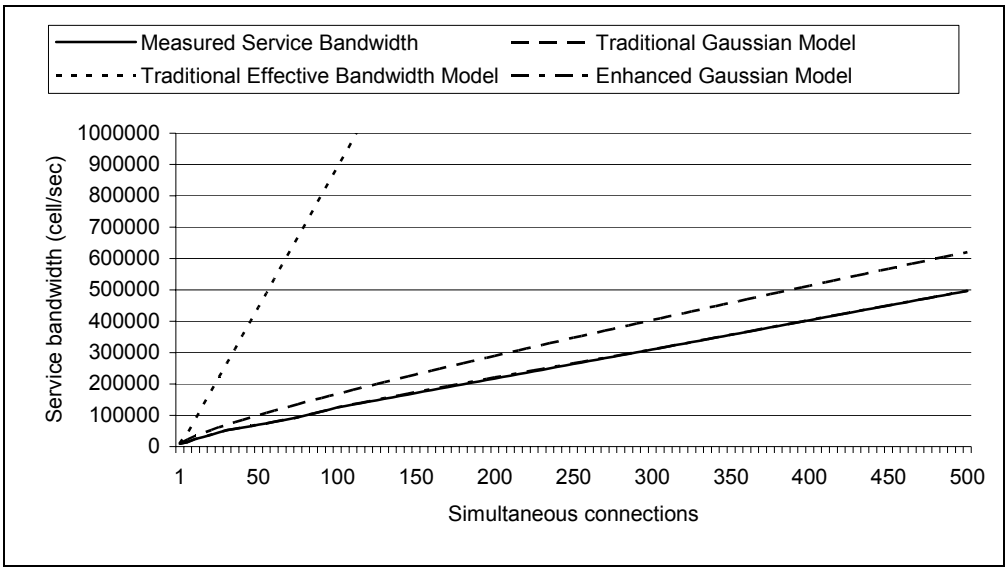


(a) Buffer size, $l = 0$.

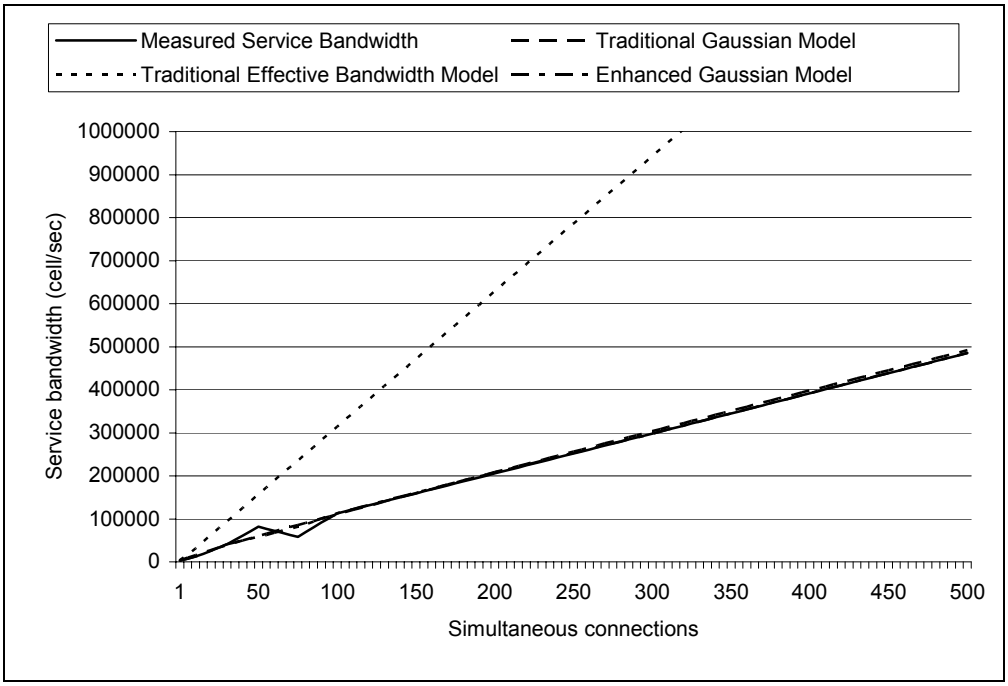


(b) Buffer size, $l = 500$.

Figure 5-1. Computed aggregate service bandwidth by the model-based CAC approaches with static number of active connections made up of homogeneous NT traffic streams.



(a) Buffer size, $l = 0$.



(b) Buffer size, $l = 500$.

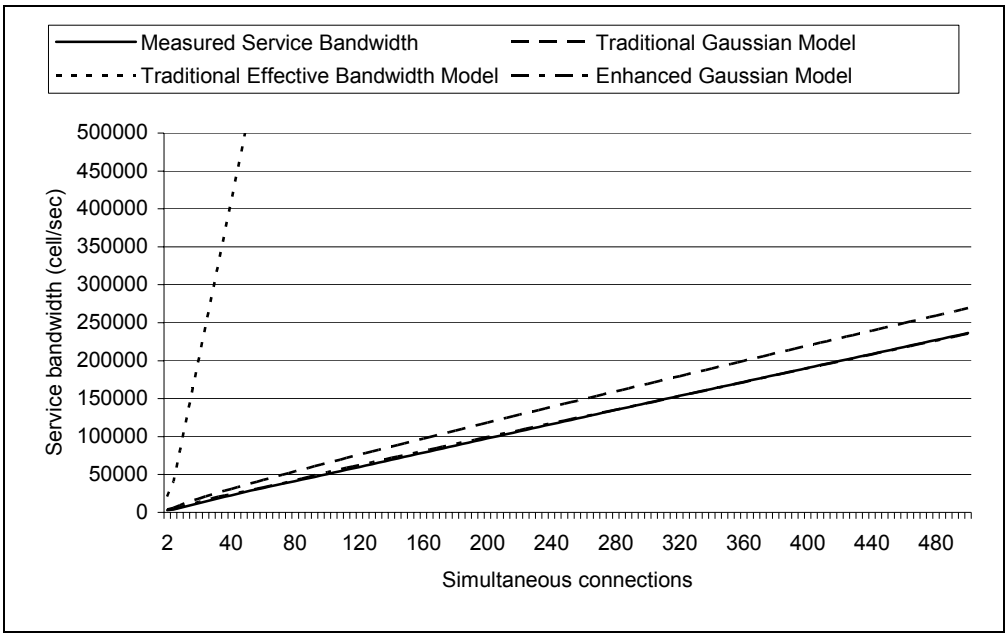
Figure 5-2. Computed aggregate service bandwidth by the model-based CAC approaches with static number of active connections made up of homogeneous VT traffic streams.

5.2.1.2 Heterogeneous Traffic Streams

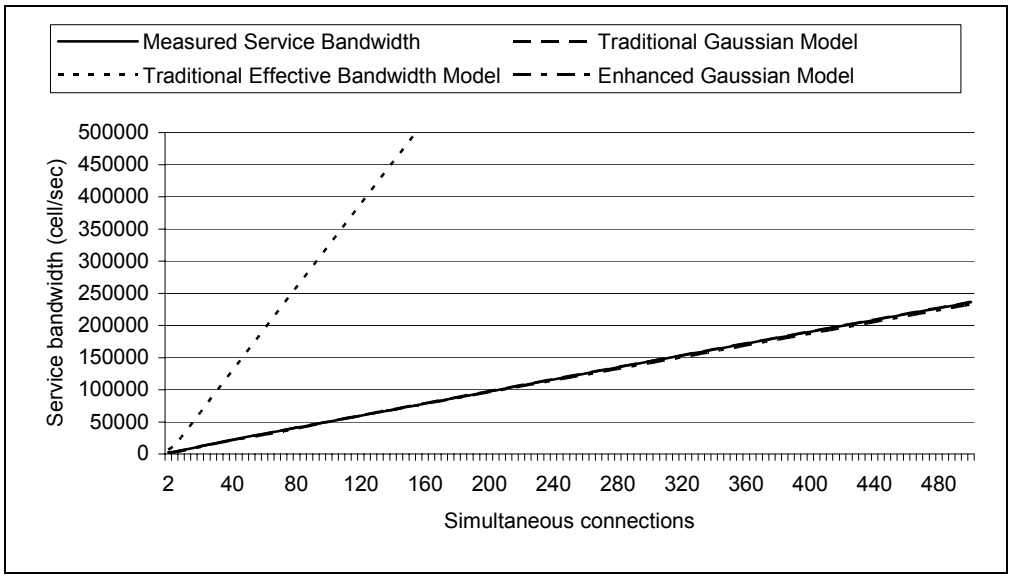
For heterogeneous traffic streams, the figures show the computed *S* value required by equal number of NT and VT connections, i.e., if there are 10 simultaneous connections in the link, five of them are NT traffic connections, while the remainders are VT traffic connections.

Figure 5-3(a) and (b) show the computed aggregate service bandwidth *S* plotted against equal number of NT and VT connections. Notice that the CAC approaches' performance patterns are almost identical to those shown in Figure 5-2(a) and (b) for VT homogeneous traffic streams. This is because VT (μ = 882.9 cell/sec) is many times larger than NT (μ = 8 cell/sec) in term of traffic load. Hence, at the SSQ bottleneck, the VT connections are viewed as the dominant users of the link.

From these figures, the eGA approach is observed to match the measured service bandwidth closely.



(a) Buffer size, $l = 0$.



(b) Buffer size, $l = 500$.

Figure 5-3. Computed aggregate service bandwidth by the model-based CAC approaches with static number of active connections made up of heterogeneous traffic streams. NT and VT are used by equal number of connections, i.e., 50-50 % share.

5.2.1.3 Study Conclusion

Based on the results of the performance studies investigating the accuracy of the model-based CAC algorithms under both homogeneous and heterogeneous traffic scenarios, we observed that:

- The GA approach over-estimates the S values for $l = 0$ case. However, for the $l = 500$ case, this approach computes quite accurate S values.
- The EB approach grossly over-estimates the S values.
- The eGA approach computes fairly accurate S values that are close to the measured service bandwidth values.

5.2.2 Model-based CAC Approaches

In this section, the performance of the traditional Gaussian (GA) and Effective Bandwidth (EB), and the enhanced Gaussian (eGA) CAC approaches are studied.

5.2.2.1 Homogeneous Traffic Streams

From Table 5-1, the most efficient model-based CAC approach is eGA because it consistently permits a large number of simultaneous connections that is very close to the maximum allowable number (see Table 5-2, whose values are derived from both Figure 5-1 and Figure 5-2), whilst still meeting the aggregate QoS requirement.

Traffic	PQ	*I*	GA	EB	eGA
NT	Util. (%)	0	6.2	1.5	61.9
		500	22.9	4.1	73.6
	CLR	0	0	0	4.9e-5
		500	0	0	2.6e-5
	MSC	0	12	3	120.8
		500	44.7	8	143.3
	MBR (%)	0	91.6	96.7	30.3
		500	74	93.9	17.5
	MSB	0	1567.1	1665.4	1514.5
		500	1532.7	1624.2	1492.9
VT	Util. (%)	0	48.3	9.3	69.5
		500	76.7	27.6	76.3
	CLR	0	0	0	6.1e-6
		500	2e-5	0	1.5e-5
	MSC	0	77.5	14.9	111.3
		500	122.8	44.5	122.1
	MBR (%)	0	49.5	90	27.4
		500	19.8	71.6	20.4
	MSB	0	5845.5	5903.6	5727.3
		500	5601.3	5963.5	5642

Table 5-1. Performance quantities – Model-based CAC approaches with homogeneous traffic streams.

Traffic	SSQ server rate		*I*	Maximum number of simultaneous connections before QoS breach
	bit/sec	cell/sec		
NT	650 K	1533	0	125
			500	150
VT	60 M	141.5 K	0	117
			500	130

Table 5-2. Maximum number of simultaneous connections before QoS breach – Model-based CAC approaches with homogeneous traffic streams.

5.2.2.2 Heterogeneous Traffic Streams

Here, we study the performance of model-based CAC approaches for heterogeneous traffic streams, where the aggregate flow is made up of two different types of traffic.

From Table 5-3, the eGA approach performs better than the other two traditional approaches.

Table 5-4 values are taken from Figure 5-3(a) and (b). Note that for $l = 0$ case, the minimum service bandwidth required to serve one VT connection, subject to meeting QoS, is enough to serve 6.7 NT connections. Thus, even though the maximum allowable connections are 142, i.e., 71 NT and 71 VT, the eGA approach is able to admit 154.9, i.e., 89.3 NT and 65.6 VT, connections instead. This is made possible because the eGA approach admits fewer VT connections, and the spare bandwidth is then channeled to service additional less bandwidth-hungry NT connections. For $l = 500$ case, the bandwidth for one VT connection can instead be used to service 13.2 NT connections.

TS	PQ	*I*	GA	EB	eGA
NT and VT	Util. (%)	0	70.8	19.3	81.6
		500	76.7	29.4	86.6
	CLR	0	0	0	1e-6
		500	0	0	7.5e-5
	MSC - NT	0	85.7	67.1	86.3
		500	84.9	78.9	86.6
	MSC - VT	0	56.4	15	65.1
		500	61.2	23	69.2
	MBR (%) - NT	0	1.5	22.3	0.7
		500	1.7	9.2	0.5
	MBR (%) - VT	0	33.5	80.5	23.9
		500	28.3	71.1	18.9
	MSB	0	68822.7	70700.2	67683.1
		500	68392.6	70242.9	67181.9

Table 5-3. Performance quantities – Model-based CAC approaches with heterogeneous traffic streams.

TS	SSQ server rate		*I*	Maximum number of simultaneous connections before QoS breach
	bit/sec	cell/sec		
NT and VT	30 M	70.75 K	0	142
			500	143

Table 5-4. Maximum number of simultaneous connections before QoS breach – Model-based CAC approaches with heterogeneous traffic streams. NT and VT are used by equal number of connections, i.e., 50-50 % share.

5.2.2.3 Study Conclusion

Based on the results of the performance studies for the three model-based CAC approaches under both homogeneous and heterogeneous traffic scenarios, the most efficient CAC approach is:

- The eGA approach.

5.2.3 Measurement-based Counterparts

In this section, the performance of the MBCAC approaches, i.e., measurement-based counterparts of the traditional Gaussian and the enhanced Gaussian, are studied; and they are denoted as m-GA and m-eGA respectively. Each MBCAC approach has two alternative implementations, i.e., with and without the AFCM proposed in section 3.4.

These MBCAC approaches conduct online measurements of the mean aggregate traffic and also estimates the performance margin *m* parameter. For approaches using the AFCM, an additional control layer is provided to adjust the level of contributions by the performance margin *m* towards the aggregate service bandwidth estimation, hence altering the admission decision behavior between conservative and daring.

The performance studies conducted here aim to investigate how reliable these MBCAC approaches are in making efficient admission decisions when given minimal a-priori traffic information.

For MBCAC approaches using the AFCM, we investigate the efficiency of these approaches using combinations of:

- Prudence level policy module – Only Adaptive Weight Feedback (AWF) method.
- Load and traffic measurements module – (1) Link Occupancy (LO), (2) Buffer Occupancy (BO), (3) Cell Loss Conservative Period (CLCP).

These combinations are based on the AFCM simulation settings mentioned in section 4.4.4.1.

5.2.3.1 Homogeneous Traffic Streams

The performance of the m-GA and m-eGA approaches with and without the AFCM are reported for homogeneous traffic streams with buffer sizes of 0 and 500 cells in Table 5-5 and Table 5-6 respectively. The fourth column lists the performance quantities without the AFCM, while columns five to ten list the MBCAC approaches that use one of the AFCM techniques.

From Table 5-5, for l = 0 case, the m-GA approach with AWF and LO=80% techniques performs better than both GA and m-GA without AFCM approaches for NT connections. This is because by setting a LO threshold that is higher than the average link utilization, the adaptive prudence level factor will be generally close to 1.0 value, thus resulting in the computed aggregate service bandwidth S to be closer to a value almost equal to *only* the aggregate mean work submitted by both the new and established connections. This is illustrated by the recorded MWei value hovering around 0.8. For l = 500 case, the approach with BO=10% technique returns good performance.

For VT connections, the m-GA approach with AWF and LO=90% techniques is efficient for l = 0 case. However, for l = 500 case, the m-GA approach with AWF and BO=20% techniques gives the best performance.

From Table 5-6, for NT connections, the m-eGA with AWF and LO=70% techniques is the better approach. When higher link occupancy threshold values are used, e.g., 80% and 90%, the results show that these approaches are not able to meet the aggregate QoS requirement. For VT connections, the m-eGA approach with AWF and LO=90% techniques is again the better approach for l = 0 case. However, for l = 500 case, the better approach is again the m-eGA approach with AWF and BO=20% techniques.

TS	PQ	*I*	m-GA	m-GA, LO=80%	m-GA, LO=90%	m-GA, BO=10%	m-GA, BO=20%	m-GA, CLCP=5	m-GA, CLCP=30
NT	Util. (%)	0	0.8	57	62	-	-	47.1	21.7
		500	8.6	62.5	68.3	73.5	74.6	59.9	37.4
	CLR	0	0	4.8e-5	1.6e-4	-	-	1.9e-4	1.7e-4
		500	0	8.1e-7	4.9e-4	9e-5	1.2e-4	1.5e-4	1.3e-4
	MSC	0	1.6	111.1	120.8	-	-	91.4	42.1
		500	16.7	121.4	133.1	142.7	145.1	116.6	72.6
	MBR (%)	0	97.2	36.2	31.7	-	-	46.3	74.5
		500	88.8	30.5	24.6	19.1	17.5	32.5	58
	MWei	0	-	0.82	0.84	-	-	0.58	0.26
		500	-	0.82	0.87	0.9	0.91	0.7	0.43
	MSB	0	1796.8	2058.4	2069.8	-	-	5560.3	15655.6
		500	1558.2	1459.7	1416.1	1420.8	1412.4	2338.9	5960.8
VT	Util. (%)	0	40.2	66.4	71.1	-	-	34.1	55.9
		500	68.1	68.7	74	79.6	80.2	68.7	64.9
	CLR	0	0	0	5e-6	-	-	5.6e-6	3.9e-7
		500	0	0	8.5e-8	1.9e-5	2.4e-5	4.7e-6	6.6e-6
	MSC	0	64.5	105.8	113.4	-	-	54.8	89.8
		500	109.1	109.6	118.3	127.2	128.1	110	103.9
	MBR (%)	0	58.4	30.4	25.4	-	-	64.6	41.6
		500	28.1	28.6	23.5	17.4	16.7	29.3	32.1
	MWei	0	-	0.66	0.85	-	-	0.36	0.1
		500	-	0.5	0.73	0.86	0.88	0.35	0.16
	MSB	0	5858.4	4884.2	5246.9	-	-	9364.9	5622.7
		500	5712.8	5017.5	4903.1	5020.6	5015.6	5339.9	5519.6

Table 5-5. Performance quantities – m-GA approaches with and without the AFCM for homogeneous traffic streams.

TS	PQ	*l*	m-eGA	m-eGA, LO=70%	m-eGA, LO=80%	m-eGA, LO=90%	m-eGA, BO=10%	m-eGA, BO=20%	m-eGA, CLCP=5	m-eGA, CLCP=30
NT	Util. (%)	0	59.7	62.2	65.7	69.5	-	-	48.4	26.8
		500	71	71.4	72.6	75.7	79.1	79.5	74.3	69.1
	CLR	0	3.2e-5	4.6e-5	1.5e-4	4e-4	-	-	2.4e-4	1.3e-4
		500	1.3e-5	1.1e-5	2e-5	5e-5	2.1e-4	2.5e-4	2.2e-4	9.1e-5
	MSC	0	116.6	121.1	127.8	135.3	-	-	94	52.5
		500	138.4	139	141.3	147.3	154	154.8	144.9	134.6
	MBR (%)	0	32.8	29.8	26.3	23.1	-	-	45.4	69
		500	20.5	19.9	18.8	15.7	12.2	11.7	17	22.6
	MWei	0	-	0.21	0.47	0.65	-	-	0.52	0.26
		500	-	0.04	0.26	0.58	0.8	0.82	0.55	0.19
	MSB	0	1517.4	1453	1377	1333.6	-	-	1492.6	2407.2
		500	1501.3	1487.4	1428.6	1353.1	1316.7	1312.8	1369	1454.3
VT	Util. (%)	0	59.1	61.5	66.4	72.5	-	-	59.9	55.9
		500	66.6	66.7	68.7	73.9	79.6	80.1	67.7	64.9
	CLR	0	0	0	0	1e-5	-	-	3.4e-6	3.9e-7
		500	0	0	0	2.8e-7	1.9e-5	2.6e-5	7e-6	6.6e-6
	MSC	0	94.8	98.5	105.8	116	-	-	95.9	89.8
		500	106.6	106.3	109.6	118.1	127.2	127.9	108.5	103.9
	MBR (%)	0	38.5	36.6	30.4	25.1	-	-	36.8	41.6
		500	29.6	30.2	28.6	23.8	17.4	18.3	28.1	32.1
	MWei	0	-	0.4	0.66	0.79	-	-	0.37	0.1
		500	-	0.12	0.5	0.75	0.86	0.86	0.44	0.16
	MSB	0	5762	5120.5	4884.2	4956.8	-	-	5255.2	5622.7
		500	5716.3	5555.9	5017.5	4908.2	5020.6	5054.9	5172.2	5519.6

Table 5-6. Performance quantities – m-eGA approaches with and without the AFCM for homogeneous traffic streams.

5.2.3.2 Heterogeneous Traffic Streams

From Table 5-7, the efficient MBCAC approach is the m-GA approach with AWF and LO=90% techniques. Even though minimal a-priori traffic information is provided, this approach is still able to achieve performance that is either slightly higher or on par with the traditional GA approach.

From Table 5-8, the m-eGA approach with AWF and LO=90% techniques is the most efficient approach, and the measured performance is only slightly lower than the eGA approach.

TS	PQ	*I*	m-GA	m-GA, LO=80%	m-GA, LO=90%	m-GA, BO=10%	m-GA, CLCP=5
NT and VT	Util. (%)	0	23.9	67.8	74.7	-	17.4
		500	33.9	69.2	75.7	80.7	17.1
	CLR	0	0	0	6.1e-7	-	5.3e-5
		500	0	0	5e-7	1.8e-4	5.3e-5
	MSC - NT	0	0.44	56.2	63.1	-	14
		500	0.44	58.1	64	68.8	14.4
	MSC - VT	0	19.3	54.3	59.8	-	14
		500	27.4	55.4	60.6	64.5	13.7
	MBR (%) - NT	0	95.6	34.2	27	-	81.5
		500	95.6	33.2	25.3	20.4	80.8
	MBR (%) - VT	0	76	36.3	29.2	-	83.2
		500	66.8	35.1	27.7	22.8	83
	MWei	0	-	0.81	0.85	-	0.16
		500	-	0.75	0.83	0.86	0.16
	MSB	0	112439.3	86512.5	84397.3	-	612787.9
		500	84727.5	71389.2	70334.4	72695	280745.2

Table 5-7. Performance quantities – m-GA approaches with and without the AFCM for heterogeneous traffic streams.

TS	PQ	*l*	m-eGA	m-eGA, LO=80%	m-eGA, LO=90%	m-eGA, BO=10%	m-eGA, CLCP=5
NT and VT	Util. (%)	0	31.3	72.9	78.6	-	18.6
		500	72	77.9	82.3	86.3	72.2
	CLR	0	0	0	4.3e-6	-	4.6e-5
		500	1.7e-6	3.3e-5	3.3e-5	3.9e-4	2.5e-5
	MSC - NT	0	23.9	59.8	65.1	-	15.5
		500	56.8	63.4	66.7	70	57.3
	MSC - VT	0	25.1	58.4	62.9	-	15
		500	57.7	62.4	65.9	69.1	57.8
	MBR (%) - NT	0	70.3	30.2	23.7	-	78.8
		500	32.8	26.5	21.5	18.4	32.3
	MBR (%) - VT	0	69.6	32.2	26	-	80.8
		500	32.1	27.2	23.4	20.4	31.8
	MWei	0	-	0.55	0.74	-	0.18
		500	-	0.25	0.54	0.73	0.07
	MSB	0	80104.3	66109.2	65410.5	-	421251.6
		500	70015.6	67857.3	66083.8	66031.3	69613

Table 5-8. Performance quantities – m-eGA approaches with and without the AFCM for heterogeneous traffic streams.

5.2.3.3 Study Conclusion

Based on the results of the performance studies for the three model-based CAC approaches under both homogeneous and heterogeneous traffic scenarios, the most efficient CAC approach is:

- The m-eGA approach with AWF and LO techniques.

5.2.4 Parameter *q(n)*: Accuracy Issue

The performance studies reported in the previous sections are for CAC approaches using a-priori traffic information derived from measurements taken of the traffic source prior to connection set-up. In other words, traffic

parameters specific to a source belonging to a particular type of traffic, are all measured prior to connection admittance. These traffic parameters are: the peak, mean and variance of the arrival traffic load, and the performance margin's multiple factor *q(n)* values. Note that for the MBCAC approaches, the only a-priori traffic parameter used is the performance margin's multiple factor *q(n)* values; while the remaining parameters are all measured real-time.

It is relatively easy to measure traffic parameters such as the peak, mean and variance of the arrival traffic load, from every traffic source. However, for the performance margin's multiple factor *q(n)* values, such per-flow measurements are operationally expensive.

Hence, this pose an interesting question, and that is if a connection is declared to be transmitting work belonging to a particular traffic genre, can a set of *q(n)* values previously measured from another traffic source that belongs to the same traffic genre, be used? If the answer is "Yes", it would imply that this set of *q(n)* values can be viewed as the *standard set of values* for that traffic genre, and hence applicable for use on connections with the same traffic type but different arrival traffic patterns.

The performance studies reported in the previous sections 5.2.2 and 5.2.3 use two sets of a-priori *q(n)* values measured from the NT trace, which is a 12 hours long network data traffic, and the VT trace, which is an MPEG-1 encoding of the 'Star Wars' movie by Garrett et al. [GF94]. These *q(n)* values are considered the two primary set of values. The CAC and MBCAC approaches will use these primary sets to help compute the aggregate service bandwidth required by connections whose traffic belong to a particular type. The first primary set of *q(n)* values will be used for connections that only transmit network data type of traffic, while the second set will be used for connections with video type of traffic. For example, the set of *q(n)* values measured from the VT trace will be used to help compute the statistical behavior of all connections that transmit work belonging to the same traffic genre, i.e., video traffic type.

In this section, results from performance studies investigating this *q(n)* accuracy issue are reported for the following CAC approaches: GA, eGA, and their measurement-based counterparts with and without the AFCM. In these studies, two real traces, i.e., NT2 and VT2, are used.

5.2.4.1 Homogeneous Traffic Streams

Table 5-9 and Table 5-10 list the results for the GA based CAC and MBCAC approaches, while Table 5-11 and Table 5-12 list for the eGA based CAC and MBCAC approaches.

We begin with the results for the GA based CAC and MBCAC approaches. For the very bursty NT2 connections (shown in Table 5-9), using the set of *q(n)* values belonging to the less bursty NT trace results in poor performance from all GA based approaches. Even though the m-GA approach with AWF and LO=60% techniques is able to meet the desired aggregate QoS, it is not efficient in utilizing the link. Note that all m-GA approaches with AWF and BO techniques are unable to meet the desired aggregate QoS. For the VT2 connections, because of its mildly bursty traffic, the use of a set of *q(n)* values belonging to the more bursty VT trace results in minimal adverse effect on the CAC performance. In fact, there is an overall improvement in the performance of all the listed CAC and MBCAC approaches in Table 5-10. For $l = 0$ case, the m-GA approach with AWF and LO=90% techniques returns good performance; while for the $l = 500$ case, the m-GA approach with AWF and BO=20% techniques is the efficient CAC approach.

Next, we summarize the results for the eGA based CAC and MBCAC approaches. For the NT2 connections, Table 5-11 shows that none of the eGA based approaches are able to meet the desired aggregate QoS. For the VT2 connections, Table 5-12 shows that for the $l = 0$ case, the m-eGA approach using the AWF and LO=90% techniques return good performance; while for the $l = 500$ case, it is the use of the AWF and BO=20% techniques.

TS	PQ	*I*	GA	m-GA	m-GA, LO=60%	m-GA, LO=70%	m-GA, CLCP=30
NT2	Util. (%)	0	3	1.1	40.3	45.8	15.7
		500	12.9	1.9	43.8	49.1	22
	CLR	0	0	0	6.1e-5	1.3e-4	4e-4
		500	0	0	1.8e-5	7.5e-5	5.2e-4
	MSC	0	7	2.6	95.9	109.2	37.2
		500	30.8	4.6	104.4	116.8	52.6
	MBR (%)	0	94.4	96.4	44	37.3	76.9
		500	81.3	95.7	39.6	34.1	68.9
	MWei	0	-	-	0.73	0.79	0.23
		500	-	-	0.69	0.76	0.31
	MSB	0	2120.9	2223.2	3234.6	2942.8	18759.4
		500	2012.9	2153.6	2080.8	1985.5	8241.5

Table 5-9. Performance quantities – GA based CAC and MBCAC approaches with and without the AFCM for homogeneous NT2 traffic streams.

TS	PQ	*I*	GA	m-GA	m-GA, LO=90%	m-GA, BO=20%	m-GA, CLCP=5
VT2	Util. (%)	0	54.5	45.3	74.2	-	53.2
		500	80	70.6	76.9	83.5	74.3
	CLR	0	0	0	1.8e-6	-	7.5e-6
		500	5.7e-7	0	0	2.3e-6	1.3e-5
	MSC	0	83.3	69.1	113.4	-	81.4
		500	122.2	107.9	117.4	127.6	113.5
	MBR (%)	0	45.8	55.6	25.9	-	45.5
		500	20.2	28.9	25.1	17.1	24.7
	MWei	0	-	-	0.84	-	0.56
		500	-	-	0.74	0.88	0.54
	MSB	0	3880.7	3901.4	3606	-	4838.1
		500	3750.6	3813.3	3358.5	3456.6	3496.6

Table 5-10. Performance quantities – GA based CAC and MBCAC approaches with and without the AFCM for homogeneous VT2 traffic streams.

TS	PQ	*I*	eGA	m-eGA	m-eGA, LO=50%	m-eGA, CLCP=5
NT2	Util. (%)	0	49.8	47.2	50.5	37.1
		500	63.7	61.7	61.8	64.1
	CLR	0	1.9e-4	1.2e-4	1.9e-4	6.7e-4
		500	9e-4	6.4e-4	6.3e-4	1.1e-3
	MSC	0	118.9	112.7	120.5	88.9
		500	152.2	147.4	147.5	153.4
	MBR (%)	0	31.3	34.8	29.6	48.5
		500	12.9	15.6	15.6	12
	MWei	0	-	-	0.09	0.51
		500	-	-	0.01	0.46
	MSB	0	1988.9	1987	1932.8	2115.3
		500	1937.2	1951.4	1948	1686.5

Table 5-11. Performance quantities – eGA based CAC and MBCAC approaches with and without the AFCM for homogeneous NT2 traffic streams.

TS	PQ	*I*	eGA	m-eGA	m-eGA, LO=90%	m-eGA, BO=20%	m-eGA, CLCP=5
VT2	Util. (%)	0	73.5	62.9	74.4	-	64
		500	79.5	69.3	76.7	83	71.5
	CLR	0	1.3e-5	0	3.4e-6	-	1e-5
		500	2.7e-7	0	0	6.4e-6	1.3e-5
	MSC	0	112.3	96.1	113.8	-	97.8
		500	121.4	105.9	117.2	126.9	109.3
	MBR (%)	0	26.9	37.1	24.5	-	34.4
		500	20.7	30.2	24	16.6	28
	MWei	0	-	-	0.79	-	0.4
		500	-	-	0.74	0.89	0.46
	MSB	0	3808.6	3839.7	3325.5	-	3513.9
		500	3769.4	3814.4	3366.1	3433	3539.5

Table 5-12. Performance quantities – eGA based CAC and MBCAC approaches with and without the AFCM for homogeneous VT2 traffic streams.

5.2.4.2 Heterogeneous Traffic Streams

For the NT2 and VT2 connections, Table 5-13 shows that amongst the GA based CAC and MBCAC approaches, the most efficient approach is the m-GA approach with AWF and LO=90% techniques. From Table 5-14, the most efficient approach amongst the eGA based CAC and MBCAC approaches is the eGA approach, and followed closely by the m-eGA approach with AWF and LO=90% techniques.

As shown from the tables, the approaches generally return respectable results even with NT2 connections in the aggregate flow. This is because the aggregate flow's traffic is dominated by work submitted from the VT2 connections.

TS	PQ	*I*	GA	m-GA	m-GA, LO=90%	m-GA, BO=10%	m-GA, CLCP=5
NT2 and VT2	Util. (%)	0	65.9	23.9	74.9	-	20.9
		500	75.5	30.7	76.5	81.1	19.1
	CLR	0	0	0	6.8e-7	-	5.2e-5
		500	0	0	1.3e-6	2.3e-4	1e-4
	MSC – NT2	0	83.3	0.34	59.2	-	15.8
		500	85.4	0.35	60.6	65	15.4
	MSC – VT2	0	49.7	18.5	56.9	-	15.9
		500	57.1	23.7	58.2	61.6	14.6
	MBR (%) – NT2	0	3.7	95.7	30.9	-	78.6
		500	1.8	95.6	29.4	25.4	79.3
	MBR (%) – VT2	0	41.7	77.3	33.1	-	80.4
		500	33.1	71.4	31.7	27.5	81.3
	MWei	0	-	-	0.83	-	0.19
		500	-	-	0.8	0.83	0.18
	MSB	0	46276.3	75023.2	58339.1	-	394850.4
		500	45888.6	57962.7	48235.4	50309.2	189302.3

Table 5-13. Performance quantities – GA based CAC and MBCAC approaches with and without the AFCM for heterogeneous NT2 and VT2 traffic streams.

TS	PQ	*I*	eGA	m-eGA	m-eGA, LO=90%	m-eGA, BO=10%	m-eGA, CLCP=5
NT2 and VT2	Util. (%)	0	82.5	11	78.7	-	18.6
		500	87.1	73.6	82.9	86.6	72.5
	CLR	0	1.1e-8	0	6e-6	-	9.8e-5
		500	2.1e-5	3e-5	2.2e-5	3.8e-4	2.5e-5
	MSC – NT2	0	86.5	7.9	63.5	-	14.8
		500	86.2	54.7	64.9	68.2	54.1
	MSC – VT2	0	62.5	8.4	59.8	-	14.1
		500	66	56	63.1	65.9	55.2
	MBR (%) – NT2	0	1	87.9	27	-	80.2
		500	1	35.5	23.8	20.8	36.8
	MBR (%) – VT2	0	26.7	88.3	29.1	-	82.1
		500	23.1	35.2	25.9	22.7	36
	MWei	0	-	-	0.71	-	0.17
		500	-	-	0.49	0.71	0.06
	MSB	0	45594.2	158428.2	44474.3	-	283531.9
		500	45083.3	46748.7	44732.8	44589.9	46593

Table 5-14. Performance quantities – eGA based CAC and MBCAC approaches with and without the AFCM for heterogeneous NT2 and VT2 traffic streams.

5.2.4.3 Study Conclusion

Amongst the GA and eGA based CAC and MBCAC approaches studied here, except for the homogeneous NT2 connections, the m-eGA approach with AWF and LO techniques consistently produced good results for both homogeneous and heterogeneous traffic streams.

Based on the results from these studies, the primary sets of performance margin's multiple factor *q(n)* values are still usable as long as the submitted work is not 'more bursty' than the primary traffic source's burstiness rate.

5.2.5 Framework Conclusion

In this section, the most efficient CAC and MBCAC approaches within the model-based framework for both homogeneous and heterogeneous traffic streams are collated and listed here. We term an efficient CAC approach to be one that ensures high link utilization whilst still maintaining the requested QoS.

Table 5-15 lists the most efficient CAC approaches when the a-priori traffic parameter – performance margin's multiple factor *q(n)* values are provided. Amongst the model-based CAC approaches listed in this table, the eGA approach performs consistently well. This is expected since it has intimate a-priori traffic information that is matched to the connections' arrival traffic. For the MBCAC case, the m-eGA approach with AWF and LO techniques produces consistent performance. For the VT connections in a buffered (l = 500) SSQ case, the use of the liberal BO technique generally results in higher link utilization performance.

Table 5-16 lists the most efficient CAC approach when minimal a-priori traffic information is provided. Even though these approaches use the traffic parameter *q(n)* to help compute the aggregate service bandwidth, these *q(n)* values are actually derived from other similar-in-traffic-type traffic sources. Because of this, some approaches did not return good performance results. This is especially the case for the very bursty NT2 connections. However for the mildly bursty VT2 connections, the recorded performance results are very promising. The most efficient CAC approaches are: the eGA approach for model-based CAC; and the m-eGA approach with AWF and LO techniques for MBCAC.

Based on the performance studies of the CAC and MBCAC approaches within the model-based framework, the approaches that produce consistently good performance outputs are:

- For model-based CAC: The eGA approach.

- For MBCAC: The m-eGA approach with AWF and LO techniques.

TS	*I*	CAC	Util. (%)	MBCAC	Util. (%)
NT	0	eGA	61.9	m-eGA, AWF, LO=70%	62.2
	500		73.6	m-GA, AWF, BO=10%	73.5
VT	0	eGA	69.5	m-eGA, AWF, LO=90%	72.5
	500		76.3	m-GA, AWF, BO=20%	80.2
NT and VT	0	eGA	81.6	m-eGA, AWF, LO=90%	78.6
	500		86.6		82.3

Table 5-15. Most efficient CAC performance – Model-based framework.

TS	*I*	CAC	Util. (%)	MBCAC	Util. (%)
NT2	0	None	-	m-GA, AWF, LO=60%	40.3
	500		-		43.8
VT2	0	eGA	73.5	m-eGA, AWF, LO=90%	74.4
	500		79.5	m-GA, AWF, BO=20%	83.5
NT2 and VT2	0	eGA	82.5	m-eGA, AWF, LO=90%	78.7
	500		87.1		82.9

Table 5-16. Most efficient CAC performance – Model-based framework using sets of *q(n)* values taken from other similar-in-traffic-type traffic sources.

5.3 Histogram-based CAC Framework: Performance Issues

The histogram-based framework consists of a variety of modules that are used to create MBCAC approaches. These modules also include the adaptive feedback controller – AFCM. By 'mixing and matching' techniques taken from every module, an MBCAC scheme can be constructed specially for use in a network with certain traffic control demands.

Other than the easy-to-define traffic parameter – peak rate, no other a-priori traffic information is required from the users before the connection set-up phase. Instead, during the admission decision process, the MBCAC approaches will use relevant traffic statistics measured real-time from the aggregate traffic arrival to predict future bandwidth consumption by the established connections.

Empirical-based performance studies of these MBCAC approaches are reported in the following sections:

- Section 5.3.1 – A new connection is admitted if and only if its user-declared peak rate is less than the link's 'Available bandwidth'. This spare/available bandwidth is derived by subtracting the estimated bandwidth required to service established connections whilst meeting QoS from the SSQ server rate. Three methods (see section 3.3.2 for further details) are used by the MBCAC approaches to estimate the minimum service bandwidth, and these methods are: (1) Peak Rate Allocation (PRA), (2) Rate Envelope Multiplexing (REM), and (3) Rate Sharing (RS). The first method is very conservative, while the latter two methods are daring. This section reports on the accuracy of these cell loss approximation methods.

- Section 5.3.2 – The amount of aggregate traffic that arrives at a SSQ bottleneck during a fixed time-slot is recorded and stored into a traffic

histogram unique to this particular time-scale. In the histogram-based framework, traffic loads are measured across different time-scales and then stored in different traffic histograms. These histograms are central to this framework because the liberal REM/RS method uses the recorded past traffic arrival loads during the admission decision process. This would imply that the accuracy of the traffic histograms' records will affect the service bandwidth value computed by the REM/RS algorithm. Hence, should the traffic histograms update its past traffic load records whenever a connection departs from the network? If the answer is "Yes", then the next question to ask is how accurate this update procedure must be to ensure better MBCAC performance without incurring huge computation and storage expenses. We answer these questions through the performance studies reported in this section.

- Section 5.3.3 – The performance studies reported in this section address the issue of removing the cell loss rate L constraint from the REM/RS algorithm. In all the performance studies, the aggregate QoS agreed upon during connection set-up phase is fixed at CLR = 10^{-4} cells. By applying this CLR value into its algorithm, the REM/RS method aims to estimate a service bandwidth value that will adequately meet this QoS requirement. However, if the algorithm's L constraint is relaxed until L = 1 cell, will this result in the aggregate QoS agreement being breached? This is a possible scenario given that more new connections will be permitted into the link since the computed service bandwidth is grossly under-estimated.

- Section 5.3.4 – This histogram-based framework includes the same AFCM used by the model-based framework. This adaptive feedback controller, made up of 'prudence level policy' and 'load and traffic measurements' modules, is used to ensure the MBCAC approaches have an additional control level to prevent QoS breach. In addition,

the AFCM adapts the admission decision process to changing traffic arrival load conditions. If the measured traffic load is low, the AFCM will in-directly cause the computed service bandwidth to be smaller in value so as to permit more new connections to increase link utilization. We conduct and report on the performance studies of the effectiveness of using AFCM during the admission decision process.

- Section 5.3.5 – As mentioned earlier, the novelty of this histogram-based framework is that it contains MBCAC approaches that are customizable to the network provider's traffic control requirements. This is achieved by 'mixing and matching' different techniques contained in this CAC framework. In this section, we report on the performance of various MBCAC approaches using different combinations of technique and different AFCM settings, in order to find the best set of techniques and AFCM settings that will give the most efficient CAC performance.

- Section 5.3.6 – Here, we list the MBCAC schemes that consistently produce promising performance results.

5.3.1 Service Bandwidth *S*: Accuracy Issue

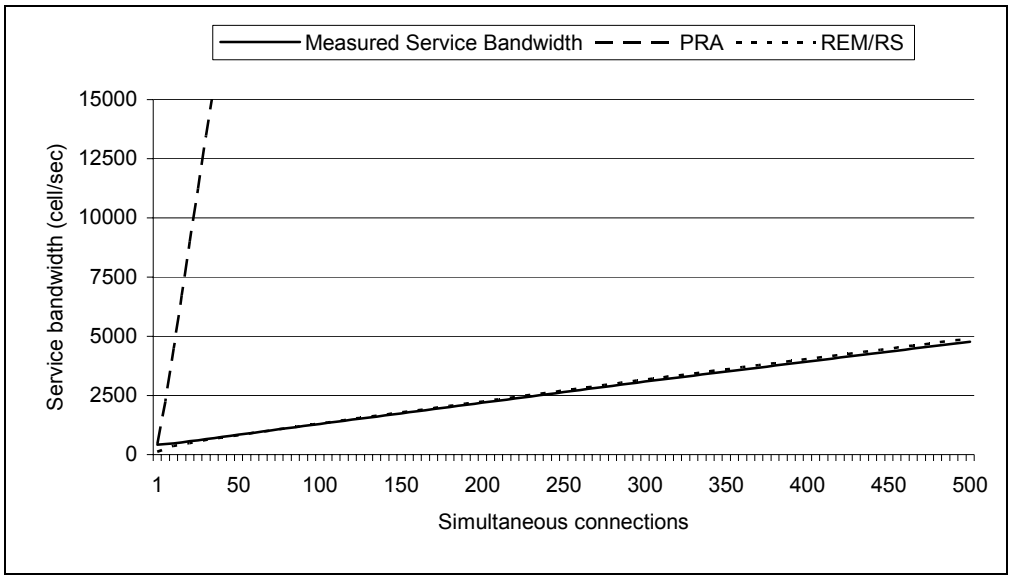
In this section, we report on the performance studies to investigate how accurate the PRA and the REM/RS methods are in estimating the aggregate service bandwidth *S* required by static number of active connections whilst meeting the desired QoS requirement.

All figures in this section show the computed *S* value plotted against fixed number of simultaneous connections for buffer sizes of 0 and 500 cells. This trace is used with random starting point, and wrapped around to the beginning when the end-of-the-trace is reached. All connections are active from the start of the simulation run, and their holding times are equal to the complete trace's duration.

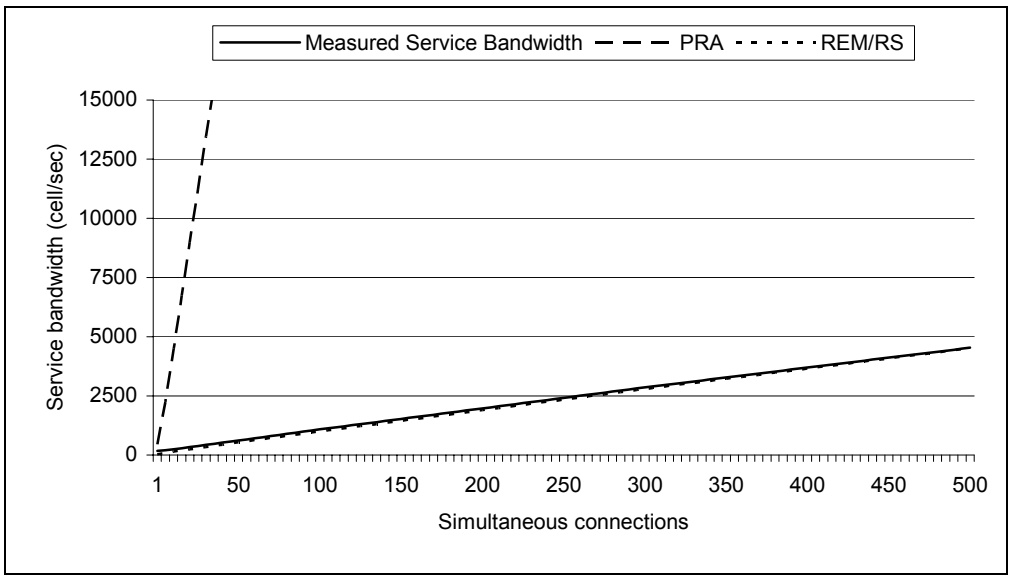
5.3.1.1 Homogeneous Traffic Streams

From Figure 5-4(a) and (b), and Figure 5-5(a) and (b), the REM/RS method produces results that closely match the measured service bandwidth. These figures clearly illustrate the algorithm does not grossly over or under estimate the required service bandwidth.

For the PRA method, it grossly over-estimates the service bandwidth. The poor CAC performance is expected since this method considers all established connections are transmitting at their user-declared peak rates.

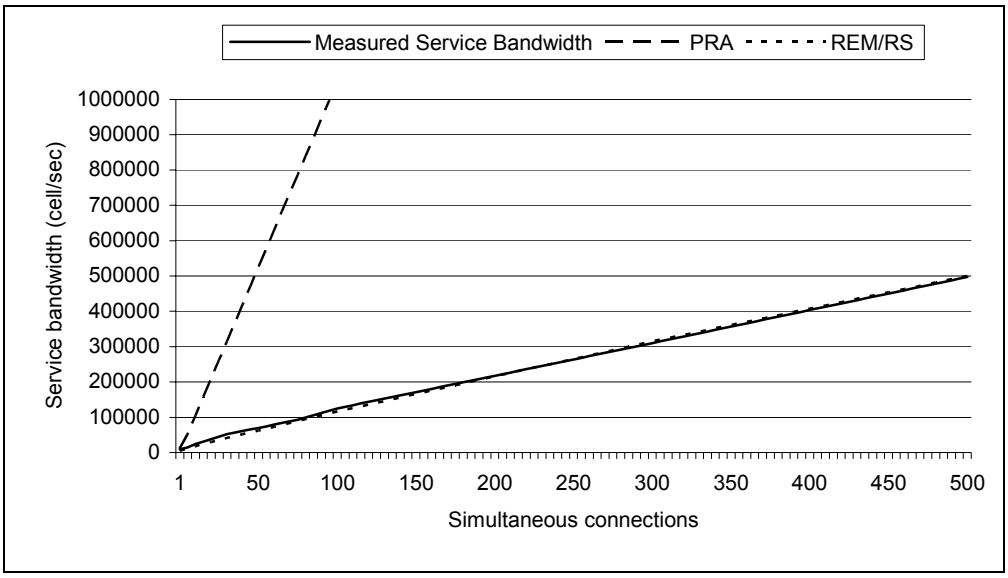


(a) Buffer size, $l = 0$.

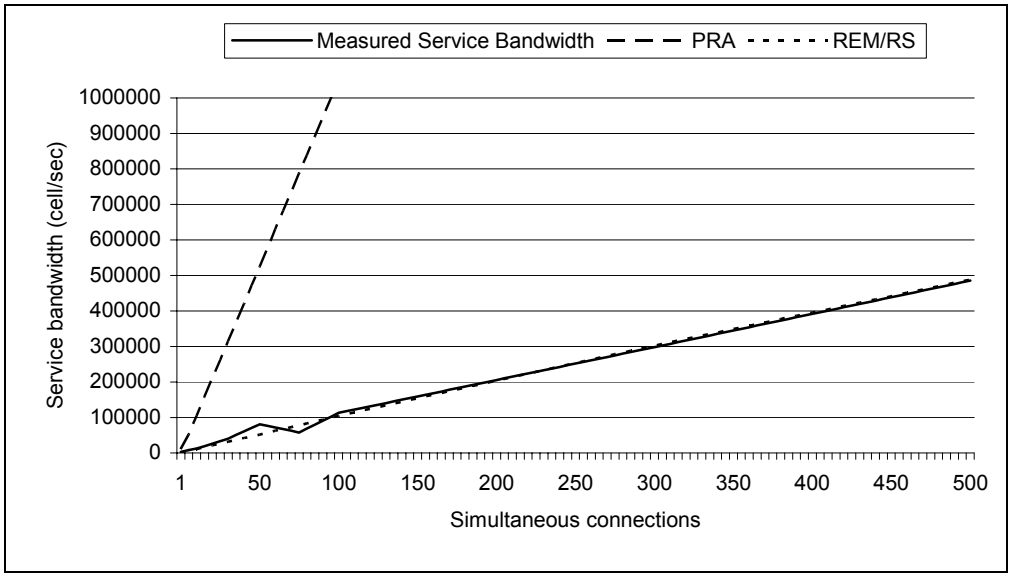


(b) Buffer size, $l = 500$.

Figure 5-4. Computed aggregate service bandwidth by the PRA, and the REM/RS methods with static number of active connections made up of homogeneous NT traffic streams.



(a) Buffer size, $l = 0$.



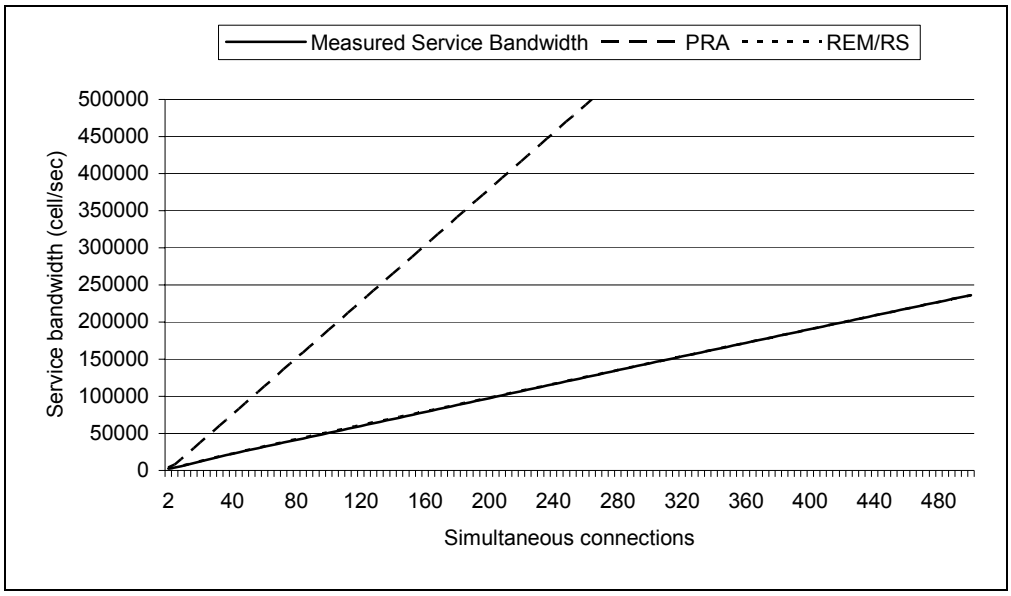
(b) Buffer size, $l = 500$.

Figure 5-5. Computed aggregate service bandwidth by the PRA, and the REM/RS methods with static number of active connections made up of homogeneous VT traffic streams.

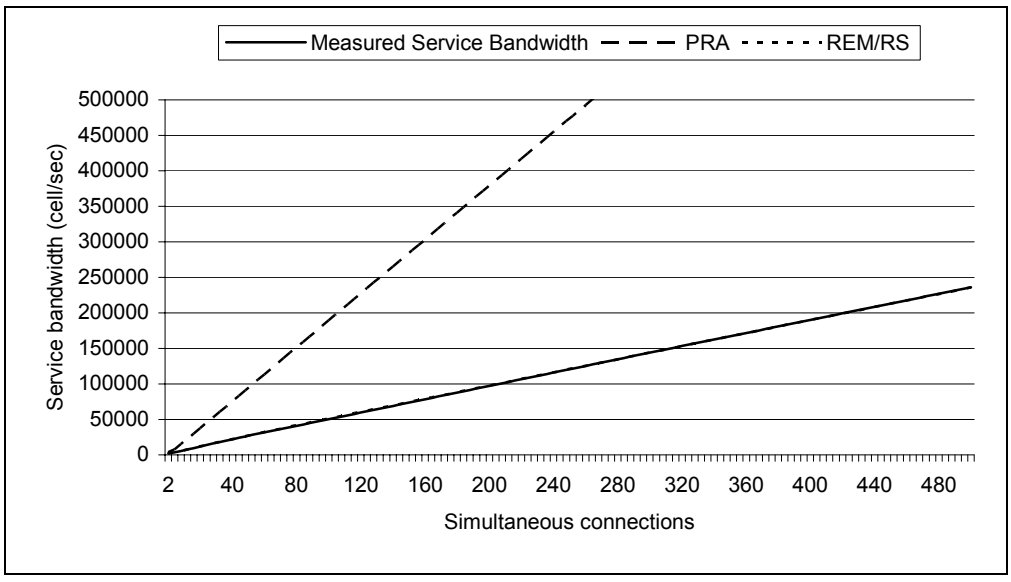
5.3.1.2 Heterogeneous Traffic Streams

For the heterogeneous traffic streams, the figures show the computed *S* value required by equal number of NT and VT connections, i.e., if there are 10 simultaneous connections in the link, five of them are NT traffic connections, while the remainders are VT traffic connections.

From Figure 5-6(a) and (b), the REM/RS method is consistent in computing fairly accurate estimates of the required service bandwidth.



(a) Buffer size, $l = 0$.



(b) Buffer size, $l = 500$.

Figure 5-6. Computed aggregate service bandwidth by the PRA, and the REM/RS methods with static number of active connections made up of heterogeneous traffic streams. NT and VT are used by equal number of connections, i.e., 50-50 % share.

5.3.1.3 Study Conclusion

Based on the results of the performance studies investigating the accuracy of the PRA and the REM/RS methods under both homogeneous and heterogeneous traffic scenarios, we observed that:

- The PRA method grossly over-estimates the *S* values.
- The REM/RS method computes fairly accurate *S* values that are close to the measured service bandwidth values.

5.3.2 Connection Departure Issue

In this section, we report on the performance results of an MBCAC approach using (1) the REM/RS method, and (2) different traffic histogram update techniques whenever a connection departs from the network. Using random connection arrivals and departures, we measure and compare the effects a traffic histogram update technique has on the overall MBCAC performance.

The traffic histogram update techniques studied in this section are:

- No histogram Update (NU).
- ZT histogram Update (ZU).
- Exact histogram Update (EU).

Before using a histogram update technique, we have to understand the effects it may have on the REM/RS computation of the *S* value. Hence to investigate this, a SSQ with unlimited link capacity is used so that the admission controller will permit all new connections. This is done to ensure the mean aggregate service bandwidth (MSB) values measured from the experiments can be compared on equal term. The remaining simulation parameters, like

mean connection arrival rate and mean holding time, are set according to the values mentioned in chapter 4.

The later part of sections 5.3.2.1 and 5.3.2.2 report the results of the performance studies for the MBCAC approach with either one of the histogram update techniques in a SSQ with finite link capacity, along with random acceptances and rejections of new connections.

5.3.2.1 Homogeneous Traffic Streams

From Table 5-17, the MBCAC approach with REM/RS method and NU technique consistently estimates higher MSB values than other histogram update techniques. This is because the algorithm uses traffic histograms that still contain traffic contributions by connections that have departed from the network. As a result, the algorithm computes higher *S* values on the assumption that these departed connections are still active in the link, and hence would continue to consume some amount of link capacity. This would imply that traffic histograms must not maintain very long traffic load history because: (1) a connection has finite holding time, and (2) to minimize the effects departed connections have on the aggregate traffic statistics.

Though the MBCAC approach with REM/RS method and EU technique can estimate *tight* MSB values, the updates of these traffic histograms whenever a connection departs is both impractical and operationally expensive. This is because for the EU technique, a connection's past traffic contributions, i.e., number of cells transmitted during a time-slot, are recorded throughout its entire holding time duration. When that connection ceases to be active, all records in a traffic histogram are updated by subtracting that connection's traffic contributions, if any. This update process is done for all traffic histograms. It is obvious that this technique requires additional storage space and valuable computing time. Nevertheless, we use the results of this technique as a performance benchmark.

The MBCAC approach with REM/RS method and ZU technique gives MSB values quite close to the values computed by the EU technique. This is achieved without excessive storage and computing requirements.

From the table, we can conclude that the MBCAC approach with REM/RS method and ZU technique is efficient in estimating the aggregate service bandwidth *S* without imposing unrealistic demands on the network switches.

Table 5-18 shows the results of the performance studies for the MBCAC approach with either one of the histogram update techniques in a SSQ with finite link capacity. From this table, the MBCAC approach with REM/RS method and EU technique produces the best performance, followed closely by the approach with the ZU technique. For the approach with the NU technique, although it may not give the best performance, it still managed to produce decent results without having to update any traffic histograms.

TS	**PQ**	***I***	**NU**	**ZU**	**EU**
NT	MSB – REM/RS	0	2040.1	1678.6	1610.1
		500	1716.4	1402.2	1340.7
VT	MSB – REM/RS	0	7053.1	6812.4	6707.8
		500	6579.4	6330.4	6217.5

Table 5-17. Mean aggregate service bandwidth – MBCAC approaches with REM/RS method and different histogram update techniques for homogeneous traffic streams. SSQ capacity is infinite.

TS	PQ	*I*	NU	ZU	EU
NT	Util. (%)	0	51.9	66.7	68.1
		500	65.6	75.5	77.1
	CLR	0	3.5e-6	4.7e-5	7.8e-5
		500	5.4e-7	2.7e-5	5.6e-5
	MSC	0	110.2	135.7	137.9
		500	134.5	150	152.8
	MBR (%)	0	19.6	9.9	9.6
		500	8.7	5.2	5.1
	MSB – REM/RS	0	1356.1	1260.7	1238.8
		500	1289.7	1164.5	1146
VT	Util. (%)	0	67.4	70.1	71.2
		500	72.6	75.2	77.2
	CLR	0	1.8e-6	7.4e-6	1.2e-5
		500	1.6e-6	4.3e-6	8.5e-6
	MSC	0	108.3	113.2	114.7
		500	116.6	120.6	123.8
	MBR (%)	0	26	23.2	21.3
		500	20.2	17.7	17
	MSB – REM/RS	0	5282.1	5302.7	5262.7
		500	5216.9	5177.7	5175

Table 5-18. Performance quantities – MBCAC approaches with REM/RS method and different histogram update techniques for homogeneous traffic streams.

5.3.2.2 Heterogeneous Traffic Streams

From Table 5-19, the *S* values computed when using different histogram update techniques are consistent with the results recorded for homogeneous traffic streams (Table 5-17).

Next, the performance of the MBCAC approach with either one of the histogram update techniques in a SSQ with finite link capacity is studied and

reported in Table 5-20. From this table, the MBCAC approach with REM/RS method and EU technique fails to meet the desired QoS. This is because the approach computes tight *S* values that do not cater for sudden bursts of arrival traffic. However, the ZU technique can meet QoS and its performance is the best amongst the MBCAC approaches studied here.

TS	PQ	*I*	NU	ZU	EU
NT and VT	MSB – REM/RS	0	94438.2	78924.3	76994.4
		500	94144.9	78459.9	76506.6

Table 5-19. Mean aggregate service bandwidth – MBCAC approaches with REM/RS method and different histogram update techniques for heterogeneous traffic streams. SSQ capacity is infinite.

TS	PQ	*I*	NU	ZU	EU
NT and VT	Util. (%)	0	48.7	85.2	86.6
		500	52.1	85.7	87.1
	CLR	0	6.3e-6	8.9e-5	1.4e-4
		500	4.8e-6	9.5e-5	1.6e-4
	MSC – NT	0	75.1	85.2	84.6
		500	78.6	85.6	85.1
	MSC – VT	0	38.9	67.9	69
		500	41.7	68.3	69.4
	MBR (%) – NT	0	13.8	3.2	3.4
		500	10.6	3.2	3.2
	MBR (%) – VT	0	47.8	16.6	15.8
		500	43.2	16.5	15.4
	MSB – REM/RS	0	67791.8	64396.3	63730.5
		500	67625.6	64269.2	63577.8

Table 5-20. Performance quantities – MBCAC approaches with REM/RS method and different histogram update techniques for heterogeneous traffic streams.

5.3.2.3 Study Conclusion

Based on the results of the performance studies investigating the effects a traffic histogram update technique has on the overall MBCAC performance, we observed that under both homogeneous and heterogeneous traffic scenarios:

- The MBCAC approach with REM/RS method and ZU technique can estimate aggregate service bandwidth values quite close to the values computed by the complex EU technique, without having excessive storage and computing requirements.

5.3.3 Constraint Liberalization Issue

In this section, we report on the performance studies investigating the effects of relaxing equation (3.25)'s cell loss rate *L* constraint on the overall MBCAC performance. In these studies, the MBCAC approaches use (1) REM/RS method, and (2) Exact histogram update technique.

5.3.3.1 Homogeneous Traffic Streams

From Table 5-21 and Table 5-22, the results clearly show that decreasing the constraint does not necessarily increase link utilization. Instead, it causes the MBCAC approaches to return unpredictable performances.

TS	PQ	*I*	*L*=1e-4	*L*=1e-2	*L*=1
NT	Util. (%)	0	68.1	68.2	68.1
		500	77.1	77.2	77.1
	CLR	0	7.8e-5	7.2e-5	7.9e-5
		500	5.6e-5	5.7e-5	5.7e-5
	MSC	0	137.9	137.9	137.8
		500	152.8	152.9	152.8
	MBR (%)	0	9.6	9.5	9.5
		500	5.1	5.1	5.1
	MSB – REM/RS	0	1238.8	1236.3	1238.1
		500	1146	1144.3	1146

Table 5-21. Performance quantities – MBCAC approaches with REM/RS method and Exact histogram update for homogeneous NT traffic streams.

TS	PQ	*I*	*L*=1e-4	*L*=1e-2	*L*=1
VT	Util. (%)	0	71.2	70.7	70.8
		500	77.2	77.2	77.2
	CLR	0	1.2e-5	1.2e-5	9.6e-6
		500	8.5e-6	1.2e-5	8.5e-6
	MSC	0	114.7	113.6	113.8
		500	123.8	124	123.8
	MBR (%)	0	21.3	21.3	22.2
		500	17	16	17
	MSB – REM/RS	0	5262.7	5254.4	5259.5
		500	5175	5150.9	5175.4

Table 5-22. Performance quantities – MBCAC approaches with REM/RS method and Exact histogram update for homogeneous VT traffic streams.

5.3.3.2 Heterogeneous Traffic Streams

From Table 5-23, the results show that the constraint liberalization technique produces inconsistent MBCAC performances.

TS	PQ	*l*	*L*=1e-4	*L*=1e-2	*L*=1
NT and VT	Util. (%)	0	86.6	86.8	86.6
		500	87.1	87.3	87.1
	CLR	0	1.4e-4	1.5e-4	1.4e-4
		500	1.6e-4	1.8e-4	1.6e-4
	MSC – NT	0	84.6	85.1	84.6
		500	85.1	84.9	85.4
	MSC – VT	0	69	69.2	69
		500	69.4	69.6	69.4
	MBR (%) – NT	0	3.4	3.2	3.4
		500	3.2	3.1	3.2
	MBR (%) – VT	0	15.8	15.6	15.8
		500	15.4	15.3	15.3
	MSB – REM/RS	0	63730.5	63661.7	63743.6
		500	63577.8	63529.9	63593.5

Table 5-23. Performance quantities – MBCAC approaches with REM/RS method and Exact histogram update for heterogeneous traffic streams.

5.3.3.3 Study Conclusion

From the performance studies reported in this section, the results clearly show that relaxing the *L* constraint does not translate to any improvements in MBCAC performance. Rather, this technique causes the MBCAC approaches to return unpredictable performances.

5.3.4 Adaptive Feedback Control Mechanism

Based on the AFCM simulation settings mentioned in section 4.4.4.2, we investigate the efficiency of the MBCAC approaches using combinations of:

- Prudence level policy module – (1) Adaptive Weight Feedback (AWF) method, (2) Adaptive Warming-up Period (AWP) method.

- Load and traffic measurements module – (1) Link Occupancy (LO), (2) Buffer Occupancy (BO), (3) Cell Loss Conservative Period (CLCP).

Note that the AWP method and the CLCP technique are not used together because both methods serve nearly the same purpose. All MBCAC approaches studied here use the Exact histogram update technique.

5.3.4.1 Homogeneous Traffic Streams

5.3.4.1.1 Adaptive Weight Feedback Method

From Table 5-24 and Table 5-25, it is obvious that the REM/RS method is fairly accurate in estimating the aggregate service bandwidth *S* value. For both NT and VT connections, the REM/RS method is able to achieve efficient CAC performance without having the need to use any AFCM techniques. From the tables, we also observed that the use of the AFCM actually suppresses the MBCAC approaches from attaining higher performances.

TS	PQ	*I*	REM/RS	AWF, LO=90%	AWF, BO=20%	AWF, CLCP=5	AWF, CLCP=30
NT	Util. (%)	0	68.1	61.9	-	55.1	33.8
		500	77.1	66.1	73	68.3	48.8
	CLR	0	7.8e-5	5e-5	-	6.4e-5	4.1e-5
		500	5.6e-5	9.1e-6	2.8e-5	4.6e-5	3.7e-5
	MSC	0	137.9	122.9	-	110	67.4
		500	152.8	128.8	143.6	134.7	96.6
	MBR (%)	0	9.6	23.1	-	30	56.8
		500	5.1	24.3	12.3	16.9	39.6
	MWei	0	-	0.92	-	0.77	0.43
		500	-	0.93	0.96	0.87	0.6
	MSB – PRA	0	-	12285.7	-	10586.3	6835.8
		500	-	15287.7	15179.4	13862.4	10018.4
	MSB – REM/RS	0	1238.8	1172.1	-	1073.1	753.5
		500	1146	1008.6	1095.7	1028.4	746.1

Table 5-24. Performance quantities – MBCAC approaches with AWF, Exact histogram update, and different Load and traffic measurements techniques. Using homogeneous NT streams.

TS	PQ	*I*	REM/RS	AWF, LO=90%	AWF, BO=20%	AWF, CLCP=5	AWF, CLCP=30
VT	Util. (%)	0	71.2	69.9	-	50.5	38.7
		500	77.2	71.8	76.6	54	35.4
	CLR	0	1.2e-5	1.9e-6	-	1.5e-6	2.1e-6
		500	8.5e-6	0	3.5e-6	3.3e-6	1.9e-6
	MSC	0	114.7	111.8	-	80.9	62.3
		500	123.8	114.5	122.5	86.1	56.9
	MBR (%)	0	21.3	24.2	-	44.4	57.7
		500	17	25.8	17.5	44.1	63.9
	MWei	0	-	0.92	-	0.62	0.43
		500	-	0.91	0.96	0.6	0.35
	MSB – PRA	0	-	32758.8	-	23772.2	18151.2
		500	-	34418.9	36101.5	25614.7	17055.2
	MSB – REM/RS	0	5262.7	5199.2	-	3917.9	3055.9
		500	5175	4838.1	5150.8	3690.8	2388.3

Table 5-25. Performance quantities – MBCAC approaches with AWF, Exact histogram update, and different Load and traffic measurements techniques. Using homogeneous VT streams.

5.3.4.1.2 Adaptive Warming-up Period Method

From Table 5-26 and Table 5-27, the most efficient approach is the MBCAC approach that uses only the REM/RS method.

TS	PQ	*l*	REM/RS	WP=15, LO=90%	WP=30, LO=90%	WP=15, BO=20%	WP=30, BO=20%
NT	Util. (%)	0	68.1	66.4	65.6	-	-
		500	77.1	73.4	71.8	76	75.6
	CLR	0	7.8e-5	7.4e-5	6.2e-5	-	-
		500	5.6e-5	1.3e-5	5.7e-6	4.4e-5	4.8e-5
	MSC	0	137.9	134.2	132.8	-	-
		500	152.8	146.2	142.8	150.6	149.7
	MBR (%)	0	9.6	12.1	13.1	-	-
		500	5.1	8.9	10.7	6.3	7.1
	MWei	0	-	0.88	0.9	-	-
		500	-	0.79	0.84	0.95	0.95
	MSB – PRA	0	-	11732.1	11675.6	-	-
		500	-	15010.4	14699.4	15117	15162.2
	MSB – REM/RS	0	1238.8	1213.4	1202.9	-	-
		500	1146	1083.5	1059.6	1129.4	1124
	MWP	0	-	1.7	3.1	-	-
		500	-	3.2	4.8	0.8	1.4

Table 5-26. Performance quantities – MBCAC approaches with AWP, Exact histogram update, and different Load and traffic measurements techniques. Using homogeneous NT streams.

TS	PQ	*l*	REM/RS	WP=15, LO=90%	WP=30, LO=90%	WP=15, BO=20%	WP=30, BO=20%
VT	Util. (%)	0	71.2	71	70.8	-	-
		500	77.2	73.5	72	77.1	76.6
	CLR	0	1.2e-5	2.4e-6	3.2e-6	-	-
		500	8.5e-6	1.2e-7	0	7e-6	6.5e-6
	MSC	0	114.7	113.9	113.9	-	-
		500	123.8	117.8	115.5	123.2	122.6
	MBR (%)	0	21.3	23.4	23.9	-	-
		500	17	20.3	21.8	18.1	18.1
	MWei	0	-	0.91	0.92	-	-
		500	-	0.83	0.88	0.96	0.96
	MSB – PRA	0	-	33432.3	33506.1	-	-
		500	-	34707.6	33893.7	36399	36169.3
	MSB – REM/RS	0	5262.7	5218.5	5198.3	-	-
		500	5175	4868.2	4820.7	5171.9	5147.1
	MWP	0	-	1.4	2.4	-	-
		500	-	2.5	3.7	0.7	1.3

Table 5-27. Performance quantities – MBCAC approaches with AWP, Exact histogram update, and different Load and traffic measurements techniques. Using homogeneous VT streams.

5.3.4.2 Heterogeneous Traffic Streams

5.3.4.2.1 Adaptive Weight Feedback Method

From Table 5-28, the MBCAC approach with AWF and LO=90% techniques produced the best performance, whilst still meeting the aggregate QoS requirement.

TS	PQ	*I*	REM/RS	AWF, LO=90%	AWF, BO=20%	AWF, CLCP=5	AWF, CLCP=30
NT and VT	Util. (%)	0	86.6	75.7	-	41	34.9
		500	87.1	73.7	81	41.2	35.5
	CLR	0	1.4e-4	1.2e-6	-	1.1e-5	3.8e-6
		500	1.6e-4	7e-7	8.2e-5	1.2e-5	1.6e-6
	MSC – NT	0	84.6	65.9	-	78.9	82.6
		500	85.1	65.9	75.1	79.7	83.3
	MSC – VT	0	69	60.2	-	32.7	27.9
		500	69.4	60.2	64.4	32.8	28.3
	MBR (%) – NT	0	3.4	25	-	7.5	3.7
		500	3.2	24.8	15.7	7	3.1
	MBR (%) – VT	0	15.8	28.7	-	56.2	59.9
		500	15.4	28.9	21.9	55.5	59
	MWei	0	-	0.85	-	0.17	0.03
		500	-	0.85	0.9	0.17	0.04
	MSB – PRA	0	-	151102.4	-	83657.5	71249.8
		500	-	151325	160532.6	83387.2	72142
	MSB – REM/RS	0	63730.5	56289.8	-	30974.1	26431.8
		500	63577.8	55858.8	59446	30619.2	26372.1

Table 5-28. Performance quantities – MBCAC approaches with AWF, Exact histogram update, and different Load and traffic measurements techniques. Traffic streams are heterogeneous.

5.3.4.2.2 Adaptive Warming-up Period Method

From Table 5-29, the MBCAC approach with WP=15 and LO=90% techniques is the most efficient approach.

TS	PQ	*l*	REM/RS	WP=15, LO=90%	WP=30, LO=90%	WP=15, BO=20%	WP=30, BO=20%
NT and VT	Util. (%)	0	86.6	83.6	82.1	-	-
		500	87.1	84	82.4	86.2	85.5
	CLR	0	1.4e-4	8.4e-6	5.3e-6	-	-
		500	1.6e-4	7.7e-6	3.7e-6	1.2e-4	1.1e-4
	MSC – NT	0	84.6	85.9	85.3	-	-
		500	85.1	86	86	85.4	85.7
	MSC – VT	0	69	66.6	65.4	-	-
		500	69.4	66.9	65.6	68.6	68.1
	MBR (%) – NT	0	3.4	2.4	2.6	-	-
		500	3.2	2.5	2.4	2.8	2.8
	MBR (%) – VT	0	15.8	18.2	20	-	-
		500	15.4	18.1	19.7	16.5	16.7
	MWei	0	-	0.45	0.59	-	-
		500	-	0.42	0.56	0.84	0.86
	MSB – PRA	0	-	165610	162618.3	-	-
		500	-	166675.1	163568.9	170918.6	169478.8
	MSB – REM/RS	0	63730.5	59933.7	58422.3	-	-
		500	63577.8	59646.9	57984.8	62679.2	62005.1
	MWP	0	-	8.3	12.4	-	-
		500	-	8.7	13.1	2.4	4.3

Table 5-29. Performance quantities – MBCAC approaches with AWP, Exact histogram update, and different Load and traffic measurements techniques. Traffic streams are heterogeneous.

5.3.4.3 Study Conclusion

Based on the results of the performance studies investigating the efficiency of the MBCAC approaches using different combinations of AFCM techniques under both homogeneous and heterogeneous traffic scenarios, we observed that:

- For the homogeneous traffic streams, the MBCAC approach with only the REM/RS method is efficient.

- For the heterogeneous traffic streams, the MBCAC approach with AWP and LO techniques is efficient.

5.3.5 'Mix and Match' Techniques

In this section, we 'mix and match' different techniques and use different AFCM settings so as to find the best combination that work well with both non-buffered and buffered SSQ. Our choices of techniques are based on the results of the performance studies reported in the previous sections. In addition, we also apply these techniques on traffic streams made up of NT2 and/or VT2 connections.

All MBCAC approaches studied here use either the No histogram update technique (section 5.3.5.1), or the ZT histogram update technique (section 5.3.5.2).

5.3.5.1 No Histogram Update

5.3.5.1.1 Homogeneous Traffic Streams

Best CAC performance:

- NT, Table 5-30: MBCAC approach with AWP, WP=15 and LO=90% techniques.

- VT, Table 5-31: MBCAC approach with only the REM/RS method.

- NT2, Table 5-32: MBCAC approach with only the REM/RS method.

- VT2, Table 5-33: For the l = 0 case, it is the MBCAC approach with AWF and LO=90% techniques. For the l = 500 case, it is the MBCAC approach with only the REM/RS method.

TS	PQ	I	REM/RS	AWF, LO=90%	WP=15, LO=90%
NT	Util. (%)	0	51.9	51.8	52.3
		500	65.6	65	65.7
	CLR	0	3.5e-6	1.6e-6	2e-6
		500	5.4e-7	0	0
	MSC	0	110.2	109.9	110.9
		500	134.5	131.4	134.4
	MBR (%)	0	19.6	20.1	19.3
		500	8.7	13.3	9.2
	MWei	0	-	0.99	0.99
		500	-	0.98	0.97
	MSB – PRA	0	-	7151.6	7184.6
		500	-	10895.8	10520.2
	MSB – REM/RS	0	1356.1	1355.6	1353.7
		500	1289.7	1252.9	1275.5
	MWP	0	-	-	0.1
		500	-	-	0.5

Table 5-30. Performance quantities – MBCAC approaches with No histogram update for homogeneous NT traffic streams.

TS	PQ	*I*	REM/RS	AWF, LO=90%	WP=15, LO=90%
VT	Util. (%)	0	67.4	66.7	66
		500	72.6	71.4	72.2
	CLR	0	1.8e-6	5.6e-7	6.3e-7
		500	1.6e-6	0	0
	MSC	0	108.3	107.4	106
		500	116.6	114	115.7
	MBR (%)	0	26	26.3	25.8
		500	20.2	23.8	22.5
	MWei	0	-	0.96	0.96
		500	-	0.93	0.86
	MSB – PRA	0	-	31138.1	30595.5
		500	-	33887.6	34091.3
	MSB – REM/RS	0	5282.1	5267	5272.1
		500	5216.9	5065.1	5100.5
	MWP	0	-	-	0.7
		500	-	-	2

Table 5-31. Performance quantities – MBCAC approaches with No histogram update for homogeneous VT traffic streams.

TS	PQ	I	REM/RS	AWF, LO=90%	WP=15, LO=90%
NT2	Util. (%)	0	49.3	49.1	49.1
		500	52.2	51.6	51.9
	CLR	0	9.6e-7	9.6e-7	9.6e-7
		500	4.6e-5	5.6e-5	6.8e-5
	MSC	0	134.8	134.1	134.5
		500	139.6	136.9	138.3
	MBR (%)	0	10	10.5	10.2
		500	8.7	11.7	10.1
	MWei	0	-	1	1
		500	-	0.99	0.99
	MSB – PRA	0	11308.9	11263	11278.7
		500	12628.2	12865.8	12678.5
	MSB – REM/RS	0	1691	1691.8	1692
		500	1624.7	1619.1	1625
	MWP	0	-	-	0.04
		500	-	-	0.19

Table 5-32. Performance quantities – MBCAC approaches with No histogram update for homogeneous NT2 traffic streams.

TS	PQ	I	REM/RS	AWF, LO=90%	WP=15, LO=90%
VT2	Util. (%)	0	70.7	71.1	70.5
		500	77.8	74.5	75.7
	CLR	0	1.5e-6	1.4e-7	6.2e-7
		500	2.5e-6	0	0
	MSC	0	107.9	108.6	107.7
		500	118.8	113.8	115.6
	MBR (%)	0	28.9	27.7	29.7
		500	20.9	25.7	23.5
	MWei	0	-	0.95	0.92
		500	-	0.92	0.77
	MSB – PRA	0	17397.4	17482.3	17430.6
		500	19142.5	18406.8	18636.8
	MSB – REM/RS	0	3578.8	3561.6	3561.7
		500	3551.6	3286.6	3312.9
	MWP	0	-	-	1.1
		500	-	-	3.5

Table 5-33. Performance quantities – MBCAC approaches with No histogram update for homogeneous VT2 traffic streams.

5.3.5.1.2 Heterogeneous Traffic Streams

Best CAC performance:

- NT and VT, Table 5-34: MBCAC approach with AWF and LO=90% techniques.
- NT2 and VT2, Table 5-35: MBCAC approach with AWP, WP=15 and LO=90% techniques.

TS	PQ	*I*	REM/RS	AWF, LO=90%	WP=15, LO=90%
NT and VT	Util. (%)	0	48.7	67	66.9
		500	52.1	69.2	68.4
	CLR	0	6.3e-6	9e-8	8e-7
		500	4.8e-6	6.2e-9	5.2e-7
	MSC – NT	0	75.1	76.7	86
		500	78.6	74.9	86.8
	MSC – VT	0	38.9	53.5	53.5
		500	41.7	55.2	54.7
	MBR (%) – NT	0	13.8	14.4	2.6
		500	10.6	15.6	1.3
	MBR (%) – VT	0	47.8	31	28.1
		500	43.2	29.5	28.6
	MWei	0	-	0.94	0.91
		500	-	0.93	0.89
	MSB – PRA	0	-	132187.8	130252.3
		500	-	136015.7	134684.8
	MSB – REM/RS	0	67791.8	66761.9	66922.7
		500	67625.6	66589.9	66870.4
	MWP	0	-	0.94	1.3
		500	-	0.93	1.7

Table 5-34. Performance quantities – MBCAC approaches with No histogram update for heterogeneous NT and VT traffic streams.

TS	PQ	I	REM/RS	AWF, LO=90%	WP=15, LO=90%
NT2 and VT2	Util. (%)	0	45.2	65.4	65.5
		500	46.6	68	68.8
	CLR	0	7.7e-6	7.8e-7	8.8e-7
		500	6.3e-6	7.2e-7	1.7e-6
	MSC – NT	0	72.5	70.5	82.3
		500	75.2	68.4	81.7
	MSC – VT	0	33.7	49.2	49.2
		500	34.8	51.2	51.7
	MBR (%) – NT	0	13.8	19	3.6
		500	12.6	20.4	4.6
	MBR (%) – VT	0	56.5	39.8	38.4
		500	55.5	38.1	35.5
	MWei	0	-	0.92	0.88
		500	-	0.91	0.85
	MSB – PRA	0	57599.6	80444.1	80924.3
		500	60022.9	82854	84855.1
	MSB – REM/RS	0	45438.9	44814	44978.8
		500	45473.9	44621	44761
	MWP	0	-	-	1.9
		500	-	-	2.3

Table 5-35. Performance quantities – MBCAC approaches with No histogram update for heterogeneous NT2 and VT2 traffic streams.

5.3.5.2 ZT Histogram Update

5.3.5.2.1 Homogeneous Traffic Streams

Best CAC performance:

- NT, Table 5-36: MBCAC approach with only the REM/RS method.
- VT, Table 5-37: MBCAC approach with only the REM/RS method.
- NT2, Table 5-38: None of the MBCAC approaches.

- VT2, Table 5-39: MBCAC approach with only the REM/RS method.

TS	PQ	*I*	REM/RS	AWF, LO=90%	WP=15, LO=90%
NT	Util. (%)	0	66.7	61.8	65.4
		500	75.5	66.2	72.1
	CLR	0	4.7e-5	3.3e-5	3.6e-5
		500	2.7e-5	1.4e-6	8e-6
	MSC	0	135.7	123.6	133.1
		500	150	129.4	143.8
	MBR (%)	0	9.9	21	12.1
		500	5.2	23	9.2
	MWei	0	-	0.93	0.91
		500	-	0.94	0.84
	MSB – PRA	0	-	11388.4	11338.2
		500	-	14923.3	14022.9
	MSB – REM/RS	0	1260.7	1221.8	1247.9
		500	1164.5	1068.7	1116.3
	MWP	0	-	-	1.3
		500	-	-	2.4

Table 5-36. Performance quantities – MBCAC approaches with ZT histogram update for homogeneous NT traffic streams.

TS	PQ	*I*	REM/RS	AWF, LO=90%	WP=15, LO=90%
VT	Util. (%)	0	70.1	69.2	69.6
		500	75.2	71.3	73.3
	CLR	0	7.4e-6	1.3e-6	1.5e-6
		500	4.3e-6	0	1.5e-7
	MSC	0	113.2	110.7	111.8
		500	120.6	114	117.7
	MBR (%)	0	23.2	25.6	23.6
		500	17.7	24.7	22.1
	MWei	0	-	0.93	0.92
		500	-	0.91	0.82
	MSB – PRA	0	-	32375.8	32621.7
		500	-	33930.2	34872.9
	MSB – REM/RS	0	5302.7	5276	5264.8
		500	5177.7	4913	4964.6
	MWP	0	-	-	1.2
		500	-	-	2.7

Table 5-37. Performance quantities – MBCAC approaches with ZT histogram update for homogeneous VT traffic streams.

TS	PQ	*I*	REM/RS	AWF, LO=90%	WP=15, LO=90%
NT2	Util. (%)	0	56.6	53.6	55.9
		500	61.5	55.9	60.1
	CLR	0	1.7e-4	2.2e-4	1.6e-4
		500	3.5e-4	2.7e-4	3.1e-4
	MSC	0	146.3	135.7	144.3
		500	153.8	135.8	149.9
	MBR (%)	0	7.9	16.3	9.4
		500	5.7	20.1	8.1
	MWei	0	-	0.94	0.94
		500	-	0.92	0.9
	MSB – PRA	0	-	15606.9	15589.2
		500	-	18692.4	18164
	MSB – REM/RS	0	1557.1	1535.7	1550.9
		500	1461.7	1401.9	1445.3
	MWP	0	-	-	1
		500	-	-	1.5

Table 5-38. Performance quantities – MBCAC approaches with ZT histogram update for homogeneous NT2 traffic streams.

TS	PQ	I	REM/RS	AWF, LO=90%	WP=15, LO=90%
VT2	Util. (%)	0	73.9	73.1	72.9
		500	82.7	74.6	76
	CLR	0	5.4e-6	1e-6	1.5e-6
		500	5.9e-6	0	0
	MSC	0	112.8	111.7	111.3
		500	126.1	113.9	116.1
	MBR (%)	0	24.1	26.2	26.1
		500	16.1	24.9	23.1
	MWei	0	-	0.91	0.89
		500	-	0.92	0.74
	MSB – PRA	0	-	18000.3	17906.7
		500	-	18364	18724.6
	MSB – REM/RS	0	3579.2	3525.6	3523.2
		500	3496.6	3157.1	3205.3
	MWP	0	-	-	1.7
		500	-	-	3.9

Table 5-39. Performance quantities – MBCAC approaches with ZT histogram update for homogeneous VT2 traffic streams.

5.3.5.2.2 Heterogeneous Traffic Streams

Best CAC performance:

- NT and VT, Table 5-40: MBCAC approach with only the REM/RS method.
- NT2 and VT2, Table 5-41: MBCAC approach with AWP, WP=15 and LO=90% techniques.

TS	PQ	*I*	REM/RS	AWF, LO=90%	WP=15, LO=90%
NT and VT	Util. (%)	0	85.2	74.9	82
		500	85.7	74.9	82.5
	CLR	0	8.9e-5	7.9e-7	8e-6
		500	9.5e-5	7.2e-7	5.5e-6
	MSC – NT	0	85.2	66	87.1
		500	85.6	66.8	86.7
	MSC – VT	0	67.9	59.5	65.3
		500	68.3	59.6	65.7
	MBR (%) – NT	0	3.2	25.4	1.4
		500	3.2	24.3	1.5
	MBR (%) – VT	0	16.6	29.4	18.7
		500	16.5	28.2	18.9
	MWei	0	-	0.86	0.57
		500	-	0.86	0.54
	MSB – PRA	0	-	149650.7	161825.3
		500	-	148649.1	163522
	MSB – REM/RS	0	64396.3	58279.3	60952.5
		500	64269.2	57823.4	60740.6
	MWP	0	-	-	6.5
		500	-	-	6.9

Table 5-40. Performance quantities – MBCAC approaches with ZT histogram update for heterogeneous NT and VT traffic streams.

TS	PQ	I	REM/RS	AWF, LO=90%	WP=15, LO=90%
NT2 and VT2	Util. (%)	0	86.3	75.5	83.1
		500	87.2	75.8	83.7
	CLR	0	1.2e-4	1.4e-6	1e-5
		500	1.4e-4	6.5e-7	8.3e-6
	MSC – NT	0	82.1	63.2	84.8
		500	82.6	62.1	84.8
	MSC – VT	0	65.1	57	62.6
		500	65.8	57.3	63.1
	MBR (%) – NT	0	4.4	28.1	2.3
		500	4.5	28.5	2.4
	MBR (%) – VT	0	21.8	32.6	24.3
		500	21.7	33.2	24.1
	MWei	0	-	0.83	0.46
		500	-	0.83	0.42
	MSB – PRA	0	-	91107.1	101182.5
		500	-	91169	101925.6
	MSB – REM/RS	0	43576.9	39259	41060.7
		500	43572.2	38924.3	40759.8
	MWP	0	-	-	8
		500	-	-	8.7

Table 5-41. Performance quantities – MBCAC approaches with ZT histogram update for heterogeneous NT2 and VT2 traffic streams.

5.3.5.3 Study Conclusion

In this section, we investigate the efficiency of a variety of MBCAC approaches using different combinations of technique and different AFCM settings. Based on the results of the performance studies, we have observed that the MBCAC approaches with the ZU technique generally produce better performances than the approaches using the NU technique.

The list below compares the performance results of the ZU technique against the NU technique:

- For the homogeneous traffic streams case, link utilization increases by up to 13%. However for the NT2 connections, the MBCAC approach with the ZU technique was unable to meet the desired QoS.

- For the heterogeneous traffic streams case, link utilization increases by up to 20%.

5.3.6 Framework Conclusion

The MBCAC approaches created within the histogram-based CAC framework are made up of different combinations of technique. Hence within this framework, a network provider can construct a customized MBCAC approach based on their traffic control requirements.

From this framework, we investigated a number of CAC performance-related issues. Based on the results of these studies, we then 'mix and match' techniques and use tested AFCM settings to create different MBCAC approaches whose performances are compared under both homogeneous and heterogeneous traffic streams scenarios. These results are then collated and listed in Table 5-42.

From this table, the MBCAC approach that produces consistent performance is:

- The MBCAC approach with REM/RS method and ZT histogram update technique.

TS	*I*	MBCAC	Util. (%)
NT	0	ZU, REM/RS	66.7
	500		75.5
VT	0	ZU, REM/RS	70.1
	500		75.2
NT and VT	0	ZU, REM/RS	85.2
	500		85.7
NT2	0	NU, REM/RS	49.3
	500		52.2
VT2	0	ZU, REM/RS	73.9
	500		82.7
NT2 and VT2	0	ZU, AWP, WP=15, LO=90%	83.1
	500		83.7

Table 5-42. Most efficient MBCAC performance – Histogram-based framework.

5.4 Model and Histogram based Frameworks: Overall CAC Performance

In this section, we compare the performance results for all CAC and MBCAC approaches from both the model-based and the histogram-based frameworks. To make fair comparisons, we only use performance results based on NT2 and/or VT2 connections. This ensures all approaches are compared under a 'minimal a-priori traffic information' scenario.

Table 5-43 lists the best overall CAC and MBCAC performances from both model and histogram based frameworks. The third column lists the most efficient approaches with their link utilization results listed in column five. Column four lists approaches whose link utilization results (column six) are lower than the results in column five by less than 4%.

TS	*I*	CAC/MBCAC		Util. (%)	
NT2	0	NU, REM/RS	-	49.3	-
	500		-	52.2	-
VT2	0	m-eGA, AWF, LO=90%	ZU, REM/RS	74.4	73.9
	500	m-GA, AWF, BO=90%		83.5	82.7
NT2 and VT2	0	ZU, AWP, WP=15, LO=90%	eGA	83.1	82.5
	500	eGA	ZU, AWP, WP=15, LO=90%	87.1	83.7

Table 5-43. Best overall CAC and MBCAC performances from both model and histogram based frameworks.

5.5 Conclusions

In this chapter, we have analyzed and reported in detail the intensive comparative performance studies of all CAC and MBCAC schemes contained within the model-based and the histogram-based frameworks. These studies address the simplicity versus efficiency tradeoff issues relevant to practical admission control methodologies.

Using realistic network data and video traces, the CAC and MBCAC schemes are studied under homogeneous and heterogeneous traffic conditions. These realistic traffic traces allow the schemes to be stress tested in real-life traffic scenarios. From these studies, a variety of performance related quantities are then measured and reported here.

In addition to the above set of traffic traces, which we term as the default set of traffic sources, we also conducted efficiency studies using a set of two new network data and video traces. When using this new set of traffic sources, minimal a-priori traffic information is provided to the CAC and MBCAC schemes.

Particularly for the model-based CAC schemes and their measurement-based counterparts, established connections using this new set of traffic sources will create a hostile operating environment because the schemes are making admission decisions based on several traffic statistics taken instead from the default traffic sources.

In this chapter, we also thoroughly examined a variety of research issues relevant to each framework, and the results of these studies are listed in the numerous tables presented here.

Furthermore, at the end of each CAC framework, we highlight the CAC and MBCAC approaches that produced the most efficient performances under

different traffic scenarios. These approaches are then summarized into a 'Best overall' list for both model and histogram based frameworks.

Lastly, we would like to note that the MBCAC approaches within the two CAC frameworks are customizable to the network providers' traffic control requirements. In other words, network providers can 'mix and match' techniques and set AFCM values, different from those used in the performance studies reported in this thesis.

Hence, network providers are not restricted to only (1) the combinations of technique and (2) the AFCM settings used by the recommended MBCAC approaches listed in the 'Best overall' list.

6 Summary and Extensions

In this chapter, we summarize the research work investigated by this thesis in section 6.1, and outline further work in section 6.2.

6.1 Summary

In this dissertation, we have investigated traffic engineering issues relevant to the design and development of efficient Connection Admission Control (CAC) strategies. Through extensive studies, results have been analyzed and presented as practical admission control methodologies for Quality of Service (QoS) assurance in a multiservice network. These research results provide the foundation for effective CAC procedures, and the principles laid here will serve to establish useful admission control guidelines for the benefits of both network providers and users.

We believe our work has addressed the issue of providing an admission control strategy that ensures: (1) established connections can experience an aggregate level of QoS agreed upon prior to connection set-up; (2) network providers can achieve good investment returns through the admittance of as many connections as the network can cater within the QoS limits; and (3) the admission decision process imposes minimal storage and computing requirements on the network switches.

Below is a list outlining the areas that have been examined in detail by this dissertation:

- We have developed a set of practical admission control methodologies through the formulation of two novel CAC frameworks proposed in sections 3.2 and 3.3. These frameworks contain different CAC and Measurement-based CAC (MBCAC) schemes. In chapter 5, these schemes are intensively studied under realistic traffic conditions and recommendations are then provided.

- We have developed a model-based CAC framework that consists of: (1) three CAC schemes based on two traditional traffic models – Gaussian and Effective Bandwidth, and the enhanced Gaussian traffic model; and (2) the measurement-based counterparts of the Gaussian and the enhanced Gaussian traffic models. This framework also includes an Adaptive Feedback Control Mechanism (AFCM), which functions as an additional control layer for QoS assurance. For the MBCAC schemes that use this AFCM, certain parameters can be configured by the network providers, so as to customize the MBCAC schemes according to their own traffic control requirements. In total, the performances of seven different CAC and MBCAC schemes are studied within this framework, and reported in section 5.2. Based on the studies reported in section 5.2.1, the Effective Bandwidth CAC scheme grossly over-estimates the bandwidth required to service established connections whilst still meeting the QoS requirement. For the Gaussian CAC scheme, it also over-estimates the service bandwidth but not as grossly inaccurate as the Effective Bandwidth CAC scheme. While for the enhanced Gaussian CAC scheme, it estimates fairly close to the measured bandwidth consumption values.

- We have developed within the model-based framework, the enhanced Gaussian CAC and MBCAC schemes. All schemes consider the level of traffic aggregation and multiplexing gain in its service bandwidth computations. From the performance studies reported in sections 5.2.1 to 5.2.4, these schemes consistently outperform the traditional

Gaussian and Effective Bandwidth CAC schemes, and the Gaussian MBCAC schemes, whilst still ensuring the aggregate QoS required.

- We have derived from the performance studies reported in section 4.4.2.2, Gaussian boundaries for two traffic genres, i.e., network data and video, for use in the enhanced Gaussian CAC and MBCAC schemes. These boundaries are expressed in term of the number of homogeneous connections required. Basically, if the total number of connections is below a certain boundary unique to that type of traffic, the enhanced Gaussian schemes assume the aggregate traffic stream is lightly aggregated and thereby exhibits non-Gaussian behavior. On the other hand, if it is greater than the boundary, because of the large number of connections in the link, aggregation will be high and hence the aggregate stream can be accurately modeled as a Gaussian process. From these studies, depending on the traffic genre, the boundaries are approximated to be in the range of 200 to 300 simultaneous homogeneous connections.

- All CAC and MBCAC schemes within the model-based framework require at least one a-priori traffic information, and this is the performance margin's multiple factor look-up table specific to a traffic genre. In section 5.2.4, we investigate and report on the effects of using a default multiple factor look-up table unique to a traffic genre, on connections transmitting work belonging to the same traffic genre but whose traffic statistics are not closely matched to that default multiple factor values. The motivation behind this study is that with a default set of tables for different traffic genres, the CAC and MBCAC schemes will be more easily deployable in a network, and it will also greatly simplify the use of these schemes. Based on the results of the performance studies, the default multiple factor values are still usable as long as the submitted work is not 'more bursty' than the default traffic source's burstiness rate. In other words, there will be no

significant CAC performance degradation if the above condition is met.

- We have developed a histogram-based CAC framework that consists of a variety of modules that are used to create MBCAC schemes. These modules also include the adaptive feedback controller – AFCM. Each module contains a collection of different techniques with common functionality. By 'mixing and matching' techniques taken from every module, an MBCAC scheme can be constructed specially for use in a network with certain traffic control demands. Hence, many customized MBCAC schemes can be created within this framework. In sections 5.3.2 to 5.3.4, the effects each module has on the overall MBCAC performance are studied and reported. Based on the results of these studies, various MBCAC schemes using different combinations of technique and different AFCM settings are also investigated and these are reported in section 5.3.5.

- We have incorporated three fundamental CAC algorithms, i.e., Peak Rate Allocation (PRA), Rate Envelope Multiplexing (REM), and Rate Sharing (RS), into a novel procedure of 'Available bandwidth' evaluation for the MBCAC schemes within the histogram-based framework. These algorithms are used to compute the bandwidth required to service the established connections, whilst still meeting the QoS requirement. Amongst these algorithms, PRA method is considered conservative while the latter two methods are considered liberal. The overall spare/available bandwidth value is then derived based on the choice of AFCM techniques and the instantaneous traffic load condition. Using results from the studies reported in section 5.3.1, the REM/RS method is fairly accurate in estimating the required service bandwidth values. From observations, these computed values are close to the measured bandwidth consumption values.

- We have incorporated different traffic histogram update techniques to handle connection departures for the MBCAC schemes within the histogram-based framework. Whenever a connection departs from the network, one of these techniques, in varying complexities and storage requirements – from no update to complete updates, will be used to alter, if applicable, all traffic histogram records. As shown by the results from the studies reported in section 5.3.2, whenever updated traffic histograms are used, it will result in more accurate available bandwidth values being computed. From these studies, we have also conducted that a simple update technique, i.e., ZT histogram update, will give performance almost on-par with the benchmark full update technique, i.e., Exact histogram update, without imposing unrealistic storage and computing requirements.

- We have investigated the effects of relaxing the cell loss rate L constraint on MBCAC schemes within the histogram-based framework. The REM/RS algorithm uses L as the desired cell loss threshold and then estimates an amount of service bandwidth that will adequately meet this particular cell loss rate. Hence, if we were to slowly remove this constraint by increasing the L value such that it results in minimal or zero service bandwidth values to be computed, then the overall effect will be higher spare bandwidth values. This will result in more new connections to be admitted into the link. As shown by the results from the studies reported in section 5.3.3, no significant improvement in the MBCAC performance is recorded whenever the L constraint is relaxed. Rather, this technique results in unpredictable MBCAC performances.

- We have developed the AFCM to be a generic component that can be used by a variety of different MBCAC schemes. This adaptive feedback controller protects the network whenever an MBCAC scheme fails to meet the aggregate QoS requirement, either because the scheme

tries to be too aggressive, or when the traffic exhibits unpredictable behavior, or both. It is simple to implement and imposes very minimal storage and computing demands on the network switches. It is basically a collection of two inter-dependent modules, i.e., (1) Prudence level policy module – which uses an active parameter to adapt the MBCAC scheme to changing traffic conditions, and (2) Load and traffic measurements module – which compares the traffic load against a choice of different threshold values.

As shown by the results from the performance studies reported in section 5.2.3 for the model-based framework, the MBCAC schemes that use the AFCM consistently achieve higher efficiency than other MBCAC schemes that do not use the feedback controller. However for the histogram-based framework, the results reported in section 5.3.4 suggested that the AFCM is only useful for heterogeneous traffic streams scenario. For homogeneous traffic streams scenario, the MBCAC schemes that do not use the AFCM produced better performance. Unlike the model-based framework, the MBCAC schemes within the histogram-based framework are purely measurement-based, except for the user-declared peak rate parameter. Hence, these schemes compute the aggregate service bandwidth based purely on real-time traffic measurements. While for the MBCAC schemes within the model-based framework, the aggregate service bandwidth is computed based on the (1) a-priori performance margin's multiple factor values, and (2) real-time traffic measurements. Because these schemes are dependent on this a-priori traffic parameter, as a result, they are not very adaptive to varying traffic load conditions, hence the need for an additional layer of control provided by the AFCM.

For the prudence level policy module, all methods are applicable to the histogram-based framework. However, for the model-based

framework, only one method is applicable. Hence, conclusions are only drawn from studies done on MBCAC schemes within the histogram-based framework. Based on the studies reported in sections 5.3.4 and 5.3.5, the adaptive warming-up period method consistently adapts the MBCAC schemes to varying traffic load conditions whilst still meeting the QoS requirement. For the load and traffic measurements module, the link occupancy technique generally provides threshold values that maximize link utilization whilst providing advance QoS breach warnings to the prudence level policy module. These conclusions are drawn from studies reported in sections 5.2.3, 5.2.4, 5.3.4 and 5.3.5.

- We have conducted an intensive comparative investigation of the performance of both CAC and MBCAC schemes within the model and histogram based frameworks. These studies address the simplicity versus efficiency tradeoff issues relevant to practical admission control methodologies. In the comprehensive performance studies reported in chapter 5, all CAC and MBCAC schemes are subjected to homogeneous and heterogeneous traffic streams scenarios. As detailed in section 4.3, four different realistic traffic traces are used to simulate real-life network data and video traffic. In addition, an M/Pareto traffic model is used to generate long connection holding-time bursts. Such long bursts are a characteristic of long range dependent traffic normally found in large networks. The uniqueness of the CAC frameworks is that it allows network providers to construct a customized MBCAC scheme based on their traffic control requirements. Hence, network providers are free to 'mix and match' different combinations of technique and different AFCM settings. In other words, network providers are not restricted to use only the combinations of technique and the AFCM settings used by the recommended MBCAC approaches mentioned in section 5.4.

6.2 Extensions

Although the work presented in this thesis provides significant research results in the area of efficient CAC strategies, we believe that there is still a need for further work in a number of areas.

Below is a list outlining the additional areas that are worth pursuing:

- *Is there a set of performance margin's multiple factor q(n) values that is generic and hence applicable to all traffic genres?* – In this thesis, we have found that the default multiple factor values are still usable as long as the submitted work is not 'more bursty' than the default traffic source's burstiness rate. However, this conclusion only holds if the submitted work belongs to the same traffic genre as the default traffic source. The consequence of this is that an a-priori set of multiple factor values must be provided for every foreseeable traffic genre. Hence, this constraint can be removed if a generic set of performance margin's multiple factor *q(n)* values can be derived.

- *How many traffic histograms are needed in order to capture intrinsic traffic load statistics?* – In this thesis, a traffic histogram holds a collection of traffic load measured from consecutive windows with each window having a fixed time-frame. Hence, it contains the amount of work submitted during a particular time-scale. In all performance studies for the histogram-based framework, we have used a fixed number of traffic histograms, i.e., five traffic histograms for five different time-scales. These time-scales are: 1, 2, 5, 10, and 100 sampling intervals. However, according to Reich's cell loss approximation algorithm, optimization is required across all time-scales. In reality, this criterion is impractical, computationally intensive, and requires large amount of storage space. Nevertheless, more work needs to be done on deriving an ideal number of time-

scales (and traffic histograms) that will capture intrinsic traffic load statistics, subject to practical computation and storage requirements.

- What is an ideal length of the smallest time-scale required to capture occasional traffic bursts, and thus enable the admission decision process to react to these bursts in the immediate future? – In this thesis, the smallest time-scale used to measure traffic arrivals is equal to one sampling interval. The length of this sampling interval is dependent on the traffic sources used by the connections in the simulated SSQ. All MBCAC approaches within the histogram-based framework use real-time traffic measurements to help compute the service bandwidth. Hence, by using time-scales that range in sizes from the smallest to the largest, these approaches will be able to react to occasional traffic bursts with higher efficiency. In light of this, more work needs to be done on deriving an ideal length of the smallest time-scale required.

- Can a standard procedure be created to ensure the 'Load and Traffic Measurements' module's Link Occupancy technique is able to give ample advance QoS breach warnings, without the possibility of suppressing achievable maximum link utilization? – In this thesis, we have proposed and studied three 'load and traffic measurements' techniques, namely, link occupancy, buffer occupancy, and cell loss conservative period. From the performance studies, the link occupancy technique generally provides threshold values that maximize link utilization whilst providing advance QoS breach warnings to the 'prudence level policy' module. However, if the chosen link occupancy threshold value is not well-suited to the arrival traffic, this threshold may actually prevent an MBCAC scheme from attaining its objective of maximum link utilization subject to meeting the QoS requirement. Hence, further work needs to be done on creating a standard procedure for deriving link occupancy threshold values that are optimized to the arrival traffic.

7 References

[Add98] R. G. Addie, "On the Applicability and Utility of Gaussian Models for Broadband Traffic," in *Proceedings of IEEE International Conference on ATM (ICATM) '98*, Colmar, France, Jun. 1998.

[Add99] R. G. Addie, "On the Weak Convergence of Long Range Dependent Traffic Processes," *Journal of Statistical Planning and Inference (JSPI)*, vol. 80, pp. 155-171, 1999.

[AMN99] R. G. Addie, P. Mannersalo, and I. Norros, "Performance Formulae for Queues with Gaussian Input," in *Proceedings of International Teletraffic Congress (ITC) 16*, Jun. 1999.

[AMS82] D. Anick, D. Mitra, and M. M. Sondhi, "Stochastic Theory of a Data-handling System with Multiple Sources," *Bell System Technical Journal*, vol. 61, no. 8, pp. 1871-1894, Oct. 1982.

[ANZ02] R. G. Addie, T. D. Neame, and M. Zukerman, "Performance Evaluation of a Queue Fed by a Poisson Pareto Burst Process," *To appear in Computer Networks*, 2002.

[AS94] S. Abe and T. Soumiya, "A Traffic Control Method for Service Quality Assurance in an ATM Network," *IEEE Journal on Selected Areas in Communications*, vol. 12, no. 2, pp. 322-331, 1994.

[ATM02a] The ATM Forum, "User-Network Interface (UNI) Signalling Specification Version 4.1," Apr. 2002.

[ATM02b] The ATM Forum, "Private Network-Network Interface (PNNI) Specification Version 1.1," Apr. 2002.

[ATM96a] The ATM Forum, "Anchorage Accord," http://www.atmforum.com/standards/anchorage.html, 1996.

[ATM99] The ATM Forum, "Traffic Management (TM) Specification Version 4.1," Mar. 1999.

[AZ94a] R. G. Addie and M. Zukerman, "An Approximation for Performance Evaluation of Stationary Single Server Queues," *IEEE Transactions on Communications*, vol. 42, no. 12, pp. 3150-3160, Dec. 1994.

[AZ94b] R. G. Addie and M. Zukerman, "Queues with Total Recall - Application to the B-ISDN," in *The Fundamental Role of Teletraffic in the Evolution of Telecommunications Networks*, vol. 1a, *Teletraffic Science and Engineering*, J. Labetoulle and J. W. Roberts, Eds.: Elsevier Science Publishers, Amsterdam, 1994, pp. 45-54.

[AZN98] R. G. Addie, M. Zukerman, and T. D. Neame, "Broadband Traffic Modeling: Simple Solutions to Hard Problems," *IEEE Communications Magazine*, vol. 36, no. 8, pp. 88-95, Aug. 1998.

[BBCDW98] S. Blake, D. Black, M. Carlson, E. Davies, and Z. Wang, "Architecture for Differentiated Services," Internet Engineering Task Force (IETF), RFC 2475, Dec. 1998.

[BBCFP02] G. Bianchi, F. Borgonovo, A. Capone, L. Fratta, and C. Petrioli, "Endpoint Admission Control with Delay Variation Measurements for QoS in IP Networks," *ACM Special Interest Group on Data Communications (SIGCOMM): Computer Communication Review*, vol. 32, no. 2, pp. 61-69, Apr. 2002.

[BCP00] G. Bianchi, A. Capone, and C. Petrioli, "Throughput Analysis of End-to-End Measurement-Based Admission Control in IP," in *Proceedings of Conference on Computer Communications (IEEE INFOCOM) '00*, Tel Aviv, Israel, Mar. 2000.

[BCS94] R. Braden, D. Clark, and S. Shenker, "Integrated Services in the Internet Architecture: an Overview," Internet Engineering Task Force (IETF), RFC 1633, Jun. 1994.

[BCV01] M. Barry, A. T. Campbell, and A. Veres, "Distributed Control Algorithms for Service Differentiation in Wireless Packet Networks," in *Proceedings of Conference on Computer Communications (IEEE INFOCOM) '01*, Anchorage, U.S.A., 22-26 Apr. 2001.

[Bel02] S. Belenki, "An Enforced Inter-admission Delay Performance-driven Connection Admission Control Algorithm," *ACM Special Interest Group on Data Communications (SIGCOMM): Computer Communication Review*, vol. 32, no. 2, pp. 31-41, Apr. 2002.

[Ben63] V. E. Benes, *General Stochastic Processes in the Theory of Queues*: Addison Wesley, 1963.

[BGRS94] B. Bensaou, J. Guibert, J. W. Roberts, and A. Simonian, "Performance of an ATM Multiplexer Queue in the Fluid Approximation Using the Benes Approach," *Annals of Operations Research*, vol. 49, pp. 137-160, 1994.

[BJS00] L. Breslau, S. Jamin, and S. J. Shenker, "Comments on the Performance of Measurement-Based Admission Control Algorithms," in *Proceedings of Conference on Computer Communications (IEEE INFOCOM) '00*, Tel Aviv, Israel, Mar. 2000.

[BJS99] L. Breslau, S. Jamin, and S. J. Shenker, "Measurement-based Admission Control: What is the Research Agenda?," in

Proceedings of IEEE/IFIP Seventh International Workshop on Quality of Service (IWQOS) '99, London, U.K., Jun. 1999.

[BKSSZ00] L. Breslau, E. W. Knightly, S. J. Shenker, I. Stoica, and H. Zhang, "Endpoint admission control: Architectural issues and performance," in *Proceedings of ACM Special Interest Group on Data Communications (SIGCOMM) '00*, Stockholm, Sweden, 28Aug.-1Sept. 2000.

[BLCT97] B. Bensaou, S. T. C. Lam, H. Chu, and D. H. K. Tsang, "Estimation of the Cell Loss Ratio in ATM Networks with a Fuzzy System and Application to Measurement-based Call Admission Control," *IEEE/ACM Transactions on Networking*, vol. 5, no. 4, pp. 572-584, Aug. 1997.

[BMN02] G. Bianchi, V. Mancuso, and G. Neglia, "Is Admission-Controlled Traffic Self-Similar?," in *Proceedings of Second International IFIP-TC6 Networking Conference (Networking 2002)*, Pisa, Italy, May 2002.

[Bou98] J. Y. Le Boudec, "Applications of Network Calculus to Guaranteed Service Networks," *IEEE Transactions on Information Theory*, vol. 44, no. 3, pp. 1087-1096, May 1998.

[BS97] F. Brichet and A. Simonian, "Measurement-based CAC for Video Applications using SBR Service," in *Proceedings of Performance Management of Complex Communication Networks (PMCCN) '97, IFIP WG 6.3 and 7.2*, Tsukuba, Japan, Nov. 1997.

[BS98] F. Brichet and A. Simonian, "Conservative Gaussian Models Applied to Measurement-based Admission Control," in *Proceedings of IEEE/IFIP Sixth International Workshop on Quality of Service (IWQOS) '98*, Napa, U.S.A., May 1998.

[BSTW95] J. Beran, R. Sherman, M. S. Taqqu, and W. Willinger, "Long-Range Dependence in Variable-Bit-Rate Video Traffic," *IEEE Transactions on Communications*, vol. 43, pp. 1566-1579, 1995.

[BW98] A. W. Berger and W. Whitt, "Effective Bandwidths with Priorities," *IEEE/ACM Transactions on Networking*, vol. 6, no. 4, pp. 447-460, Aug. 1998.

[CC96] R. G. Cheng and C. J. Chang, "Design of a Fuzzy Traffic Controller for ATM Networks," *IEEE/ACM Transactions on Networking*, vol. 4, no. 3, pp. 460-469, Jun. 1996.

[CCL99] R. 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, pp. 111-121, Feb. 1999.

[CDS02] C. A. Courcoubetis, A. Dimakis, and G. D. Stamoulis, "Traffic Equivalence and Substitution in a Multiplexer With Applications to Dynamic Available Capacity Estimation," *IEEE/ACM Transactions on Networking*, vol. 10, no. 2, pp. 217-231, Apr. 2002.

[Cha00] C. S. Chang, "Performance Guarantees in Communication Networks," Springer Verlag, 2000.

[CK00] C. Cetinkaya and E. W. Knightly, "Egress Admission Control," in *Proceedings of Conference on Computer Communications (IEEE INFOCOM) '00*, Tel Aviv, Israel, Mar. 2000.

[CKK01] C. Cetinkaya, V. Kanodia, and E. W. Knightly, "Scalable Services via Egress Admission Control," *IEEE Transactions on Multimedia*, vol. 3, no. 1, pp. 69-81, 2001.

[CL97] H. Che and S. Q. Li, "Fast Algorithms for Measurement-Based Traffic Modeling," in *Proceedings of Conference on Computer Communications (IEEE INFOCOM) '97*, Kobe, Japan, Apr. 1997.

[CLG95] S. Chong, S. Q. Li, and J. Ghosh, "Predictive Dynamic Bandwidth Allocation for Efficient Transport of Real-time VBR Video Over ATM," *IEEE Journal on Selected Areas in Communications*, vol. 13, no. 1, pp. 12-23, 1995.

[CLLRTM97] S. Crosby, I. Leslie, J. T. Lewis, R. Russell, F. Toomey, and B. McGurk, "Practical Connection Admission Control for ATM networks Based on On-line Measurements," in *Proceedings of IEEE ATM '97*, Lisboa, Portugal, Jun. 1997.

[CLMLRT97] S. Crosby, I. Leslie, B. McGurk, J. T. Lewis, R. Russell, and F. Toomey, "Statistical Properties of a Near-optimal Measurement-based CAC Algorithm," in *Proceedings of IEEE ATM '97*, Lisboa, Portugal, 1997.

[CLW94] G. L. Choudhury, D. M. Lucantoni, and W. Whitt, "On the Effective Bandwidths for Admission Control in ATM Networks," in *Proceedings of International Teletraffic Congress (ITC) 14*, Jun. 1994.

[CLW96] G. L. Choudhury, D. M. Lucantoni, and W. Whitt, "Squeezing the Most Out of ATM," *IEEE Transactions on Communications*, vol. 44, no. 2, 1996.

[CM84] H. Chen and A. Mandelbaum, "Discrete Flow Networks: Bottleneck Analysis and Fluid Approximation," *Mathematics of Operations Research*, vol. 16, no. 2, pp. 408-446, May 1984.

[Cru91a] R. Cruz, "A Calculus for Network Delay, Part I: Network Elements in Isolation," *IEEE Transactions on Information Theory*, vol. 37, no. 1, pp. 114-131, Jan. 1991.

[Cru91b] R. Cruz, "A Calculus for Network Delay, Part II: Network Analysis," *IEEE Transactions on Information Theory*, vol. 37, no. 1, pp. 132-141, Jan. 1991.

[CS02] C. T. Chou and K. G. Shin, "Analysis of Combined Adaptive Bandwidth Allocation and Admission Control in Wireless Networks," in *Proceedings of Conference on Computer Communications (IEEE INFOCOM) '02*, New York, U.S.A., Jun. 2002.

[CS98] J. Choe and N. B. Shroff, "A Central-limit-theorem-based Approach for Analyzing Queue Behavior in High-speed Networks," *IEEE/ACM Transactions on Networking*, vol. 6, no. 5, pp. 659-671, Oct. 1998.

[CSZ92] D. D. Clark, S. J. Shenker, and L. Zhang, "Supporting Real-time Applications in an Integrated Services Packet Network: Architecture and Mechanism," in *Proceedings of ACM Special Interest Group on Data Communications (SIGCOMM) '92*, Aug. 1992.

[CT95] C. S. Chang and J. A. Thomas, "Effective Bandwidths in High Speed Networks," *IEEE Journal on Selected Areas in Communications*, vol. 13, pp. 1091-1100, 1995.

[CW95] C. Courcoubetis and R. Weber, "Effective Bandwidths for Stationary Sources," *Probability in Engineering and Informational Sciences*, vol. 9, no. 2, pp. 285-294, 1995.

[DCLM89] Z. Dziong, J. Choquette, K. Q. Liao, and L. G. Mason, "Admission Control and Routing in ATM Networks," in *Proceedings of ITC Specialist Seminar*, Adelaide, Australia, Sept. 1989.

[DD95] C. Douligeris and G. Develekos, "A Fuzzy Logic Approach to Congestion Control in ATM Networks," in *Proceedings of IEEE International Conference on Communications (ICC) '95*, Seattle, U.S.A., Jun. 1995.

[DJM97] Z. Dziong, M. Juda, and L. G. Mason, "A Framework for Bandwidth Management in ATM Networks - Aggregate

Equivalent Bandwidth Estimation Approach," *IEEE/ACM Transactions on Networking*, vol. 5, no. 1, pp. 134-147, Feb. 1997.

[DKPS95] M. Degermark, T. Kohler, S. Pink, and O. Schelen, "Advance Reservations for Predicted Service," in *Proceedings of Fifth International Network and Operating Systems Support for Digital Audio and Video Workshop*, 1995.

[DLCRT95] N. G. Duffield, J. T. Lewis, N. O' Connel, R. Russell, and F. Toomey, "Entropy of ATM Traffic Streams: A Tool for Estimating QoS Parameters," *IEEE Journal on Selected Areas in Communications*, vol. 13, pp. 981-990, 1995.

[DMRW94] D. E. Duffy, A. A. McIntosh, M. Rosenstein, and W. Willinger, "Statistical Analysis of CCSN/SS7 Traffic Data from Working CCS Subnetworks," *IEEE Journal on Selected Areas in Communications*, vol. 12, no. 3, pp. 544-551, Apr. 1994.

[Duf00] N. G. Duffield, "A Large Deviation Analysis of Errors in Measurement Based Admission Control to Buffered and Bufferless Resources," *Queueing Systems*, vol. 34, no. 1-4, pp. 131-168, 2000.

[EHLMW95] A. Elwalid, D. Heyman, T. V. Lakshman, D. Mitra, and A. Weiss, "Fundamental Bounds and Approximations for ATM Mutliplexers with Applications to Video Teleconferencing," *IEEE Journal on Selected Areas in Communications*, vol. 13, pp. 1004-1016, 1995.

[EKR00] V. Elek, G. Karlsson, and R. Ronngren, "Admission Control Based on End-to-End Measurements," in *Proceedings of Conference on Computer Communications (IEEE INFOCOM) '00*, Tel Aviv, Israel, Mar. 2000.

[EM93] A. I. Elwalid and D. Mitra, "Effective Bandwidth of General Markovian Traffic Sources and Admission Control of High Speed

Networks," *IEEE/ACM Transactions on Networking*, vol. 1, no. 3, pp. 329-343, Jun. 1993.

[ENW96] A. Erramilli, O. Narayan, and W. Willinger, "Experimental Queueing Analysis with Long-range Dependent Packet Traffic," *IEEE/ACM Transactions on Networking*, vol. 4, no. 2, pp. 209-223, Apr. 1996.

[EP00] K. M. Elsayed and H. G. Perros, "Comparative Performance Analysis of Call Admission Control Schemes in ATM Networks," in *Performance Evaluation and Application of ATM Networks*, D. Kouvatsous, Ed.: Kluwer Academic Publishers, 2000, pp. 113-140.

[EP97] K. M. Elsayed and H. G. Perros, "Analysis of an ATM Statistical Multiplexer with Heterogeneous Markovian On/Off Sources and Applications to Call Admission Control," *Journal of High Speed Networks*, vol. 6, no. 2, pp. 123-139, 1997.

[EP98] K. M. Elsayed and H. G. Perros, "Traffic Management: A Review of Call Admission Control Schemes for ATM Networks (Invited Paper)," in *Froehlich/Kent Encyclopedia of Telecommunications*, M. Dekker, Ed., 1998, pp. 75-88.

[Flo96] S. Floyd, "Comments on Measurement-based Admissions Control for Controlled-load Service," Technical Report, Lawrence Berkeley Laboratory, Jul. 1996.

[FMN99] N. L. S. Fonseca, G. S. Mayor, and C. A. V. Neto, "Policing and Statistical Multiplexing of Self-Similar Sources," in *Proceedings of IEEE Global Communications Conference (GLOBECOM) '99*, Rio de Janeiro, Brazil, Dec. 1999.

[FNM00] N. L. S. Fonseca, C. A. V. Neto, and G. S. Mayor, "Statistical Multiplexing of Self-Similar Sources," in *Proceedings of IEEE*

Global Communications Conference (GLOBECOM) '00, San Francisco, U.S.A., Nov. 2000.

[FV90] D. Ferrari and D. C. Verma, "A Scheme for Real-time Channel Establishment in Wide-area Networks," *IEEE Journal on Selected Areas in Communications*, vol. 8, no. 3, pp. 368-379, 1990.

[GAN91] R. Guerin, H. Ahmadi, and M. Naghshineh, "Equivalent Capacity and its Application to Bandwidth Allocation in High-speed Networks," *IEEE Journal on Selected Areas in Communications*, vol. 9, no. 7, pp. 968-981, 1991.

[GB98] M. W. Garrett and M. Borden, "Interoperation of Controlled-Load Service and Guaranteed Service with ATM," Internet Engineering Task Force (IETF), Network Working Group, RFC 2381, Status: Proposed Standard, Aug. 1998.

[GF94] M. W. Garrett and A. Fernandez, "Variable Bit Rate Video Bandwidth Trace Using MPEG Code," Telcordia Technologies, Inc. ftp://ftp.research.telcordia.com/pub/vbr.video.trace/MPEG.data, 1994.

[GG92] R. Guerin and L. Gun, "A Unified Approach to Bandwidth Allocation and Access Control in Fast Packet Switched Networks," in *Proceedings of Conference on Computer Communications (IEEE INFOCOM) '92*, Florence, Italy, 1992.

[GGPS94] L. Georgiadis, R. Guerin, V. Peris, and K. N. Sivarajan, "Efficient Network QoS Provisioning based on Per Node Traffic Shaping," *IEEE/ACM Transactions on Networking*, vol. 4, no. 4, pp. 182-501, Aug. 1996.

[GH91] R. J. Gibbens and P. J. Hunt, "Effective Bandwidths for the Multi-type UAS Channel," *Queuing Systems*, vol. 9, pp. 17-28, 1991.

[Gib96] R. J. Gibbens, "Traffic Characterization and Effective Bandwidths for Broadband Network Traces," in *Stochastic Networks: Theory and Applications*, F. P. Kelly, S. Zachary, and I. B. Ziedins, Eds.: Royal Statistical Society, Lecture Note Series 4, Oxford University Press, 1996, pp. 169-179.

[GK97] R. J. Gibbens and F. P. Kelly, "Measurement-based Connection Admission Control," in *Proceedings of International Teletraffic Congress (ITC) 15*, Jun. 1997.

[GK99] R. J. Gibbens and F. P. Kelly, "Distributed Connection Acceptance Control for a Connectionless Network," in *Proceedings of International Teletraffic Congress (ITC) 16*, Edinburgh, U.K., Jun. 1999.

[GKK95] R. J. Gibbens, F. P. Kelly, and P. B. Key, "A Decision-theoretic Approach to Call Admission Control in ATM Networks," *IEEE Journal on Selected Areas in Communications*, vol. 13, no. 6, pp. 1101-1114, 1995.

[GKT97] M. Grossglauser, S. Keshav, and D. N. C. Tse, "RCBR: A Simple and Efficient Service for Multiple Time-scale Traffic," *IEEE/ACM Transactions on Networking*, vol. 5, no. 6, pp. 741-755, Dec. 1997.

[GT99a] M. Grossglauser and D. N. C. Tse, "A Framework for Robust Measurement-based Admission Control," *IEEE/ACM Transactions on Networking*, vol. 7, no. 3, pp. 293-309, 1999.

[GT99b] M. Grossglauser and D. N. C. Tse, "A Time-scale Decomposition Approach to Measurement-based Admission Control," in *Proceedings of Conference on Computer Communications (IEEE INFOCOM) '99*, New York, U.S.A., Mar. 1999.

[GT99c] R. J. Gibbens and Y. C. Teh, "Critical Time and Space Scales for Statistical Multiplexing in Multiservice Networks," in *Proceedings*

of International Teletraffic Congress (ITC) 16, Edinburgh, U.K., Jun. 1999.

[GTKT96] M. Grossglauser, D. N. C. Tse, J. Kurose, and D. Towsley, "A New Algorithm for Measurement-based Admission Control in Integrated Services Packet Networks," in *Proceedings of Fifth International Workshop on Protocols for High-Speed Networks*, Antipolis, France, Oct. 1996.

[GW00] R. Geist and J. Westall, "Practical Aspects Of Simulating Systems Having Arrival Processes With Long-Range Dependence," in *Proceedings of 2000 Winter Simulation Conference*, 2000.

[GW94] M. W. Garrett and W. Willinger, "Analysis, Modeling and Generation of Self-Similar VBR Video Traffic," in *Proceedings of ACM Special Interest Group on Data Communications (SIGCOMM) '94*, London, U.K., Sept. 1994.

[Hir90] A. Hiramatsu, "ATM Communications Network Control by Neural Networks," *IEEE Transactions on Neural Networks*, vol. 1, no. 1, pp. 122-130, Mar. 1990.

[HLP93] J. M. Hyman, A. A. Lazar, and G. Pacifici, "A Separation Principle Between Scheduling and Admission Control for Broadband Switching," *IEEE Journal on Selected Areas in Communications*, vol. 11, no. 4, pp. 605-616, May 1993.

[HTY97] J. M. Hah, P. L. Tien, and M. C. Yuang, "Neural-Network-based Call Admission Control for ATM Networks with Heterogeneous Arrivals," *Computer Communications*, vol. 20, no. 9, pp. 732-740, Sept. 1997.

[Hui88] J. Y. Hui, "Resource Allocation for Broadband Networks," *IEEE Journal on Selected Areas in Communications*, vol. 6, no. 9, pp. 1598-1608, 1988.

[HY97] J. M. Hah and M. C. Yuang, "Estimation-based Call Admission Control with Delay and Loss Guarantees in ATM Networks," *IEE Proceedings on Communications*, vol. 144, no. 2, pp. 85-92, Apr. 1997.

[ISO93] ISO/IEC, "Coding of Moving Pictures and Associated Audio for Digital Storage Media up to 1.5 Mbit/s - Part 2: Video," International Standard for Standardization (ISO) / International Electrotechnical Commission (IEC): IS 11172-2, 1993.

[ITU00a] ITU-T, "Methods for Cell Level Traffic Control in B-ISDN," *ITU-T Study Group 2, Recommendation E.736*, Mar. 2000.

[ITU00b] ITU-T, "Information Technology - Generic Coding of Moving Pictures and Associated Audio Information: Video," *ITU-T Recommendation H.262*, Feb. 2000.

[ITU00c] ITU-T, "B-ISDN ATM Layer Cell Transfer Performance," *ITU-T Study Group 13, Recommendation I.356*, Mar. 2000.

[ITU00d] ITU-T, "Traffic Control and Congestion Control in B-ISDN," *ITU-T Study Group 13, Recommendation I.371*, Feb. 2000.

[Jai95] R. Jain, "Congestion Control and Traffic Management in ATM Networks: Recent Advances and a Survey," *Computer Networks ISDN System*, Jan. 1995.

[JDSZ95] S. Jamin, P. B. Danzig, S. J. Shenker, and L. Zhang, "A Measurement-based Admission Control Algorithm for Integrated Services Packet Networks," in *Proceedings of ACM Special Interest Group on Data Communications (SIGCOMM) '95*, Cambridge, U.S.A., 28Aug.-1Sept. 1995.

[JDSZ97] S. Jamin, P. B. Danzig, S. J. Shenker, and L. Zhang, "A Measurement-based Admission Control Algorithm for Integrated Services Packet Networks," *IEEE/ACM Transactions on Networking*, vol. 5, no. 1, pp. 56-70, Feb. 1997.

[JS97] S. Jamin and S. J. Shenker, "Measurement-based Admission Control Algorithms for Controlled-load Service: A Structural Examination," Computer Science Department, University of Southern California CSE-TR-333-97, Apr. 1997.

[JSD97] S. Jamin, S. J. Shenker, and P. B. Danzig, "Comparison of Measurement-based Admission Control Algorithms for Controller-Load Service," in *Proceedings of Conference on Computer Communications (IEEE INFOCOM) '97*, Kobe, Japan, Apr. 1997.

[JSSWZ02] D. R. Jeske, B. Samadi, K. Sohraby, Y. T. Wang, and Q. Zhang, "QoS with an Edge-Based Call Admission Control in IP Networks," in *Proceedings of Second International IFIP-TC6 Networking Conference (Networking 2002)*, Pisa, Italy, May 2002.

[Kel91] F. P. Kelly, "Effective Bandwidths at Multi-class Queues," *Queuing Systems*, vol. 9, pp. 5-15, 1991.

[KGC94] V. G. Kulkarani, L. Gun, and P. F. Chimento, "Effective Bandwidth Vector for Two-priority ATM Traffic," in *Proceedings of Conference on Computer Communications (IEEE INFOCOM) '94*, Toronto, Canada, Jun. 1994.

[KGC95] V. G. Kulkarani, L. Gun, and P. F. Chimento, "Effective Bandwidth Vector for Multiclass Traffic Multiplexed in a Partitioned Buffer," *IEEE Journal on Selected Areas in Communications*, vol. 13, no. 6, pp. 1039-1047, 1995.

[KKZ00] F. P. Kelly, P. B. Key, and S. Zachary, "Distributed Admission Control," *IEEE Journal on Selected Areas in Communications*, vol. 18, no. 12, pp. 2617-2628, Dec. 2000.

[KM98] M. M. Krunz and A. M. Makowski, "Modeling Video Traffic Using M/G/∞ Input Processes: A Compromise between Markovian and LRD Models," *IEEE Journal on Selected Areas in Communications*, vol. 16, no. 5, Jun. 1998.

[Kni97] E. W. Knightly, "On the Accuracy of Admission Control Tests," in *Proceedings of IEEE International Conference on Network Protocols (ICNP) '97*, Atlanta, U.S.A., Oct. 1997.

[Kos84] L. Kosten, "Stochastic Theory of Data-handling Systems with Groups of Multiple Sources," in *Performance of Computer-Communication Systems*, H. Rudin and W. Bux, Eds.: Amsterdam, The Netherlands: Elsevier, 1984, pp. 321-331.

[KQ98] E. W. Knightly and J. Qiu, "Measurement-based Admission Control with Aggregate Traffic Envelopes," in *Proceedings of IEEE International Tyrrhenian Workshop on Digital Communications (ITWDC) '98*, Ischa, Italy, Sept. 1998.

[KS00] S. H. Kang and D. K. Sung, "A CAC Scheme Based on Real-time Cell Loss Estimation for ATM Multiplexers," *IEEE Transactions on Communications*, vol. 48, no. 2, pp. 252-258, Feb. 2000.

[KS93] S. H. Kang and D. K. Sung, "A Trial Multilayer Perceptron Neural Network for ATM Connection Admission Control," *IEICE Transaction on Communications*, vol. E76-B, no. 3, Mar. 1993.

[KS99] E. W. Knightly and N. B. Shroff, "Admission Control for Statistical QoS: Theory and Practice," *IEEE Network*, vol. 13, no. 2, pp. 20-29, 1999.

[KWC93] G. Kesidis, J. Walrand, and C. S. Chang, "Effective Bandwidths for Multiclass Markov Fluids and Other ATM Sources," *IEEE/ACM Transactions on Networking*, vol. 1, no. 4, pp. 424-428, Aug. 1993.

[LC99] V. Lemaire and F. Clerot, "Estimation of the Blocking Probabilities in an ATM Network Node Using Artificial Neural Network for Connection Admission Control," in *Proceedings of International Teletraffic Congress (ITC) 16*, Edinburgh, U.K., Jun. 1999.

[LCH95] S. Q. Li, S. Chong, and C. L. Hwang, "Link Capacity Allocation and Network Control by Filtered Input Rate in High Speed Networks," *IEEE/ACM Transactions on Networking*, vol. 3, no. 1, pp. 10-25, 1995.

[LEWW95] W. C. Lau, A. Erramilli, J. L. Wang, and W. Willinger, "Self-similar Traffic Generation: The Random Midpoints Displacement Algorithm and its Properties," in *Proceedings of IEEE International Conference on Communications (ICC) '95*, Seattle, U.S.A., Jun. 1995.

[LH93] S. Q. Li and C. L. Hwang, "Queue Response to Input Correlation Functions: Discrete Spectral Analysis," *IEEE/ACM Transactions on Networking*, vol. 1, no. 5, pp. 522-533, 1993.

[LH97] S. Q. Li and C. L. Hwang, "On the Convergence of Traffic Measurement and Queueing Analysis: A Statistical-match Queueing (SMAQ) Tool," *IEEE/ACM Transactions on Networking*, vol. 5, no. 1, pp. 95-110, Feb. 1997.

[Lin91] K. Lindberger, "Analytical Methods for the Traffic Problems with Statistical Multiplexing in ATM-networks," in *Proceedings of International Teletraffic Congress (ITC) 13*, Copenhagen, Holland, Jun. 1991.

[LLD96] 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, pp. 758-469, Oct. 1996.

[LM94] K. Q. Liao and L. G. Mason, "A Congestion Control Framework for Broadband ISDN Using Selective Window Control," in *Proceedings of Broadband Communications '94*, Paris, France, Mar. 1994.

[LRTMCL98] J. T. Lewis, R. Russell, F. Toomey, B. McGurk, S. Crosby, and I. Leslie, "Practical Connection Admission Control for ATM

Networks Using On-line Measurements," *Computer Communications*, vol. 21, no. 17, pp. 1585-1596, Nov. 1998.

[LTG95] N. Likhanov, B. Tsybakov, and N. D. Georganas, "Analysis of an ATM Buffer with Self-Similar ('Fractal') Input Traffic," in *Proceedings of Conference on Computer Communications (IEEE INFOCOM) '95*, Boston, U.S.A., Apr. 1995.

[LTWW93] W. E. Leland, M. S. Taqqu, W. Willinger, and D. V. Wilson, "On the Self-Similar Nature of Ethernet Traffic," in *Proceedings of ACM Special Interest Group on Data Communications (SIGCOMM) '93*, San Francisco, California, Sept. 1993.

[LZ00] T. K. Lee and M. Zukerman, "Connection Admission Control Techniques With and Without Real-time Measurements," *IEICE Transactions on Communications, IEICE/IEEE Joint Special Issue on Recent Progress in ATM Technologies*, vol. E83-B, no. 2, 25 Feb. 2000.

[LZ98] T. K. Lee and M. Zukerman, "Simple Measurement-based Connection Admission Control for Heterogeneous Traffic Sources," in *Proceedings of Fifth International Conference on Telecommunications (ICT) '98*, Chalkidiki - Porto Carras, Greece, 22-24 Jun. 1998.

[LZ99a] T. K. Lee and M. Zukerman, "An Efficiency Study of Different Model-based and Measurement-based Connection Admission Control Techniques Using Heterogeneous Traffic Sources," in *Proceedings of IEEE ATM Workshop '99*, Kochi City, Japan, 24-27 May 1999.

[LZ99b] T. K. Lee and M. Zukerman, "Practical Approaches for Connection Admission Control in Multiservice Networks," in *Proceedings of IEEE International Conference on Networks (ICON) '99*, Brisbane, Australia, 28 Sept.-1 Oct. 1999.

[LZ99c] T. K. Lee and M. Zukerman, "Efficiency Comparisons between Different Model-based and Measurement-based Connection Admission Control Schemes under Heavy Traffic," in *Proceedings of IEEE Global Communications Conference (GLOBECOM) '99*, Rio de Janeiro, Brazil, 5-9 Dec. 1999.

[LZA01] T. K. Lee, M. Zukerman, and R. G. Addie, "Admission Control Schemes for Bursty Multimedia Traffic," in *Proceedings of Conference on Computer Communications (IEEE INFOCOM) '01*, Anchorage, U.S.A., 22-26 Apr. 2001.

[LZC99] T. K. Lee, M. Zukerman, and F. Cameron, "Utilization Comparisons between Several Admission Control Schemes under Realistic Traffic Conditions," in *Proceedings of Seventh International Federation for Information Processing (IFIP) Workshop on Performance Modelling and Evaluation of ATM Networks*, Antwerp, Belgium, 28-30 Jun. 1999.

[MASKR88] B. Maglaris, D. Anastassiou, P. Sen, G. Karlsson, and J. Robbins, "Performance Models of Statistical Multiplexing in Packet Video Communications," *IEEE Transactions on Communications*, vol. 36, Jul. 1988.

[MH02] G. Mao and D. Habibi, "Loss Performance Analysis for Heterogeneous ON/OFF Sources With Application to Connection Admission Control," *IEEE/ACM Transactions on Networking*, vol. 10, no. 1, pp. 125-138, Feb. 2002.

[Mit88] D. Mitra, "Stochastic Theory of a Fluid Model of Producers and Consumers Coupled by a Buffer," *Advanc. Appl. Prob.*, vol. 20, pp. 646-676, Sept. 1988.

[Mit92] D. Mitra, "Asymptotically Optimal Design of Congestion Control for High Speed Data Networks," *IEEE Transactions on Communications*, vol. 40, pp. 301-311, Feb. 1992.

[Miy91] Y. Miyao, "A Call Admission Control Scheme in ATM Networks," in *Proceedings of International Conference on Communications (ICC) '91*, Denver, U.S.A., Jun. 1991.

[MPCC00] R. Mortier, I. Pratt, C. Clark, and S. Crosby, "Implicit Admission Control," *IEEE Journal on Selected Areas in Communications*, vol. 18, no. 12, pp. 2629--2639, Dec. 2000.

[MR99a] L. Massoulie and J. W. Roberts, "Arguments in Favour of Admission Control for TCP Flows," in *Proceedings of International Teletraffic Congress (ITC) 16*, Edinburgh, U.K., 1999.

[MV96] M. Montgomery and G. de Veciana, "On the Relevance of Time Scales in Performance Oriented Traffic Characterizations," in *Proceedings of Conference on Computer Communications (IEEE INFOCOM) '96*, San Francisco, U.S.A., Mar. 1996.

[NELC01] B. Nandy, J. Ethridge, A. Lakas, and A. Chapman, "Aggregate Flow Control: Improving Assurances for Differentiated Services Network," in *Proceedings of Conference on Computer Communications (IEEE INFOCOM) '01*, Anchorage, U.S.A., 22-26 Apr. 2001.

[Nor94] I. Norros, "A Storage Model with Self-similar Input," *Queuing Systems*, vol. 16, pp. 387-396, 1994.

[NRSV91] I. Norros, J. W. Roberts, A. Simonian, and J. Virtamo, "The Superposition of Variable Bitrate Sources in ATM Multiplexers," *IEEE Journal on Selected Areas in Communications*, vol. 9, no. 3, pp. 378-387, Apr. 1991.

[NZA02] T. D. Neame, M. Zukerman, and R. G. Addie, "Modeling Bursty Multimedia Traffic Streams Using the Poisson Pareto Burst Process," *Submitted to ACM Transactions on Modeling and Computer Simulation (TOMACS)*, Jul. 2002.

[NZA99] T. D. Neame, M. Zukerman, and R. G. Addie, "Application of the M/Pareto Process to Modeling Broadband Traffic Streams," in *Proceedings of IEEE International Conference on Networks (ICON) '99*, Brisbane, Australia, Sept. 1999.

[PE95] P. Pruthi and A. Erramilli, "Heavy-tailed On/Off Source Behavior and Self-Similar Traffic," in *Proceedings of IEEE International Conference on Communications (ICC) '95*, Seattle, U.S.A., Jun. 1995.

[PE96] H. G. Perros and K. M. Elsayed, "Call Admission Control Schemes: A Review," in *IEEE Communications Magazine*, vol. 34, 1996, pp. 82-91.

[PF95] V. Paxson and S. Floyd, "Wide-area Traffic: The Failure of Poisson Modeling," *IEEE/ACM Transactions on Networking*, vol. 3, no. 3, pp. 226-244, 1995.

[PG94] A. Parekh and R. Gallager, "A Generalized Processor Sharing Approach to Flow Control in Integrated Services Networks: The Multiple Node Case," *IEEE/ACM Transactions on Networking*, vol. 2, no. 2, pp. 137-150, Apr. 1994.

[QK01] J. Qiu and E. W. Knightly, "Measurement-based Admission Control with Aggregate Traffic Envelopes," *IEEE/ACM Transactions on Networking*, vol. 9, no. 2, pp. 199-210, 2001.

[QK98] J. Qiu and E. W. Knightly, "QoS Control via Robust Envelope-based MBAC," in *Proceedings of IEEE/IFIP Sixth International Workshop on Quality of Service (IWQOS) '98*, Napa, U.S.A., May 1998.

[Rat91] E. P. Rathgeb, "Modelling and Performance Comparison of Policing Mechanisms for ATM Networks," *IEEE Journal on Selected Areas in Communications*, vol. 9, no. 3, pp. 325-334, Apr. 1991.

[Rat93] E. P. Rathgeb, "Policing of Realistic VBR Video Traffic in an ATM Network," *International Journal of Digital and Analog Communication System*, vol. 6, pp. 213-226, Oct.-Dec. 1993.

[Reg94] K. M. Rege, "Equivalent Bandwidth and Related Admission Criteria for ATM Systems - A Performance Study," *International Journal of Communications Systems*, vol. 7, pp. 181-197, 1994.

[Rei01] M. Reisslein, "Measurement-Based Admission Control for Bufferless Multiplexers," *International Journal of Communication Systems*, vol. 14, no. 8, pp. 735-761, Oct. 2001.

[Rei58] E. Reich, "On the Integrodifferential Equation of Takacs. I.," *Ann. Math. Stat.*, vol. 29, pp. 563-570, 1958.

[Rei84] M. I. Reiman, "Open Queueing Networks in Heavy Traffic," *Mathematics of Operations Research*, vol. 9, pp. 441-458, 1984.

[RKT02] D. Rubenstein, J. Kurose, and D. Towsley, "Detecting Shared Congestion of Flows Via End-to-End Measurement," *IEEE/ACM Transactions on Networking*, vol. 10, no. 3, pp. 381-395, Jun. 2002.

[RMV96] J. Roberts, U. Mocci, and J. Virtamo, "Broadband Network Teletraffic," in *Final Report of Action Cost 242*, J. Roberts, U. Mocci, and J. Virtamo, Eds.: Berlin: Springer Verlag, 1996.

[Rob97] J. W. Roberts, "Realizing Quality of Service Guarantees in Multi-service Networks," in *Proceedings of Performance Management of Complex Communication Networks (PMCCN) '97, IFIP WG 6.3 and 7.2*, Tsukuba, Japan, Nov. 1997.

[Ros95] O. Rose, "Statistical Properties of MPEG Video Traffic and Their Impact on Traffic Modeling in ATM systems," in *Institute of Computer Science Research Report Series, Report No. 101*, vol. 20: University of Wuerzburg, 1995, pp. 397-406.

[RRR02] M. Reisslein, K. W. Ross, and S. Rajagopal, "A Framework for Guaranteeing Statistical QoS (Extended Version)," *IEEE/ACM Transactions on Networking*, vol. 10, no. 1, pp. 27-42, Feb. 2002.

[RZKJ01] A. Raha, W. Zhao, S. Kamat, and W. Jia, "Admission Control for Hard Real-time Connections in ATM LANs," *IEE Proceedings on Communications*, vol. 148, no. 4, pp. 217-228, Aug. 2001.

[Sai92] H. Saito, "Call Admission Control in an ATM Network Using Upper Bound of Cell Loss Probability," *IEEE Transactions on Communications*, vol. 40, no. 9, pp. 1512-1521, Sept. 1992.

[SCY98] K. Shiomoto, 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, pp. 625-634, Oct. 1998.

[Sim91] A. Simonian, "Stationary Analysis of a Fluid Queue with Input Varying as an Orenstein-Uhlenbeck Process," *SIAM Journal on Applied Mathematics*, vol. 51, no. 3, Jun. 1991.

[SKG92] A. Skliros, P. B. Key, and T. R. Griffiths, "CDV Tolerance Values for ATM Connections Passing Through a FIFO ATM Multiplexer," in *Practical Application Schemes, COST 242 Technical Document 062*, 1992.

[Skl93] A. Skliros, "Dimensioning the CDV tolerance parameter for ATM Multiplexers," in *Proceedings of 10th UK Teletraffic Symposium*, 1993.

[SL02] K. S. Seo and B. G. Lee, "Measurement-based Admission Control Using Maximum Burstiness," *IEEE Communications Letters*, vol. 6, no. 9, pp. 403-405, Sept. 2002.

[SPG97] S. Shenker, C. Partridge, and R. Guerin, "Specification of Guaranteed Quality of Service," Internet Engineering Task Force (IETF), RFC 2212, Sept. 1997.

[SR01] J. Siwko and I. Rubin, "Connection Admission Control for Capacity-varying Networks with Stochastic Capacity Change Times," *IEEE/ACM Transactions on Networking*, vol. 9, no. 3, pp. 351-362, Jun. 2001.

[SRLL95] C. Shim, I. Ryoo, J. Lee, and S-B. Lee, "Modeling and Call Admission Control Algorithm of Variable Bit Rate Video in ATM Networks," *IEEE Journal on Selected Areas in Communications*, vol. 2, no. 2, pp. 332-344, Feb. 1995.

[SS91] H. Saito and K. Shiomoto, "Dynamic Call Admission Control in ATM Networks," *IEEE Journal on Selected Areas in Communications*, vol. 9, no. 7, pp. 982-989, 1991.

[TG97] D. N. C. Tse and M. Grossglauser, "Measurement-Based Call Admission Control: Analysis and Simulation," in *Proceedings of Conference on Computer Communications (IEEE INFOCOM) '97*, Kobe, Japan, Apr. 1997.

[Tur87] J. S. Turner, "The Challenge of Multi-point Communication," in *Proceedings of Fifth ITC Seminar*, Lake Como, Italy, May 1987.

[UH97] K. Uehara and K. Hirota, "Fuzzy Connection Admission Control for ATM Networks Based on Possibility Distribution of Cell Loss Ratio," *IEEE Journal on Selected Areas in Communications*, vol. 15, no. 2, pp. 179-190, Feb. 1997.

[VKW95] G. de Veciana, G. Kesidis, and J. Walrand, "Resource Management in Wide-area ATM Networks Using Effective Bandwidth," *IEEE Journal on Selected Areas in Communications*, vol. 13, pp. 1081-1090, 1995.

[VW94] G. de Veciana and J. Walrand, "Effective Bandwidths: Call Admission, Traffic Policing and Filtering for ATM Networks," University of California, Berkeley, Technical Report ERL M93/47, 1994.

[WCKG94] R. Warfield, S. Chan, A. Konheim, and A. Guillaume, "Real-time Traffic Estimation in ATM Networks," in *Proceedings of International Teletraffic Congress (ITC) 14*, Jun. 1994.

[Whi93] W. Whitt, "Tail Probabilities with Statistical Multiplexing and Effective Bandwidth for Multi-class Queues," *Telecommunications System*, vol. 2, pp. 71-107, 1993.

[WKFR89] G. Woodruff, R. Kositpaiboon, G. Fitzpatrick, and P. Richards, "Control of ATM Statistical Multiplexing Performance," in *Proceedings of ITC Specialist Seminar*, Adelaide, Australia, Sept. 1989.

[Wro97] J. Wroclawski, "Specification of the Controlled-load Network Element Service," Internet Engineering Task Force (IETF), RFC 2211, Sept. 1997.

[WTE96] W. Willinger, M. S. Taqqu, and A. Erramilli, "A Bibliographical Guide to Self-similar Traffic and Performance Modeling for Modern High-speed Networks," in *Stochastic Networks*, F. P. Kelly, S. Zachary, and I. Ziedins, Eds.: Oxford University Press, Oxford, 1996, pp. 339-366.

[YHS96] S. Y. Youssef, I. W. Habib, and T. N. Saadawi, "A Neural Network Control for Effective Admission Control in ATM Networks," in *Proceedings of IEEE International Conference on Communications (ICC) '96*, Dallas, U.S.A., Jun. 1996.

[ZA94] Z. Zhang and A. S. Acampora, "Equivalent Bandwidth for Heterogeneous Sources in ATM Networks," in *Proceedings of IEEE International Conference on Communications (ICC) '94*, New Orleans, U.S.A., May 1994.

[ZF94a] H. Zhang and D. Ferrari, "Improving Utilization for Deterministic Service in Multimedia Communication," in *Proceedings of IEEE*

International Conference in Multimedia Computing and Systems, 1994.

[ZF94b] H. Zhang and D. Ferrari, "Rate-controlled Service Disciplines," *Journal of High Speed Networks*, vol. 3, no. 4, pp. 389-412, 1994.

[ZL98a] M. Zukerman and T. K. Lee, "A Measurement-based Connection Admission Control for ATM Networks," in *Proceedings of IEEE International Conference on ATM (ICATM) '98*, Colmar, France, 22-24 Jun. 1998.

[ZL98b] M. Zukerman and T. K. Lee, "A Framework for Real-time Measurement-based Connection Admission Control in Multi-service Networks," in *Proceedings of IEEE Global Communications Conference (GLOBECOM) '98*, Sydney, Australia, 8-12 Nov. 1998.

[ZT97] M. Zukerman and P. W. Tse, "An Adaptive Connection Admission Control Scheme for ATM Networks," in *Proceedings of IEEE International Conference on Communications (ICC) '97*, Montreal, Canada, Jun. 1997.