1-1-2001

Statistical multiplexing and connection admission control in ATM networks

Guoqiang Mao
Edith Cowan University

STATISTICAL MULTIPLEXING AND CONNECTION ADMISSION CONTROL IN ATM NETWORKS

By
Guoqiang Mao

A thesis submitted for the degree of
Doctor of Philosophy
at
School of Engineering and Mathematics
Faculty of Communications, Health and Science
Edith Cowan University

Principal Supervisor : Dr. Daryoush Habibi
Co-Supervisor : Professor Kamran Eshraghian

November 2001

USE OF THESIS

The Use of Thesis statement is not included in this version of the thesis.

DECLARATION

I certify that this thesis does not, to the best of my knowledge and belief:

(i) incorporate without acknowledgement any material previously submitted for a degree or diploma in any institution of higher education;

(ii) contain any material previously published or written by another person except where due reference is made in the text; or

(iii) contain any defamatory material.

Signature *Date* Nov. 30, 3[illegible]

Acknowledgements

First and foremost, I would like to express my sincerest gratitude to my principal supervisor Dr. Daryoush Habibi and co-supervisor Professor Kamran Eshraghian, for their encouragement and unconditional support. I would like to thank them particularly for the independence they gave me to choose the new challenges in my research work, and the financial support they gave me to choose the hardware and software which made my research easier and more enjoyable.

I am grateful to my parents Huapu Mao and Yuanxiu Wan, my wife Qiqian Huang for their love and encouragement.

Finally, I would like to thank the Postgraduate School and the School of Engineering and Mathematics at Edith Cowan University for providing a friendly and creative environment.

This work would not have been possible without the financial support of the following organizations:

- Edith Cowan University, through the postgraduate student research scholarship.
- Australian government, through the overseas postgraduate research scholarship.

Abstract

Asynchronous Transfer Mode (ATM) technology is widely employed for the transport of network traffic, and has the potential to be the base technology for the next generation of global communications. Connection Admission Control (CAC) is the most effective traffic control mechanism which is necessary in ATM networks in order to avoid possible congestion at each network node and to achieve the Quality-of-Service (QoS) requested by each connection. CAC determines whether or not the network should accept a new connection. A new connection will only be accepted if the network has sufficient resources to meet its QoS requirements without affecting the QoS commitments already made by the network for existing connections. The design of a high-performance CAC is based on an in-depth understanding of the statistical characteristics of the traffic sources.

The aim of this research is to investigate the statistical characteristics of the ATM traffic, and develop a computationally efficient and robust CAC scheme based on a novel closed-loop architecture which is capable of achieving high network resources utilization while satisfying the QoS requirements of all connections.

First, loss performance analysis of heterogeneous traffic sources is performed. A cell loss rate function is defined which is used to characterize the cell loss of heterogeneous traffic sources in the bufferless fluid flow model. Cell loss rate function is a convenient and effective tool to study the loss performance. The stochastic ordering theory is used to analyze the cell loss rate function. The loss performance of heterogeneous traffic sources, especially heterogeneous on-off

sources, is discussed. A set of theorems on the loss performance of heterogeneous traffic sources are presented and proved which have great significance in real applications.

Second, the application of the loss performance analysis in connection admission control is presented. A computationally efficient, measurement-based connection admission control scheme is proposed which only needs simple parameters from traffic sources, i.e. peak cell rate and/or mean cell rate. Extensive simulation is performed which is indicative of good performance of the CAC scheme. The CAC scheme is shown to perform well with regard to QoS guarantees, robust against inaccuracies in user-declared traffic parameters, and capable of achieving a high link utilization.

Third, based on the asymptotic relationship between cell loss ratio and buffer size, an in-service QoS monitoring and estimation scheme (ISME) is proposed which uses virtual buffer techniques. The proposed ISME based on virtual buffers significantly shortens the monitoring period required for making a valid QoS estimation. Simulation studies show that the proposed ISME achieves better performance than that reported in the literature.

Finally, a novel closed-loop architecture for connection admission control is proposed, which is able to overcome some inherent drawbacks of the open-loop architecture adopted by almost all CAC schemes in the literature. Based on the proposed CAC and ISME, a method of implementing the closed-loop architecture for CAC is presented. Simulation is carried out which indicates that the CAC scheme based on the closed-loop architecture is able to overcome the drawbacks of the open-loop architecture and achieve better performance.

Publications

The followings are a list of papers published during my PhD study:

1. Guoqiang Mao and Daryoush Habibi, "Loss Performance Analysis for Heterogeneous ON-OFF Sources with Application to Connection Admission Control", to appear in *IEEE/ACM Transaction on Networking*.

2. Guoqiang Mao and Daryoush Habibi, "A Cell Loss Upper Bound for Heterogeneous On-Off sources with Application to Connection Admission Control", to appear in *Computer Communications*.

3. Guoqiang Mao and Daryoush Habibi, "A Tight Upper Bound for Heterogeneous ON-OFF Sources", *IEEE Global Telecommunications Conference*, December, 2000, pp. 636-640.

4. Guoqiang Mao and Daryoush Habibi, "Heterogeneous ON-OFF Sources in the Bufferless Fluid Flow Model", *IEEE International Conference on Networks*, September, 2000, pp. 307-312.

5. Daryoush Habibi and Guoqiang Mao, "A Hybrid ATM Connection Admission Control Scheme Based on On-line Measurements and User Traffic Descriptors", *IEEE International Conference on Networks*, September, 2000, pp. 248-254.

6. Guoqiang Mao, "On Input State Space Reduction for Heterogeneous ON-OFF Sources", *Inter-University Postgraduate Electrical Engineering Symposium*, July, 2000, pp. 97-100.

Contents

Acknowledgements **v**

Abstract **vi**

Publications **viii**

1 Introduction **1**

1.1 ATM Overview 1

1.2 Traffic Control and Congestion Control 4

1.2.1 The Challenge 4

1.2.2 Objectives and Generic Functions 6

1.2.3 Preventive Versus Reactive Control 10

1.2.4 Time-Scale Decomposition 11

1.3 Connection Admission Control and Statistical Multiplexing 13

1.4 Problems Addressed by This Thesis 15

1.5 Organization of The Thesis 18

2 Literature Review **19**

2.1 Traffic-Descriptor based CAC 19

2.1.1 Effective Bandwidth Approach 20

2.1.2 Large Deviation Approach 21

2.1.3 Bufferless System Model 25

2.2 Measurement-based CAC 30

2.2.1 Distribution Measurement Approach 31

2.2.2 Leaky Bucket Approach 32
2.2.3 Asymptotic Matching Approach 34
2.2.4 Decision-Theoretical Approach 35
2.2.5 Aggregate Effective Bandwidth Approach 36
2.2.6 Instantaneous Rate Approach 41
2.2.7 Virtual Buffer Approach 43
2.2.8 Comparison 46
2.3 Self-Similarity in Network Traffic 47
2.3.1 Existence of Self-Similarity - Evidence and Possible Causes 47
2.3.2 Definition and Inference of Self-Similarity 50
2.3.3 Relevance of Self-Similarity - *The Pros and Cons* 53
2.4 Summary 57

3 Loss Performance Analysis 61
3.1 Fluid Flow Model 61
3.2 Bufferless Fluid Flow Model 63
3.3 Cell Loss Rate Function 64
3.4 Stochastic Ordering Theory 67
3.5 Traffic Source Model 69
3.6 Loss Performance Analysis of Heterogeneous On-Off Sources . . 74
3.6.1 Loss Performance Analysis of Individual On-Off Source . 75
3.6.2 A Cell Loss Upper Bound for Bernoulli Sources 78
3.6.3 A Cell Loss Upper Bound for Heterogeneous Traffic Sources 82
3.7 Summary 92

4 A Measurement-based Connection Admission Control Scheme 94
4.1 CAC Scheme 96
4.1.1 Estimation of The Activity Parameter p 98
4.1.2 Cell Loss Ratio Estimation 104
4.2 Cell Loss Ratio of Individual Connections 107
4.3 Simulation Study 109
4.3.1 Simulation Model 110

4.3.2 Saturation Scenario 111
4.3.3 Moderate Scenario 113
4.4 Robustness of The CAC Scheme 115
4.4.1 Impact of Inaccuracies in the Declared Mean Cell Rate . . 117
4.4.2 Impact of Inaccuracies in the Declared Peak Cell Rate . . 118
4.5 Application of The CAC Scheme to Real Traffic Sources 122
4.6 Summary 131

5 In-Service QoS Monitoring and Estimation 133
5.1 Relationship Between CLR and Buffer Size in ATM Networks . . 134
5.1.1 Markovian Traffic Process 135
5.1.2 Long-Range Dependent Traffic Model 137
5.2 A CLR Estimation Algorithm 139
5.2.1 Estimation of Parameters α and β 142
5.2.2 Low-pass FIR Filter 144
5.2.3 Choice of Parameters 145
5.3 Simulation Study 149
5.3.1 Markovian Scenario 150
5.3.2 Self-Similar Scenario 154
5.4 Summary 156

6 Connection Admission Control - Closing the Loop 158
6.1 Architecture of the Closed-Loop CAC Scheme 160
6.2 Introduction to Fuzzy Systems 164
6.3 Control System Design 168
6.4 Simulation Study 172
6.4.1 Simulation Using Exponential On-Off Sources 174
6.4.2 Simulation Using Real Traffic Sources 179
6.5 Summary 183

7 Conclusion 189
7.1 Contributions of The Thesis 190

7.2 Future Work 192
7.2.1 Future Work on Traffic Measurements 192
7.2.2 Future Work on FKBC 193
7.2.3 Future Work on Engineering The Buffer Size 194

List of Tables

4.1 Parameters of the three traffic types in saturation scenario 112
4.2 Parameters of the three traffic types in moderate scenario 116
4.3 Traffic rate of the M-JPEG encoded movies ($bytes/frame$) 126
5.1 Traffic parameters of the exponential on-off source 151
6.1 Parameters of the three traffic types in the saturation scenario . . . 175
6.2 Parameters of the three traffic types in the moderate scenario . . . 177
6.3 Traffic rate of the M-JPEG encoded movies ($bytes/frame$) 183

List of Figures

1.1 ATM-based network hierarchy 2
1.2 ATM connection relationship 3
1.3 Hierarchical layer-to-layer relationship 4
1.4 Call establishment using virtual paths 5
1.5 Hierarchy of time scales 12
1.6 Effect of statistical multiplexing 15
2.1 Architecture of traditional CAC scheme 59
2.2 Architecture of CAC scheme using in-service monitoring 59
3.1 Fluid flow model 62
3.2 Cell loss in bufferless fluid flow model 64
3.3 On-off source model 72
3.4 Illustration of $F(y)$ and $G(y)$ 76
4.1 Variance time plot of the aggregate traffic rate in the saturation scenario 102
4.2 Relationship between update interval and measurement interval . . 103
4.3 Utilization achieved in the saturation scenario 113
4.4 Cell loss ratio observed in the saturation scenario 114
4.5 The number of the three traffic types on the link observed in the saturation scenario 115
4.6 Comparison between admissible region and actually admitted calls 116
4.7 Statistical multiplexing gain achieved in the saturation scenario . . 117
4.8 Cell loss ratio observed in the moderate scenario 118

4.9 Number of the three traffic types on the link observed in the moderate scenario 119
4.10 Call blocking ratios of the three traffic types in the moderate scenario 120
4.11 Statistical multiplexing gain achieved in the moderate scenario . . 121
4.12 Impact of inaccuracies in declared mcr on utilization 122
4.13 Impact of inaccuracies in declared mcr on cell loss ratio 123
4.14 Impact of inaccuracies in declared pcr on utilization 124
4.15 Impact of inaccuracies in declared pcr on cell loss ratio 125
4.16 Cell loss ratio observed in the video source scenario 127
4.17 Number of each type of calls multiplexed on the link in the video source scenario Part I 128
4.18 Number of each types of calls multiplexed on the link in the video source scenario Part II 129
4.19 Variance time plot of beauty and the beast vbr video source . . . 130
4.20 Variance time plot of the aggregate traffic rate 131
5.1 System model of the CLR estimation algorithm 141
5.2 Amplitude-frequency response of the FIR filter 145
5.3 Architecture of the CLR estimation algorithm 146
5.4 Cell level, burst level and call level cell loss ratio 147
5.5 Simulation model for the CLR estimation algorithm 150
5.6 CLR estimation in the simulation with a buffer size of 300 cells . 152
5.7 Cell loss ratio estimation in the Markovian scenario 153
5.8 Parameters α and β in the simulation with a buffer size of 300 cells 154
5.9 CLR estimation in a simulation with a buffer size of 500 cells . . . 155
5.10 Cell loss ratio estimation in the Self-similar scenario 156
6.1 Architecture of the open-loop CAC scheme Part I 159
6.2 Architecture of the open-loop CAC scheme Part II 159
6.3 Controlling effect of the safety margin 163
6.4 Closed-loop architecture for CAC 164
6.5 Architecture of a fuzzy knowledge based controller 165
6.6 Architecture of the control system 168

6.7 The fuzzy sets ZO and NZ on the domain [0, 1] for call blocking ratio 169
6.8 Fuzzy sets NL, NS, ZO, PS and PL for ε 170
6.9 Moving average of utilization achieved in the saturation scenario . 175
6.10 CLR observed in the saturation scenario 176
6.11 Call number in the saturation scenario 177
6.12 Safety margin parameter in the saturation scenario 178
6.13 Moving average of utilization achieved in the moderate scenario . 179
6.14 Call number in the moderate scenario 180
6.15 Call blocking ratio in the moderate scenario 181
6.16 Safety margin in the moderate scenario 182
6.17 Moving average of utilization achieved in the real traffic source scenario 184
6.18 CLR observed in the real traffic source scenario 185
6.19 Number of each type of calls multiplexed on the link in the real traffic source scenario Part I 186
6.20 Number of each types of calls multiplexed on the link in the real traffic source scenario Part II 187
6.21 Safety margin in the real traffic source scenario 188

Chapter 1

Introduction

1.1 ATM Overview

Bandwidth-hungry computer and communication applications are on the rise with a variety of services, such as the World Wide Web, video conferencing, video on demand, and high-definition television (HDTV), which all require large amounts of network resources for their transmission. Several types of high-speed/high-bandwidth networks exist. Among them, asynchronous transfer mode (ATM) technology is proving to be one of the best technical and commercial solutions, and is able to satisfy this growing demand for heterogeneous communications and dramatically increased bandwidths.

ATM is a universal communication service based on small 53-octet cells which are routed through fixed paths with shared and multiplexed links and nodes. It provides a sequence-preserving, connection-oriented cell transfer service between source and destination with an agreed Quality of Service (QoS) and throughput [1, pp. 6]. ATM has already been widely employed for the transport of network traffic, and has the potential to be the base technology for the next generation of global communications.

The advantage of ATM lies in its efficient use of network resources, high throughput switches, and its potential ability to guarantee QoS. ATM technology is expected to support a wide variety of services and applications, and satisfy

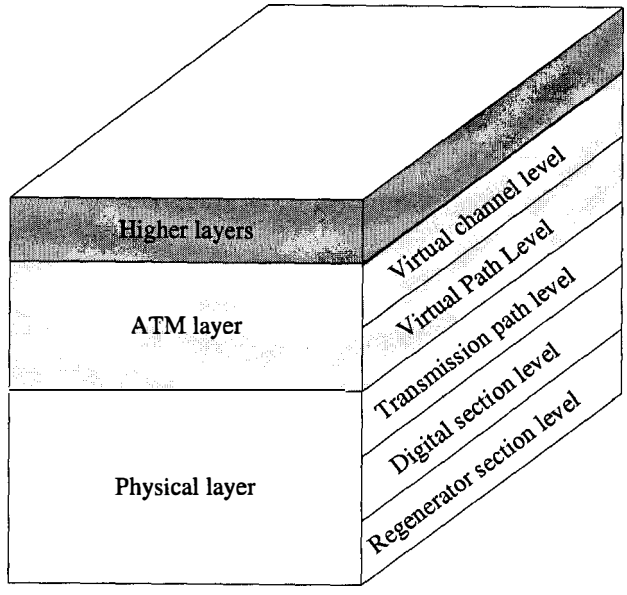


Figure 1.1: ATM-based network hierarchy

a range of user needs and network performance objectives. It allows flexibility in the choice of connection bit rates and enables the statistical multiplexing of variable bit rate (VBR) traffic streams. Thus, ATM provides a universal bearer service for broadband integrated services digital networks (B-ISDN), which can carry voice, data, and video with the same cell transport arrangement.

Fig. 1.1 shows the overall functional hierarchy of an ATM-based network. The ATM layer consists of virtual channel (VC) and virtual path (VP) levels. Logical connections in ATM are referred to as virtual channel connections (VCCs). A VCC is set up between two end users through the network, and a full-duplex flow of ATM cells is exchanged over the connection. VCCs are also used for user-network exchange (for control signalling) and network-network exchange (for network management and routing).

For ATM, a second sublayer of processing provides for virtual paths. It is illustrated in Fig. 1.2. A virtual path connection (VPC) is a bundle of VCCs that have the same endpoints. Thus, all the cells flowing over all of the VCCs in a single VPC are switched along the same route. Fig. 1.3 shows the hierarchical relationship between VCC and VPC, as well as sublayers in physical layer of

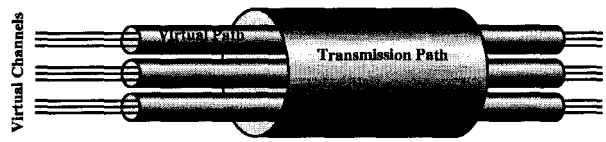


Figure 1.2: ATM connection relationship

ATM networks.

The virtual path concept was developed in response to a trend in high speed networking in which the control cost of the network is becoming an increasingly higher proportion of the overall network cost [2]. The virtual path technique helps containing the control cost by grouping connections sharing common paths through the network into a single unit. Network management actions can then be applied to a small number of connection groups instead of a large number of individual connections. Some of the advantages which can be realized through the use of the virtual path concept are [3]:

- simplified connection admission (involving only VPC terminators),
- simplified routing at transit nodes (i.e. based on Virtual Path Identifier (VPI) only),
- adaptability to varying traffic and network failures through dynamic resource management [2],
- the ability to implement priority control by segregating traffic with different quality of service requirements.

Fig. 1.4 shows the general call establishment process using virtual channels and virtual paths [4, Chapter 16]. First, a connection setup request is sent from the customer terminal to its local network node. Control intelligence in the network node analyzes the source model and destination address to determine if a VPC satisfying the requirements of the connection request exists. The capacity available on the VPC is checked. If there are sufficient network resources to support the QoS requirements of the connection request, as well as all VCCs in the VPC,

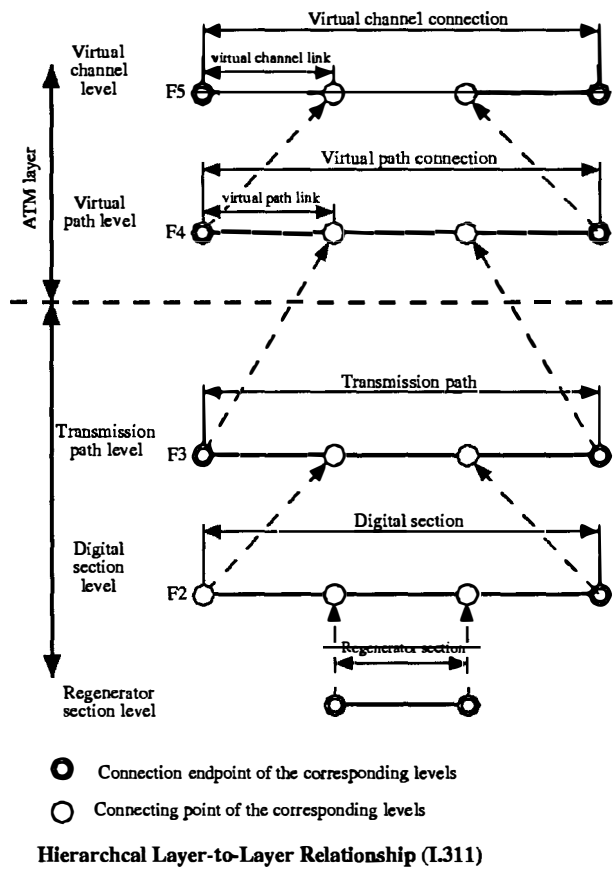


Figure 1.3: Hierarchical layer-to-layer relationship

the VCC is admitted. In this case there is no need for any call processing at transit network nodes. VCCs supported on core-network VPCs can be established with less effort than VCCs carried on the network using only virtual channel switching.

1.2 Traffic Control and Congestion Control

1.2.1 The Challenge

The principle of ATM itself guarantees neither high utilization nor stringent QoS without traffic and congestion control. However, it enables sophisticated traffic control and congestion control which are able to achieve high utilization and

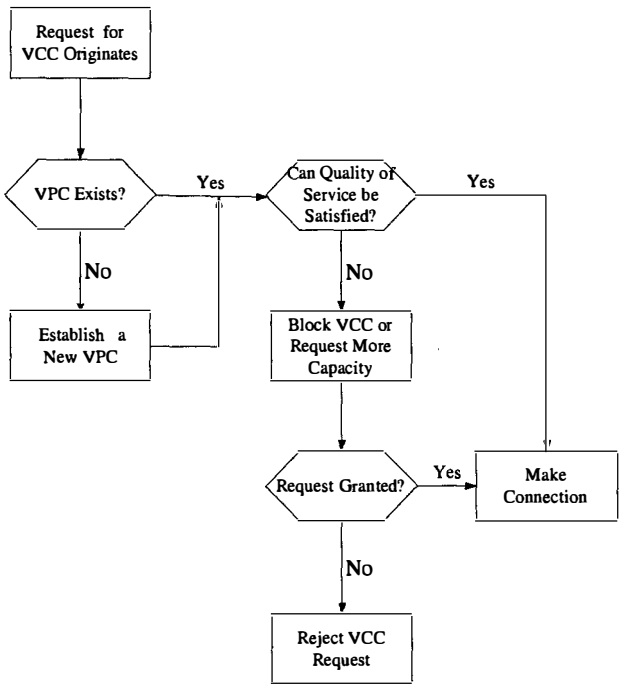


Figure 1.4: Call establishment using virtual paths

stringent QoS.

Congestion in its various forms is the basic problem of traffic control in the ordinary telephone network, in packet network, as well as in the ATM networks. Congestion occurs when the demand is greater than the available resources. According to Jain [5], congestion is caused in packet networks:

- by a shortage of a buffer space,
- by slow links, or
- by slow processors.

This may lead to a belief that, when some or all of these problems are solved by technical development (cheap memory, high speed links and fast processors), the congestion problem is eliminated. However, contrary to this belief, without proper traffic and congestion control scheme, technical development may lead to more congestion and thus reduce performance [5]. This is indisputably the situation

in ATM networks as well. A vast effort is directed towards the development of proper control schemes for ATM. The challenge is to design simple and efficient controls while still achieving reasonable bandwidth utilization through statistical multiplexing [6].

1.2.2 Objectives and Generic Functions

This section depicts the role of traffic control and congestion control as defined in ATM User Network Interface (UNI) Specification Version 3.1 [1] by ATM Forum. The primary role of traffic control and congestion control is to protect the network and the user in order to achieve network performance objectives. An additional role is to optimize the utilization of network resources.

Traffic control refers to the set of actions taken by the network to avoid congestion. Congestion control refers to the actions taken by the network to minimize the intensity, spread and duration of congestion. Congestion is defined as a state of network elements in which the network is not able to meet the network performance objectives. It is to be distinguished from the state where buffer overflow is causing cell losses, but still meets the negotiated QoS.

Under normal operation, i.e. when no network failures occur, traffic control functions are intended to avoid network congestion. However, congestion may still occur because of inability of traffic control functions to deal with unpredictable statistical fluctuations of traffic flows, or because of network failures. Therefore, congestion control functions are intended to react to network congestion in order to minimize its intensity, spread and duration.

Traffic control functions include [1]:

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

- Connection Admission Control (CAC): the set of actions taken by the network during the call set-up phase (or during call re-negotiation phase) in

order to establish whether a virtual channel/virtual path request can be accepted or rejected (or whether a request for re-allocation can be accommodated).

- Usage Parameter Control (UPC): the set of actions taken by the network to monitor and control the traffic to ensure that it does not violate the traffic contract. The main purpose of UPC is to protect network resources from malicious as well as unintentional misbehavior, which can affect the QoS of other established connections, by detecting violation of negotiated parameters and taking appropriate actions.

- Selective Cell Discarding: network may selectively discard cells with lower cell loss priority while still meeting network performance objectives on both the low priority flow and high priority flow.

- Traffic Shaping: traffic shaping mechanism may be used to achieve a desired modification of the traffic characteristics.

- Explicit Forward Congestion Indication (EFCI): the EFCI is a congestion notification mechanism that the ATM layer service user may make use of to improve the utility that can be derived from the ATM layer.

Congestion control functions include [1]:

- Selective Cell Discarding: A congested network element may selectively discard cells identified as belonging to a non-compliant ATM connection and/or those cells with lower Cell Loss Priority.

- Explicit Forward Congestion Indication (EFCI).

VPCs plays a key role in Network Resource Management. By reserving capacity on VPCs, the processing required for establishing individual VCCs is reduced. Individual VCCs can be established by making simple connection admission decisions at nodes where VPCs are terminated. Moreover, service separation can be realized by using VPCs. Services with similar traffic characteristics and service

requirements can be grouped together using VPCs. The service separation has an important effect on CAC because the multiplexing process is more regular if the traffic characteristics of aggregated streams, such as peak rate and burst length, resemble each other. Furthermore, it might be possible to use simpler CAC methods if sources are grouped into few service classes. It is worth noting that service separation and capacity reservation may decrease the network resources utilization. Strategies for service separation and capacity reservation will be determined by evaluating the trade-off between increased capacity costs and reduced control costs.

The UPC is required at the UNI to protect network resources from malicious as well as unintentional misbehavior. Traffic sent by user is monitored to check whether the traffic contract between source and network is conformed. Non-conforming cells are either discarded or tagged. If a cell is tagged, it is still transmitted to the end user when network is not congested. However, in case of congestion, the tagged cell will be discarded prior to other conforming cells.

ATM networks make it possible to use traffic shaping to change the traffic stream before multiplexing, mainly in order to increase the utilization of network links, in particular when the burstiness of offered traffic is very high [7], [8]. The usefulness of traffic shaping depends on the time-scale of variations and on the delay requirements of application.

In EFCI, a network element in an impending-congested state or a congested state may set an explicit forward congestion indication in the cell header. The end user's Customer Premises Equipment (CPE) examining this indication may use this indication to implement protocols that adaptively lower the cell rate of the connection during congestion or impending congestion. Congestion indication can also be sent backwards by using Backwards Explicit Congestion Notification (BECN). When a queue in an ATM switch exceeds a certain threshold the network node sends BECN cells back to the sources causing the congestion. In case of severe congestion, the network node sends BECN cells to all sources sending their cells to the congested queue. On receipt of a BECN cell to a particular virtual channel, a source must reduce its transmission rate for the indicated channel.

If no BECN cells are received for a certain period of time, a source may gradually restore its transmission rate. According to Newman [9], BECN could be applicable for very bursty sources without specifying traffic characteristic for every individual data source when the transmission delay is limited, as in LAN. But considerable problems might arise if the network size is large [10]. A good performance level might be expected in a network of up to several hundred kilometers.

In addition to the above traffic control and congestion control functions, there are some additional control functions:

- connection admission control that reacts to and takes account of the measured load on the network.
- variation of usage monitored parameters by the network. For example, reduction of the peak rate available to the user.
- other traffic control techniques (e.g. re-routing, connection release, OAM functions) are for further study.
- fast resource management (FRM).

CAC that reacts to and takes account of the measured load on the network, or called measurement-based CAC, has attracted a lot of research efforts in recent years. It is the subject of this thesis and will be introduced later in detail. Traffic contract re-negotiation is also under intense research. Traffic source or network may re-negotiate the traffic contract to increase or reduce network resources available to sources to make efficient utilization of network resources.

The idea of FRM is to increase the multiplexing efficiency by implementing admission control at the burst scale (or at the rate-variation scale) in addition to the call scale. When a new burst is to be sent it is necessary to obtain a new resource allocation by means of a rapid in-band signalling exchange between user and successive network nodes. The bandwidth used by a connection is relinquished at the end of a burst. According to Roberts [11], the drawbacks of this approach are the time needed to obtain a new resource allocation which reduces

efficiency particularly for short bursts, the need to implement a sophisticated protocol and the low network utilization realizable when the connection peak rate is high. If a network node rejects the burst, it can be either buffered at the network interface or discarded depending on the application. In both cases the rejection probability should be reasonably low to avoid enormous buffers or degradation of QoS.

1.2.3 Preventive Versus Reactive Control

Traffic control and congestion control functions can be classified into two basic approaches: preventive and reactive. The preventive approach depends mainly on traffic control functions while the reactive approach utilizes primarily congestion control functions.

Preventive approach prevents the occurrence of congestion by predicting the network behavior when a new connection is admitted into the network. The new connection is admitted into the network when there are sufficient network resources available, otherwise, it is rejected or required to reduce its traffic rate. Through this mechanism, the preventive approach attempts to avoid the occurrence of network congestion.

In reactive approach, traffic sources are allowed to increase their traffic rates when there are enough network resources. When congestion occurs or is pending, the traffic sources are notified to reduce their traffic rates to relieve the network congestion. The cells that are lost due to congestion are re-transmitted.

Reactive control is useful in data networks where traffic sources are suitable for traffic rate re-allocation, buffers in network nodes are typically large, and traffic rates are not very high. However, reactive control is inadequate for ATM networks for the following reasons:

- ATM networks typically support a wide range of applications with quite different bandwidth and QoS requirements. Much of the traffic is not amenable to flow control [12]. For example, video and voice traffic sources cannot stop generating cells when the network is congested.

- Feedback that is required for reactive approach is slow compared to propagation delays across the network, due to the drastically reduced cell transmission time. By the time a reactive scheme responds to the network congestion, a large amount of traffic may have already been injected into the network, which makes a reactive scheme obsolete.

- The very high speeds in switching and transmission make ATM networks more volatile in terms of congestion and traffic control. A scheme that relies heavily on reacting to changing conditions will produce extreme and wasteful fluctuations in routing policy and flow control.

Therefore, ATM networks are dependent on the capabilities of preventive control methods, but reactive control functions can still be useful in minimizing the intensity and duration of congestion. In addition, reactive functions may have an important role when exploiting the free capacity in ATM networks. Because of the statistical properties of traffic in ATM networks, the mean load of high priority traffic may remain low. Network operators may attempt to make better utilization by offering the remaining capacity to traffic sources with much less stringent QoS requirements, e.g. these traffic that can tolerate occasional long delays.

A good compromise, as Ramamurthy and Dighe [13] have proposed, would be an aggressive preventive control strategy that uses network resources optimally, with reactive control mechanism as backup to relieve congestion in the unlikely event of the network experiencing congestion.

1.2.4 Time-Scale Decomposition

Hui suggests a time-scale decomposition approach for traffic and congestion control [14]. This time-scale decomposition framework is adopted by many later literature [15], [16], [17]. This decomposition is based on the qualitatively different nature of the system at three different time scales along with the partial decoupling of the problems at these scales. The hierarchy of time scales - call scale, burst scale and cell scale is illustrated in the Fig. 1.5.

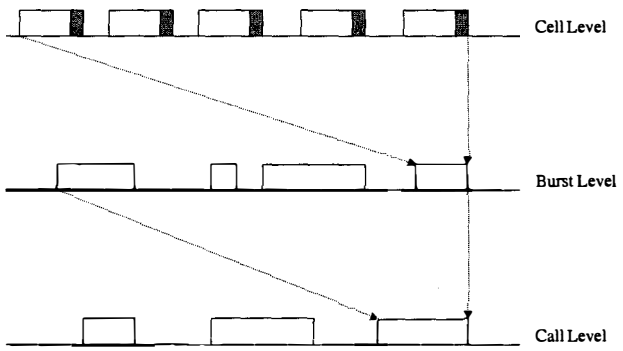


Figure 1.5: Hierarchy of time scales

At the cell time scale, the traffic consists of discrete entities-the cells, produced by each source at a rate which is often orders of magnitude lower than the transmission rate of the output trunk. Most ATM source streams, even from variable bit rate sources, are *locally periodic* when considered at cell scale (the rate of a traffic source is considered to be constant during a tiny time period, therefore it generates cells periodically into the network) [18]. Thus the randomness of the total input emerges from the independence of the phase of the locally periodic source streams, not from rate fluctuations of these streams. Cell level fluctuations due to the randomness of cell arrivals from traffic sources can be effectively absorbed by a small buffer, and is not a major concern in traffic and congestion control.

At the burst time scale the finer cell scale granularity is ignored and the input process is characterized by its instantaneous rate. Consequently, fluid flow model appears as a natural modeling tool. At this time scale problems arise from possible excesses of the total input rate over the output rate. Two approaches can be distinguished in this context depending on the multiplexing option adopted in the system design. Using *Rate Envelop Multiplexing*, buffers are provided only for cell scale queueing and the system appears as a bufferless system at the burst scale. In the *Rate Sharing Multiplexing*, buffers are provided to absorb at least part of the burst scale traffic fluctuations and the system appears as a buffered system at the burst scale. Both approaches are adopted in performance analysis of network traffic. Traffic and congestion control functions at burst time scale include CAC, UPC, traffic shaping and EFCI. These preventive and reactive control

functions cooperate with each other to control traffic in the network such that QoS requirements, such as cell loss ratio (CLR), cell transfer delay (CTD) and cell delay variation (CDV), of connections are satisfied while at the same time network resources utilization is maximized.

Call time scale is characterized by the holding time of connections and represents the longest of the time scales, and stochastic arrival pattern of connections. Traffic and congestion control functions at call time scale include NRM. By network resources engineering and reservation, load balancing, etc., call level QoS requirements such as call blocking ratio of each traffic class are satisfied.

This research follows the time-scale decomposition approach suggested by Hui [14]. Since connection admission control is the subject of this research, our network performance analysis naturally rests at burst time scale, or sometimes referred to as rate variation level in literature, and fluid flow model is used in our performance analysis. Although as shown later call level QoS parameter, i.e. call blocking ratio, is considered in our CAC scheme, it is assumed that CAC cannot control the call arrival pattern and the available network resources. Therefore, the call level QoS parameter is controlled by some call level traffic and congestion control functions like NRM, not by CAC. Some other burst time scale traffic and congestion control functions may improve the performance of CAC, however, they are not subject of this research.

1.3 Connection Admission Control and Statistical Multiplexing

Connection admission control is a preventive control mechanism and it proves to be the most effective traffic control mechanism. CAC determines whether the network should admit a new connection. A new connection will only be admitted if the network has sufficient resources to meet its QoS requirements without affecting the QoS commitment already made by the network for existing connections. After a new connection is admitted, network resources, i.e. bandwidth and buffers, are allocated to it. Resources allocation is tightly associated with

CAC. CAC determines whether the network has enough resources to meet the QoS requirements of the new connection as well as all existing connections, then resources allocation allocates resources to the new connection accordingly, if the new connection is admitted. There are two alternative approaches for CAC and resources allocation: deterministic and statistical multiplexing.

In deterministic multiplexing, each connection is allocated a bandwidth equal to its peak cell rate (pcr). Deterministic multiplexing is simple to implement. However since much of the traffic in ATM networks is bursty, it may cause large amounts of bandwidth to be wasted, particularly for those traffic with large peak to mean cell rate ratios.

An alternative approach is statistical multiplexing. In ATM networks all services are provided through the same cell transport arrangement. This makes statistical multiplexing feasible. In statistical multiplexing approach, the amount of bandwidth allocated in the network to a variable bit rate source is less than its peak cell rate, but necessarily greater than its mean cell rate (mcr). Statistical multiplexing approach is more difficult to implement than deterministic multiplexing but it leads to huge savings in network resources. Statistical multiplexing approach enables ATM networks to achieve high network resources utilization.

Fig. 1.6 illustrates the effect of statistical multiplexing. In deterministic multiplexing approach, bandwidth $pcr_1 + pcr_2$, which is the sum of peak cell rates of traffic source 1 and source 2, is allocated to source 1 and source 2 respectively. However, as shown in Fig. 1.6, when source 1 and source 2 are statistically multiplexed, their bandwidth requirement is actually pcr_{12}, which is less than $pcr_1 + pcr_2$. This behavior, referred to as statistical multiplexing effect, is utilized in statistical multiplexing approach to make efficient utilization of network resources. When there are many connections multiplexed together, the bandwidth savings that can be achieved by statistical multiplexing approach becomes very large.

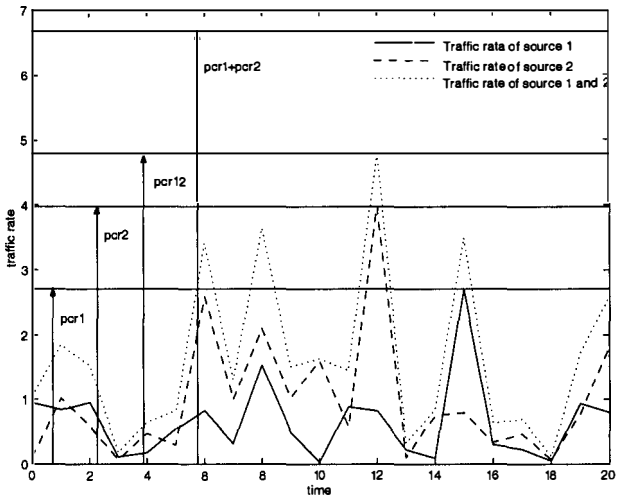


Figure 1.6: Effect of statistical multiplexing

1.4 Problems Addressed by This Thesis

In statistical multiplexing approach the network resources required to satisfy the QoS requirements of a connection depend not only on traffic characteristics of the connection itself, but also on the characteristics of other connections in the network. Therefore, a high performance CAC is based on an in-depth understanding of the statistical characteristics of multiplexing of traffic sources, and the relationship between these traffic characteristics and QoS parameters. This thesis investigates the statistical characteristics of the multiplexing of traffic sources and its application in CAC. The objective of this thesis is to improve our understanding of the statistical characteristics of ATM traffic and develop a time-efficient, robust CAC scheme based on a novel closed-loop architecture, which is capable of achieving high network resources utilization while at the same time satisfying the QoS requirements of all connections.

First, the characteristics of the CAC schemes in the literature are analyzed. It is pointed out that these CAC schemes, although varying in details, all employ open-loop control. There are fundamental drawbacks in the open-loop architecture. Because of the heterogeneity and complexity of network traffic, it is very difficult to model network traffic and obtain model parameters accurately. CAC scheme

based on the open-loop architecture lacks the ability to adjust its performance to account for model errors.

In this thesis, starting from a simple on-off traffic source model and a bufferless fluid flow model, the statistical characteristics and loss performance of heterogeneous network traffic are investigated. A set of theorems on the loss performance of heterogeneous network traffic are proposed and proved which have great significance in real applications. On the basis of the loss performance analysis, a measurement-based CAC scheme is proposed. The proposed measurement-based CAC scheme is able to provide robust QoS guarantee, and achieve a significantly higher utilization than what can be achieved by CAC schemes using effective bandwidth approach. The proposed CAC scheme is also robust against inaccuracies in user declared traffic parameters.

Moreover a novel closed-loop architecture is proposed for implementing the CAC scheme. An in-service QoS monitoring algorithm is developed which is employed to provide feedback to the closed-loop CAC scheme on its performance. A fuzzy knowledge based controller (FKBC) is designed which uses the estimated CLR from the in-service QoS monitoring algorithm as input. The FKBC controls the performance of the CAC scheme by controlling the safety margin parameter in the measurement-based CAC scheme. Simulation indicates that the closed-loop architecture is able to overcome the inherent drawbacks of the open-loop architecture and achieve significantly higher network resources utilization. More importantly, our research reveals a new direction in the development of CAC schemes. We believe that this is a promising approach to eventually solve the problem of CAC in ATM networks.

More specifically, the following four areas are investigated in the thesis:

1. Loss performance of heterogeneous network traffic is investigated. A methodology is proposed which facilitates loss performance analysis of network traffic. This methodology is applied to analyze the loss performance of heterogeneous network traffic. A set of theorems about loss performance of heterogeneous network traffic are proposed and a cell loss upper bound is developed for the heterogeneous network traffic. The application of the loss

performance analysis is demonstrated.

2. A measurement-based CAC (MBAC) scheme is designed based on the loss performance analysis. Due to consideration on practical applications of a MBAC scheme, principles are proposed which govern the design of a MBAC. It is proposed that robustness, flexibility and simplicity are major concerns in CAC scheme design. A MBAC scheme is designed which only needs simple traffic parameters from sources. It is easy to implement, robust with regard to QoS guarantees, robust against inaccuracies in user-declared traffic parameters, and capable of achieving a high network utilization in a network with small buffers. All these features make the proposed MBAC scheme an attractive option for real implementation. Moreover, the impact of self-similarity on the performance of the MBAC scheme is discussed and analyzed.

3. An in-service QoS monitoring and estimation scheme (ISME) is developed. In ATM networks, a prime concern is to ensure that there are adequate resources to meet QoS requirements of traffic sources. In order to achieve good traffic and congestion control, and effective network management, in-service monitoring and estimation has to be employed as opposed to the conventional out-of-service monitoring and testing techniques [19]. One potential problem for ISME is that some QoS indicators are specified in terms of the probability of occurrence of certain rare events. Monitoring of such rare events needs quite a long time. The statistical information obtained after a long monitoring period may be obsolete and the network management system reaction may be too late. In this thesis an efficient and reliable QoS monitoring and estimation scheme is developed which reduces the required monitoring period significantly.

4. A closed-loop connection admission control architecture is proposed. To date, almost all CAC schemes employ open-loop control. The performance of these CAC schemes relies on the precision of traffic models and traffic

parameters. However, in ATM networks, due to heterogeneity and complexity of network traffic, it is very difficult to model traffic sources and obtain model parameters accurately. This research introduces ISME into CAC design and develops a closed-loop CAC scheme which is able to overcome inherent drawbacks of an open-loop architecture.

In ATM networks, QoS is usually measured by CLR, CTD and CDV. CLR is defined as the ratio of lost cells due to buffer overflow to total transmitted cells. CTD is defined as the elapsed time between a cell exit event at the source UNI and the corresponding cell entry event at the destination UNI for a particular connection. CDV is the variation in cell transfer delay. In this thesis we consider that CTD and CDV can be controlled within the desired bounds by properly engineering the network buffer size [20], [21], [22], [23], [24], [25]. That is, the network buffer size is determined according to the delay constraint of the traffic. Therefore CLR, which is a measure of the portion of traffic overflowing at the buffer, is used as the QoS index.

1.5 Organization of The Thesis

The rest of the thesis is organized as follows: Chapter 2 presents a review of related research in this area. Chapter 3 presents the traffic model and network model used in this study. A methodology for loss performance analysis of heterogeneous network traffic is proposed. A set of theorems are proposed regarding the loss performance of heterogeneous network traffic. In Chapter 4, the analytical results obtained in Chapter 3 are used to design a CAC scheme which suits real-time implementation in ATM networks. An in-service QoS monitoring scheme is developed in Chapter 5. Chapter 6 proposes a closed-loop CAC architecture. Finally, thesis summary and conclusions are given in Chapter 7, as well as suggestions for further extensions to the work presented in this thesis.

Chapter 2

Literature Review

The design of a high-performance CAC is based on an in-depth understanding of the statistical characteristics of the multiplexing of traffic sources. To date, many connection admission control schemes have been proposed [26], [27], [28], [23], [29], [30], [16], [22]. These schemes can be classified into two basic categories:

1. traffic descriptor-based CAC, and
2. measurement-based CAC.

2.1 Traffic-Descriptor based CAC

Traffic descriptor-based CAC uses the *a priori* traffic characterizations provided by sources at connection setup phase to determine whether or not the new connection can be admitted in addition to all existing ones. It is able to achieve high network utilization when traffic descriptions required by the CAC scheme are tight. Earlier CAC schemes are almost all traffic descriptor-based. Among them, a large family of traffic descriptor-based CAC schemes adopt the so-called equivalent bandwidth approach.

2.1.1 Effective Bandwidth Approach

Anick *et al.* lay the mathematical basis for an effective bandwidth approach in [31], which is one of the earliest papers discussing the statistical characteristics of the multiplexing of traffic sources. Based on the analysis of the statistical characteristics of homogeneous independent exponential on-off sources in a network model referred to as the fluid flow model, they derive the differential equations describing the equilibrium buffer distribution and the solutions to these differential equations. Their analysis shows that the overflow probability of traffic sources in a queue whose buffer size is x can be reasonably approximated by an exponential function:

$$P(Q > x) \sim Ae^{-rx}, \tag{2.1}$$

where Q is steady state queue length, A is a constant called the *asymptotic constant*, and $f(x) \sim g(x)$ means that

$$lim_{x\to\infty} f(x)/g(x) = 1,$$

r is called the *asymptotic decay rate* and is given by:

$$r = \frac{(1-\rho)(1+\lambda)}{1-C/N},$$

where ρ denotes link utilization, λ is the ratio of the average on period to the average off period of the on-off source, C is the link capacity normalized by the peak rate of the on-off source and N is the number of on-off sources.

Gibbens *et al.* [32] develop the effective bandwidth for exponential on-off sources. Guerin *et al.* independently obtain the formulas in [32] through insightful interpretations of the results in [31] and extend them through heuristics [33]. By approximating the asymptotic constant A by 1, they give effective bandwidth of an exponential on-off source explicitly. The effective bandwidth of a single on-off source in a network with buffer size x is given by:

$$\widehat{c} = \frac{\alpha b(1-p) \times pcr - x + \sqrt{[\alpha b(1-p)pcr - x]^2 + 4x\alpha bp(1-p)pcr}}{2\alpha b(1-p)}.$$

where $\alpha = ln(1/\varepsilon)$, and ε is the overflow probability. The exponential on-off source is characterized by its peak rate pcr, activity parameter p and mean burst length b. The *activity parameter* p is the ratio of mean rate to peak rate. Moreover, by modeling the aggregate traffic process as Gaussian process with mean rate m and standard deviation σ, Guerin *et al.* propose that the effective bandwidth of the aggregate traffic can be determined:

$$\widehat{C} = min\left\{m + \alpha' \times \sigma, \quad \sum_{i=1}^{N} \widehat{c_i}\right\},$$

where N is the number of multiplexed connections, α' is given by

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

and

$$\sigma^2 = \sum_{i=1}^{N} \sigma_i^2.$$

σ_i is the standard deviation of instantaneous traffic rate of the ith connection and it is provided by the traffic source at the connection setup phase.

Elwalid *et al.* [34] improve the results in [31], [32], [33] by investigating the effective bandwidth of general Markovian traffic sources. They conclude that for general Markovian traffic sources, a notional effective bandwidth can be obtained for each source from the dominant real eigenvalue of the infinitesimal generator Matrix of a Markov chain characterizing the traffic source. In their approach also, the asymptotic constant A in Eq. 2.1 is assumed to be 1. In another work, Kelly [35] shows the existence of the effective bandwidth in $M/G/1$ queue.

2.1.2 Large Deviation Approach

In addition to the above research which is based on investigating the eigenvalue problem, large deviations theory is also widely used to determine the effective bandwidth of a traffic source [36], [37], [38], [39], [40], [41], [42], [43]. A large deviation asymptotic analysis provides strong support for the simple effective bandwidth, because it shows that it is asymptotically correct that as the buffer

size gets larger the tail probability gets smaller, and because it provides a basis for assigning effective bandwidth values to different traffic sources (voice, data, video, etc.).

Kesidis *et al.* use large deviation approach to determine the effective bandwidth [37]. They show the existence of effective bandwidth for some traffic source models that are often used for modeling ATM traffic, which include constant bit rate source model, Poisson source model, discrete and continuous Markov source models and Markov Modulated Poisson Process (MMPP) model. In [37], the admission criteria is identical to that of [34]. Chang *et al.* [38] study the stability problems in a queueing network using the so-called Minimum Envelop Rate (MER) approach. By associating the definition of MER with the large deviations theory through the Gärtnet-Ellis theorem [44], they develop a set of rules which provides a method to reconcile two types of stability problems of queueing networks:

- conditions for queueing networks that render bounded queue lengths and bounded delay for traffic sources,
- conditions for queueing networks in which the tail of queue length distribution is bounded exponentially with respect to r, the asymptotic decay rate in Eq. 2.1.

Berger *et al.* [43] consider the case of network nodes that use a priority-service discipline to support multiple classes of service. They present the effective bandwidth for services with priorities.

In [36] Shwartz *et al.* sum up some applications of large deviations theory in performance analysis. To summarize the effective bandwidth approach using large deviations theory, assume that the arrival stream consists of J types of traffic, with N_j sources of type j $(j = 1, \ldots J)$, all having a Markovian statistical nature. Let $A_j(0,t)$ denote the number of arrivals of type j source over the time interval $(0,t)$. Assume that the cumulant log moment generating function of $A_j(0,t)$, given by

$$\wedge_j(\theta) = lim_{t\to\infty}\frac{1}{t}logEe^{\theta A_j(0,t)}, \tag{2.2}$$

exists for all real θ, and that $\wedge_j(\theta)$ is differentiable and convex. By large deviations theory:

$$P(Q > x) \sim e^{-\theta^* x}, \tag{2.3}$$

where θ^* is determined by

$$\theta^* = sup\{\theta | \sum_{j \in J} N_j \wedge_j (\theta)/\theta \leqslant C\}. \tag{2.4}$$

Thus $e_j(\theta^*) = \wedge_j(\theta^*)/\theta^*$ is the effective bandwidth of type j source subject to the condition that the tail distribution of the queue length has the decay rate θ^*.

An attractive feature of the effective bandwidth approach is that the effective bandwidths of several traffic sources are additive. That is, the effective bandwidth of the aggregation of several traffic sources is equal to the sum of the effective bandwidth of each traffic source. The notion of effective bandwidth is very natural and desirable. The idea is to assign each source an effective bandwidth. Then, any subset of sources is admissible if the sum of the required effective bandwidths is less than the total available bandwidth.

The additive property is achieved by assuming that the asymptotic constant A in Eq. 2.1 is equal to one. While such arrangement results in simple CAC scheme, as a penalty it also makes effective bandwidth approach very conservative. The effective bandwidth of a traffic source is determined only by its own characteristics. Other sources in the network have no impact on the effective bandwidth of this traffic source. This is equivalent to having no traffic smoothing from multiplexing. In particular, it is known that when many separate bursty sources are multiplexed, the total is less bursty than individual components, i.e. there is a basis for more statistical multiplexing gain than what can be achieved by effective bandwidth approach. This is theoretically supported by the classical limit theorem stating that superposition of arrival processes, suitably scaled, converges to a Poisson process as the number of arrival processes increases [45]. For related performance studies see Heffes *et al.* [46], Sriram *et al.* [47], and Fendick *et al.* [48]. In contrast, with the effective bandwidth approximation, the burstiness of N superposed independently and identically distributed sources is the same as a single source [49, pp. 76]. This implies that the effective bandwidth approximation

will predict greater congestion for any fixed arrival rate than it should. Therefore effective bandwidth approach cannot effectively investigate the utilization gain due to statistical multiplexing. Approximation in Eq. 2.3 tends to get worse when the number of sources increases, the channel utilization decreases, the buffer size decreases, and the traffic source gets further from the Poisson, either more bursty or less bursty [50]. Under some circumstances, the asymptotic constant A can be as low as 10^{-7} which renders the effective bandwidth approach meaningless for cell loss estimation.

Research has been carried out to investigate the loss penalty incurred by effective bandwidth approach and the method for choosing an appropriate asymptotic constant [39], [40], [41], [50]. Simonian *et al.* show heuristically that the asymptotic constant A in Eq. 2.1 can be approximated by the cell loss in the bufferless system [41].

Botvich *et al.* investigate the economy of scale in large buffers [40] using large deviations theory. They analyze a queue fed with N homogeneous Markov sources. They show that under very general conditions, the tail of the queue length distribution satisfies the following relationship:

$$lim_{N\to\infty} N^{-1} logP\,(Q > x) = -I(\frac{x}{N}),$$

where the shape function I is expressed in terms of the cumulant generating function [36] of the input traffic:

$$I(x) = inf_{t>0}\left[t \times \lambda_t^*(x)\right]. \tag{2.5}$$

In Eq. 2.5, λ_t^* is the Legendre-Fenchel transform of λ_t, which is defined by

$$\lambda_t^*(x) = sup_\theta\left[x\theta - \lambda_t(\theta)\right],$$

where λ_t is the finite time cumulant of a traffic source with arrival process $A(0,t)$:

$$\lambda_t(\theta) = t^{-1} logE\left[e^{(\theta A(0,t) - c\times t)}\right].$$

$A(0,t)$ denotes the number of arrivals in the time interval $(0,t)$ and the traffic source is serviced at rate c.

The difference $I(\frac{x}{N}) - \theta^* x$ determines the economy of scale which is to be obtained in large buffers through statistical multiplexing, where θ^* is the asymptotic decay rate.

In [50], Choudhury *et al.* give numerical examples showing that the asymptotic constant A for a buffer fed with N Markovian traffic sources is asymptotically exponential in N, i.e.

$$A \sim \beta e^{-N\gamma} \quad as\ n \rightarrow \infty,$$

where $\beta > 0$. For sources more bursty than Poisson, $\gamma > 0$, while for sources less bursty than Poisson, $\gamma < 0$. Moreover, $|\gamma|$ tends to be larger when the burstiness gets further from Poisson and the traffic intensity decreases, which is the likely operating condition for ATM networks.

In another research by Elwalid *et al.* [39] the overflow probability for N Markov sources serviced at rate c is approximated by:

$$P(Q > x) \approx Le^{-\theta^* x}.$$

They show that L is the loss in bufferless multiplexing as estimated from Chernoff's theorem [51].

Their research results in better network resources utilization. However as a penalty, the additive property of effective bandwidth approach is lost and the CAC scheme becomes more complex.

2.1.3 Bufferless System Model

In addition to the CAC research which is based on the analysis of queueing systems or fluid flow model, there is another class of CAC research which builds its basis on the bufferless system model [14], [52], [21], [53]. Bufferless system model assumes that there is no buffer in the system at the burst time scale [14]. Therefore, cell loss occurs if and only if the aggregate traffic rate exceeds the link capacity. Since the bufferless system model assumes that there is no buffer in the system, it generates conservative estimate of cell loss ratio. However, bufferless

model significantly simplifies the theoretical analysis and enables us to concentrate on the traffic process itself. Therefore it is widely applied to both traffic descriptor-based CAC and measurement-based CAC [14], [54], [52], [21], [53], [16].

Hui proposes a resource allocation method in [14]. He suggests that congestion should be evaluated at different levels, namely, the packet level, the burst level and the call level (refer to section 1.2.4 for details). He also gives a method for evaluating congestion at the burst level. He obtains an upper bound of the burst level blocking probability, where the burst level blocking probability is defined as the probability that the aggregate instantaneous bit rate of calls exceeds the capacity of the trunk group:

$$P(W(t) > c) \leq e^{-(s^* c - \mu_W(s^*))},$$

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

$$c = \mu_W'(s^*),$$

$\mu_W(s)$ is the log moment generating function of W. $W(t)$ is the offered traffic load at time t:

$$W(t) = \sum_{k=1}^{K} R_k(t),$$

where R_k is the instantaneous burst level rate of kth call and K is the set of calls assigned at the call level to the trunk group G with total bandwidth c. For Poisson traffic, the offered traffic load is modeled by jumps of different amplitudes $a_i > 0$, which arrives with Poisson rate γ_i and lasts for a duration b_i. $\mu_W(s)$ is then given by

$$\mu_W(s) = \sum_i \gamma_i b_i (e_i^{sa_i} - 1).$$

For other non-Poisson traffic, it is assumed that traffic source k has a steady state probability p_i for $R_k(t) = a_i$, then $\mu_k(s)$ is given by:

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

The log moment generating function of W is given by $\mu_W(s) = \Sigma_k \mu_k(s)$.

Murase *et al.* propose a CAC scheme based on a method which is able to estimate cell loss probability for each on-off source. In their approach, a traffic source is modeled as an on-off source with peak cell rate pcr, mean cell rate mcr and activity parameter p, where p is defined as the ratio of mcr to pcr. Assume that there are K types of traffic sources on the link, and that N_j is the number of sources of type j. They show that under bufferless system model, the cell loss ratio of type j on a link with capacity C is given by the following equation:

$$clr_j = \frac{OF_j}{\rho_j},$$

where

$$OF_j = \sum_{n_i \in D}^{n_1=N_1} \cdots \sum^{n_K=N_K} \left[\prod_{i=1}^{K} p_i(n_i) \frac{\left(\left(\left(\sum_{i=1}^{K} n_i \times pcr_i \right) - C \right) pcr_j \times n_j \right)}{\sum_{i=1}^{K} n_i \times pcr_i} \right],$$

$$\rho_j = N_j \times pcr_j,$$

and

$$p_i(n_i) = \begin{pmatrix} N_i \\ n_i \end{pmatrix} p_i^{n_i} (1 - p_i)^{N_i - n_i}.$$

The parameters pcr_i and p_i denote the peak cell rate and the activity parameter of type j connections. D is a set defined by:

$$D = \left\{ (n_1, \ldots, n_K) \mid \left(\Sigma_{i=1}^{K} n_i \times pcr_i - C \right) \geq 0 \right\}.$$

The above clr equation is used in their approach to determine an admissible region. On arrival of a new connection if the call numbers of each type lies within the admissible region, the new connection is admitted, otherwise it is rejected.

Lee *et al.* propose a CAC scheme suitable for real-time calculation based on on-off traffic source model [21]. They model a traffic source by an on-off source with peak cell rate pcr and mean cell rate mcr. A traffic rate unit u is chosen such that all traffic rates are normalized with regard to u. If the peak cell rate

of a connection is not an integer value, it is quantized. For example, if the ith connection has mean cell rate mcr_i and peak cell rate pcr_i, after quantization it becomes a quantized source with mean cell rate mcr_i and peak cell rate $\lceil pcr_i \rceil$, where $\lceil \bullet \rceil$ denotes the minimum integer that is greater than $\bullet$. The probability mass function (pmf) of the quantized traffic source is given by:

$$f_i(x) = \begin{cases} p_i & if\ x = \lceil pcr_i \rceil \\ 1 - p_i & if\ x = 0 \end{cases},$$

where p_i is the activity parameter of the quantized source, which is defined as the ratio of its mean cell rate mcr_i to its peak cell rate $\lceil pcr_i \rceil$. Assuming there are n connections in the network, define a series $I(m)$ as:

$$I(m) = \sum_k (k - m)^+ q(k),$$

where $(\bullet)^+$ denotes $max\ \{\bullet, 0\}$, $m = 0, 1, \ldots$, and

$$q(k) = f_1 * \cdots * f_n(k),$$

where $*$ denotes convolution. When a new connection with mean cell rate mcr_{n+1} and peak cell rate pcr_{n+1} arrives, it is first quantized. The quantized source has mean cell rate mcr_{n+1}, peak cell rate $\lceil pcr_{n+1} \rceil$, activity parameter p_{n+1} and pmf $f_{n+1}(x)$. The cell loss ratio when the new connection is admitted is estimated as:

$$clr = \frac{p_{n+1} I\left(C - \lceil pcr_{n+1} \rceil\right) + (1 - p_{n+1}) I(C)}{I(0) + mcr_{n+1}}.$$

If the estimated cell loss ratio is less than the CLR objective, the connection is admitted, otherwise it is rejected. Here C denotes link capacity, which is assumed to be an integer. If the connection is admitted, $I(m)$ will be updated:

$$\begin{aligned} \widehat{I}(m) &= \sum_k (k - m)^+ q * f_{n+1}(k) \\ &= \begin{cases} (1 - p_{n+1}) I(m) + p_{n+1} \left(\lceil pcr_{n+1} \rceil - m + I(0)\right) & if\ m < \lceil pcr_{n+1} \rceil \\ (1 - p_{n+1}) I(m) + p_{n+1} I\left(m - \lceil pcr_{n+1} \rceil\right) & if\ m \geq \lceil pcr_{n+1} \rceil \end{cases} \end{aligned}$$

If a connection with mean cell rate mcr_i, peak cell rate pcr_i leaves the network, $I(m)$ will be updated according to the equation:

$$\widehat{I}(m) = \frac{I(m) - p_i \widehat{I}(m - \lceil pcr_i \rceil)}{1 - p_i},$$

where p_i is the activity parameter of the corresponding quantized source.

Elwalid *et al.* consider the extremal, periodic, on-off sources regulated by leaky-buckets [53]. In their approach, a traffic source is modeled by a periodic on-off source which is the worst case of traffic source in terms of cell loss [55], [56], [57]. A periodic on-off source regulated by a leaky bucket is also a periodic on-off source. Assuming that the periodic on-off source leaving the leaky bucket has peak cell rate pcr, mean cell rate mcr and burst length B_T, then in lossless multiplexing, its effective bandwidth is given by

$$e_0 = \begin{cases} \frac{pcr}{1+\frac{B/C}{B_T}(pcr-mcr)} & if\ mcr \leq \frac{B_T}{B/C} \\ mcr, & if\ \frac{B_T}{B/C} \leq mcr < pcr \end{cases},$$

where B is the buffer size, C is the link capacity.

Chernoff bound [51] is used to obtain the effective bandwidth of the traffic source when statistical multiplexing is considered. Assume that there are J types of connections, and that each type has K_j traffic sources. Assuming that the instantaneous load u_{ij} of the ith connection of type j is an independent, nonnegative random variable, its moment generating function is then given by:

$$M_j(s) = E\left[exp\left(su_{ij}\right)\right].$$

Chernoff's bound gives [51]

$$log\left(P_{loss}\right) \leq -F_{\vec{K}}(s^*),$$

where

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

and

$$F_{\vec{K}}(s^*) = sup_{s \geq 0} F_{\vec{K}}(s).$$

Here P_{loss} is cell loss probability of the aggregate traffic. For periodic on-off source, it can be shown that

$$F_{\overrightarrow{K}}(s) = sC - \sum_{j=1}^{J} K_j log\left\{1 - p_j + p_j exp\left(s \times pcr_j\right)\right\},$$

and s^* is obtained by solving the equation

$$\sum_{j=1}^{J} \frac{K_j \times p_j \times pcr_j \times exp\left(s \times pcr_j\right)}{1 - p_j + p_j exp\left(s \times pcr_j\right)} = C, \tag{2.6}$$

where p_j and pcr_j are the activity parameter and the peak cell rate of type j connection respectively. The above equations are used to determine an admissible region A in their approach:

$$A = \left\{\overrightarrow{K} \ : \ \sum_{j=1}^{J} K_j e_j \leq C\right\},$$

the parameters e_j which define A are obtained:

$$e_j = \frac{logM_j(s^*)}{s^* + (logL)/C},$$

where L is the cell loss ratio objective and s^* is obtained from Eq. 2.6.

2.2 Measurement-based CAC

Measurement-based CAC (MBAC) uses the *a priori* traffic characterizations only for the incoming connection and uses measurements to characterize existing connections. Under measurement-based CAC scheme, network utilization does not suffer significantly if traffic descriptions are not tight. Because Measurement-based CAC bases connection admission decisions on measurements of traffic rather than on worst-case bounds of traffic, it is able to achieve much higher network utilization than traffic descriptor-based schemes, while still providing acceptable QoS [26], [58]. Of course, traffic measurements are not always good predictors of future behavior of traffic, therefore the measurement-based approach to admission

control can lead to occasional QoS violations that exceed the desired levels. However, such occasional service failures are acceptable for many real-time services without stringent QoS requirements.

There are mainly two objectives to be achieved in designing a measurement-based admission control scheme. The first is to provide a scheme that accurately estimates *a priori* the QoS that will result based on traffic parameters from traffic measurements and/or traffic descriptors. This scheme is able to achieve the highest possible utilization for given QoS requirements. The second is to design a measurement scheme that is implementable and able to obtain accurate estimates of parameters from measurements which are required by the MBAC.

Many measurement-based admission control schemes have been proposed in the literature [26], [23], [16], [22], [42], [59], [60], [15], [61] and they implicitly or explicitly seek to achieve one or both of these goals.

2.2.1 Distribution Measurement Approach

Saito *et al.* propose a MBAC which is based on measuring the marginal cell arrival distribution of the aggregate traffic [54]. The marginal distribution of the instantaneous rate is measured. The admission decision is made by applying a CLR upper bound formula [62], [63]. When a new connection arrives, if the estimated cell loss ratio clr is less that the cell loss ratio objective, the connection is admitted; otherwise it is rejected. The instantaneous rate is defined as the number of cells arriving in a window equivalent to the buffer size. Denote by T the time window size, L the length of a cell, C the link capacity, and B the buffer size. The CLR upper bound formula is similar to the virtual CLR formula in [52], i.e. if the instantaneous traffic rate exceeds the link capacity, the excess traffic is lost. Assume that there are n existing connections in the network. When a new connection with mean cell rate mcr_{n+1} and peak cell rate pcr_{n+1} arrives, the upper-bound cell loss ratio is estimated as:

$$\widehat{clr} = \frac{1}{\left(\widehat{mcr}(t) + \frac{T \times mcr_{n+1}}{L}\right)} \sum_{k=0}^{\infty} \left[k - \frac{C \times T}{L}\right]^{+} \widehat{p}(\bullet; t) * \theta_{n+1}(k),$$

where $\widehat{mcr}(t)$ denotes the estimated mean traffic rate of existing connections at time t, $*$ denotes convolution between two distributions, $[\bullet]^+$ denotes $max\{\bullet, 0\}$, $\widehat{p}(k;t)$ denotes the estimated marginal cell-arrival distribution for the existing connections at time t, and θ_{n+1} denotes the worst-case cell-arrival distribution of the new connection. θ_{n+1} is given by

$$\theta_{n+1}(k) = \begin{cases} \frac{\phi_{n+1}}{\xi_{n+1}}, & k = \xi_{n+1} \\ \frac{1-\phi_{n+1}}{\xi_{n+1}}, & k = 0 \\ 0, & otherwise \end{cases},$$

where

$$\xi_{n+1} = \left\lceil \frac{pcr_{n+1} \times T}{L} \right\rceil,$$

and

$$\phi_{n+1} = \frac{mcr_{n+1} \times T}{L}.$$

If the new connection is admitted, the aggregate cell-arrival distribution is replaced by the convolving distribution, and the normalized mean cell rate of the new connection $\frac{T \times mcr_{n+1}}{L}$ is added to $\widehat{mcr}(t)$.

The renewal period consists of N measurement windows of length T. Denote the measured distribution in the N measurement windows at the tth renewal period by $\{q(k;t), k = 0, 1, \ldots\}$, and the measured mean rate of the aggregate traffic by $m(t)$. If the number of connections on the link does not change during a renewal period, the estimated $\widehat{p}$ and $\widehat{mcr}(t)$ are renewed in the following way:

$$\begin{aligned} \widehat{p}(k;t+1) &= \alpha \times q(k;t) + (1-\alpha)\widehat{p}(k;t) \\ \widehat{mcr}(t+1) &= \alpha \times m(t) + (1-\alpha)\widehat{mcr}(t) \end{aligned}$$

where α $(0 \leq \alpha < 1)$ is a smoothing factor determining the weight of the current measurement.

2.2.2 Leaky Bucket Approach

Tedijanto *et al.* base their CAC scheme on measurements of leaky-bucket regulated individual flows [64]. The mean rate mcr and the leaky-bucket tagging

probability ξ of individual flow are monitored. Dynamic bandwidth allocation is performed. If (mcr, ξ) of a connection is outside a certain region, an adjustment to the reserved bandwidth for the connection is issued to satisfy the changed traffic condition. According to the AMS model [31], [33], [34], effective bandwidth c of a connection in a network with buffer size B under the constraint of loss probability ε is calculated as

$$c = pcr\frac{y - B + \sqrt{(y-B)^2 + 4By\rho}}{2y},$$

where $y = log(1/\varepsilon)b(1-\rho)$, $\rho = mcr/pcr$, pcr denotes the peak rate of a connection, and b denotes the burst size. The burst size is obtained from measurements of the state (mcr, ξ):

$$b = \frac{\eta}{log\frac{pcr(\gamma - mcr) + \xi mcr(pcr - \gamma)}{\xi\gamma(pcr - mcr)}},$$

where γ is the token generation rate of the leaky bucket with bucket size M and

$$\eta = \frac{M(\gamma - mcr)pcr^2}{(pcr - mcr)(pcr - \gamma)\gamma}.$$

The admission decision is made as follows. Assume there are n connections on the link. A new connection will be accepted if its requested bandwidth, denoted by c_{new}, satisfy the condition:

$$c_{new} \leq C - \sum_{i=1}^{n} c_i.$$

Otherwise there are two cases. If the mean rate of the new connection

$$mcr_{new} \leq C - \sum_{i=1}^{n} c_i < c_{new},$$

the connection is accepted if its traffic parameters can be re-shaped to fit into the available bandwidth, otherwise it is not rejected because the average rate cannot be maintained. If the connection is admitted, bucket size M is determined such that the tagging probability of the connection is lower than a certain target value ξ_T:

$$M = \phi log\left[\frac{pcr(\gamma - mcr) + mcr \times \xi_T(pcr - \gamma)}{\xi_T\gamma(pcr - mcr)}\right],$$

where

$$\phi = \frac{b(pcr - mcr)\gamma(pcr - \gamma)}{pcr(\gamma - mcr)}.$$

2.2.3 Asymptotic Matching Approach

Baiocchi *et al.* [65] propose an asymptotic matching method, which is used to investigate the loss performance of homogeneous on-off sources. In their approach, the aggregate traffic of on-off sources is modeled by a two-state Markov Modulated Poisson Process (MMPP). The MMPP consists of two states: the overload state and the underload state. The overload state is defined as a state where the traffic rate of the aggregate traffic exceeds the link capacity. The underload state is defined as the complement of the overload state.

Assume that N on-off sources are multiplexed on the link and that M is the maximum number of active connections to ensure that overload does not occur. Parameters of the two-state MMPP are given as:

$$\begin{aligned}
r_1 &= \eta, \\
r_0 &= r_1 \frac{Np\Lambda - \lambda_0}{\lambda_0 - Np\Lambda}, \\
\lambda_1 &= pcr \times \sum_{i=M+1}^{N} i \frac{\pi_i}{\Sigma_{j=M+1}^{N} \pi_j}, \\
\lambda_0 &= pcr \times \sum_{i=0}^{M} i \frac{\pi_i}{\Sigma_{j=M+1}^{N} \pi_j},
\end{aligned}$$

where η is the maximal real part eigenvalue of an infinitesimal generator of a rate transition Markov Matrix obtained from the transient states of $\Re$. $\Re$ is obtained from the phase process considering only the states $\{M, M+1, \ldots, N\}$. pcr is the peak cell rate of the on-off source, p is the activity parameter of the on-off source. r_1 is the mean transition rate out of the overload state. r_0 is the mean transition rate out of the underload state. λ_1 is the mean arrival rate of the MMPP process in the overload state. λ_0 is the mean arrival rate of the MMPP process in the underload state. λ_1 and λ_0 are measured in the same unit as pcr. The asymptotic matching method gives physical meaning to the fitting process. The

approach estimates the CLR of a statistical multiplexer loaded with homogeneous on-off sources [65]. Kang *et al.* extend the application of the asymptotic matching method to heterogeneous on-off sources [66].

Based on the work of Baiocchi *et al.* Shiomoto and Chaki propose a real-time MMPP parameter estimation method referred to as measurement-based asymptotic matching [29]. Measurement-based asymptotic matching employs a fictitious queue and a window to detect the overload state. The overload state is detected when the fictitious queue has not been empty for a certain period. If no cell arrives and the fictitious queue is empty, underload state is detected. On arrival of cells, at the beginning of a time slot, the fictitious queue is incremented by the number of cells. Inside the time slot, the state does not change, but it becomes susceptible to overloading.

The arrival rates and mean duration time for each state are measured. When a connection request arrives, the measured arrival rates and duration time are modified and the CLR is calculated from the $MMPP/D/1/K$ queueing model. The estimated CLR is compared with the target value: if it is lower than the target value, the connection is admitted; otherwise, it is rejected. In contrast to the method in [65] which needs a set of traffic descriptors for all the connections to calculate the parameters by asymptotic matching, the measurement-based asymptotic matching method in [29] requires measurement of actual traffic.

2.2.4 Decision-Theoretical Approach

Gibbens *et al.* consider call level dynamics in their MBAC which uses a decision-theoretical approach [16]. The instantaneous rate measured at the decision epoch is used to make admission decisions. In their scheme, when a new connection arrives, the measured instantaneous rate is compared with a threshold. If it is smaller than the threshold, the connection is admitted; otherwise, it is rejected. The originality of their approach lies in how they decide the threshold value for the instantaneous rate. The threshold value, determined using the decision-theoretical approach, is a function of the number of existing connections. The connection acceptance boundary, denoted by s, is the set of the traffic rates, which depends

on the number of connections n: $s = (s(n), n = 0, 1, \ldots)$. That is, the new connection is admitted if the measured instantaneous rate is lower than $s(n)$ and the number of existing connections is n. Creating the connection acceptance boundary is the concern. Gibbens *et al.* approach this as a decision-theoretical problem. They take into account not only the burst-level characteristics (activity parameter) but also the call-level characteristics (connection arrival rate). Assuming a prior function $f(p, \lambda)$, where p denotes the activity parameter and λ denotes the connection arrival rate normalized by the departure rate, they pre-compute the connection acceptance boundary to maximize the expected reward per unit time. If each offered cell attracts a reward of one unit while each lost cell incurs a penalty of y units, the expected reward per unit time is defined as:

$$\int\int \left[\sum_{n=1}^{\infty} \pi(n; p, \lambda)\,(np - yM(n; p))\right] f(p, \lambda) dp d\lambda,$$

where $\pi(n; p, \lambda)$ denotes the stationary probability that the number of connections is n, and $M(n; p)$ denotes the cell loss rate, given that the activity parameter is p and the connection arrival rate is λ (λ (π $(n; p, \lambda)$) is determined by pre-computed thresholds for the instantaneous rate). Because the penalty of y units measures the marginal CLR, it is set large to maintain a low CLR. They demonstrate that their approach controls the CLR around the target value except for regions of low burstiness (high p value), assuming that the uniform prior function $f(p, \lambda)$ is given.

2.2.5 Aggregate Effective Bandwidth Approach

The method proposed by Dziong *et al.* uses a Kalman filter to estimate the mean and variance of the instantaneous rate [42]. Their approach is based on the aggregate effective bandwidth. If the probability of the aggregate effective bandwidth exceeding the link capacity is lower than a threshold, a new connection request is admitted, otherwise, it is rejected.

The aggregate effective bandwidth G is approximated by a linear function of

both the mean M and the variance V of aggregate instantaneous rates:

$$G = \gamma M + \theta V.$$

Coefficient γ is assumed to be independent from the current state and can be calculated off-line using the link speed, QoS constraint, and average traffic mixture. Then the coefficient θ, for a given state, is given by

$$\theta = \frac{G - \gamma M}{V}.$$

An estimate of the aggregate effective bandwidth $\widehat{G}$ is obtained by using the estimated mean $\widehat{M}$ and variance $\widehat{V}$ of aggregate instantaneous rates:

$$\widehat{G} = \gamma \widehat{M} + \theta \widehat{V}.$$

The measurement process is formulated as a state-estimation problem. The state is defined by a vector of the mean and the variance of the aggregate instantaneous rates, i.e. the state of the system is defined as $X_k = [M_k, V_k]^T$. The state estimation employs a two-state Kalman filter. The Kalman filter estimates the state at each connection arrival or departure epoch.

The state estimation is performed according to the first-order recursive equation:

$$\widehat{X}_k = \widehat{X}_k^e + K_k \left[Z_k - \widehat{X}_k^e \right],$$

where K_k denotes the Kalman gain factor, and

$$\widehat{X}_k^e = \widehat{X}_{k-1} + x_k$$

denotes the state estimation extrapolation, and x_k denotes either the declared mean and variance of the accepted connection or the normalized declared mean and variance of the released connection. The Kalman gain factor is determined such that the conditional probability that the measured state Z_k occurs becomes the highest, given the previous estimate and the Gaussian measurement and estimation errors.

Given that the measured

$$Z_k = \left[\overline{M_k}, \overline{V_k}\right]^T$$

and measurement error covariance

$$Y_k = \left[\overline{v}^m_{e,k}, \overline{v}^v_{e,k}\right],$$

where $\overline{v}^m_{e,k}$ and $\overline{v}^v_{e,k}$ denote the variance of the estimation errors, the Kalman gain is defined by

$$K_k = P^e_k \left[P^e_k + Y_k\right]^{-1},$$

where P^e_k denotes the estimate error covariance matrix extrapolation given by

$$P^e_k = F_{k-1} P_{k-1} F^T_{k-1} + Q_k.$$

F_{k-1} is a transition matrix obtained from Eq. 2.7:

$$X_k = F_k X_{k-1} + e_k, \tag{2.7}$$

where e_k is the model error. P_{k-1} is the updated estimate error covariance matrix obtained from

$$P_{k-1} = \left[I - K_{k-1}\right] P^e_{k-1},$$

and Q_k is the estimation error covariance matrix, which is given by traffic sources at connection setup time. In their approach, the model and measurement errors are assumed to be Gaussian, therefore estimation error distributions are also Gaussian with zero mean and variance defined by diagonal elements of the error covariance matrix.

The instantaneous rate is sampled regularly, and the mean, the variance and the error variance are measured. The measurement error covariance matrix

$$Y_k = \left[\overline{v}^m_{e,k}, \overline{v}^v_{e,k}\right]$$

is given by

$$\begin{aligned}\overline{v}^m_{e,k} &= \frac{\overline{V}_k}{N_k} \\ \overline{v}^v_{e,k} &= \frac{\overline{S}_k - \overline{V}_k^2}{N_k},\end{aligned}$$

where

$$\begin{aligned}\overline{V}_k &= \frac{\Sigma_i (d_i - \overline{M}_k)^2}{N_k - 1} \\ \overline{S}_k &= \frac{\Sigma_i (d_i - \overline{M}_k)^4}{N_k - 1}.\end{aligned}$$

d_i denotes the ith sample of the instantaneous rate, and N_k denotes the number of sample.

They reserve a bandwidth R for the estimation error for the aggregate effective bandwidth of the existing connections:

$$R = U(\varepsilon_1)\sqrt{\gamma^2 \widehat{v}^m_e + \theta^2 \widehat{v}^v_e},$$

where $\widehat{v}^m_e$ and $\widehat{v}^v_e$ denote diagonal elements of the error covariance matrix P, $U(\varepsilon_1)$ denotes a coefficient derived from a normalized Gaussian distribution, which ensures that

$$P\left\{G > \widehat{G} + R\right\} \leq \varepsilon_1.$$

For a new connection, the peak rate of the new connection is used for the sake of simplicity and safety. Thus, a new connection is admitted if

$$g'_k + \widehat{G} + R < L,$$

where g'_k and L denotes the peak rate of the new connection and the link capacity; otherwise it is rejected. In this method the mean and variance of the instantaneous rate are monitored. Traffic sources are required to declare the mean and variance of instantaneous rates, and their declaration errors.

In another MBAC proposed by Ren *et al.* also based on the aggregate effective bandwidth, they adopt the fuzzy logic control. For a bufferless system, they model

the aggregate traffic process $R(t)$ as a stationary Gaussian process with mean $\overline{M}$ and variance σ^2. The aggregate bandwidth required to support all existing connections with the specified cell loss ratio ε_{tar} is approximately given by

$$C = \overline{M} + \xi\sigma, \tag{2.8}$$

where

$$\xi \approx 1.8 - 0.46 log_{10}(\eta),$$

and

$$\eta = \frac{\overline{M}\sqrt{2\pi}}{\sigma}\varepsilon_{tar}.$$

For a buffered system with buffer size B, the aggregate traffic process is modeled as a diffusion process referred to as the Ornstein-Uhlenbeck process

$$\frac{dR(t)}{dt} = -\beta(R_t - \overline{M}) + \alpha\frac{dW_t}{dt}$$

where W_t denotes the standard Wiener process, β denotes the drift coefficient and α denotes the diffusion coefficient. Then the aggregate bandwidth required to support all existing connections can be obtained:

$$C = \overline{M} + \gamma\sigma. \tag{2.9}$$

The coefficient γ is given implicitly in the equation

$$\frac{e^{-(\gamma^2/2)}}{\gamma\sqrt{2\pi}}e^{-(\beta/\sigma)\gamma B} < \varepsilon_{tar}.$$

The parameters $\overline{M}$ and σ are obtained from traffic measurements.

The effective bandwidth in this dynamic CAC scheme takes the form

$$C_d = s \times C_u + (1 - s) \times C_m,$$

where C_u is the effective bandwidth calculated from either Eq. 2.9 or Eq. 2.8, in which the mean and the variance of the aggregate traffic is determined from user-declared parameters. The traffic source is modeled as an independent on-off

source. C_m is the effective bandwidth calculated using the measured aggregate traffic parameters, and s is a smoothing factor.

To determine s, they use the fuzzy logic control. They obtain a real-time estimate of the CLR actually achieved, denoted by ε_{mes}, using the in-service QoS monitoring technique in [67]. The in-service QoS monitoring technique is based on the asymptotic relationship between CLR and buffer size, and uses virtual buffers. ε_{mes}, together with the measured mean and variance of the aggregate traffic, are used to obtain an estimate of the effective bandwidth, denoted by C_a. The ratio of C_a to c, the link capacity, becomes the input to the fuzzy logic control, and the output is s.

2.2.6 Instantaneous Rate Approach

Li *et al.* investigate the performance of the capacity allocation method in the frequency domain and find that low-frequency traffic rate components remain intact after the traffic passes through the buffer [68]. They suggest that buffering cannot support low-frequency components; thus, support must be furnished by assigning adequate link capacity. The cutoff frequency, which defines the low frequency region, is derived through analysis of an actual VBR video signal for different buffer sizes. They also suggest that routing control and admission control can be formulated as a linear system with only low-frequency components by overlooking the queueing process related to high-frequency components. Their research supports the MBAC proposed by Shiomoto *et al.*

The method proposed by Shiomoto *et al.* employs low-pass-filtered traffic rate measurement [60]. The instantaneous rate is measured using a low-pass filter, and admission decision is made using the effective bandwidth derived from the measured instantaneous rate. The effective bandwidth is set to the maximum instantaneous rate observed in a monitoring period. The instantaneous link utilization can be estimated by applying a recursive low-pass filter to the observed number of cells arriving during a time slot Δ:

$$\lambda(t) = \alpha n(t) + (1-\alpha)\lambda(t-\Delta), \qquad 0 \leq \alpha \leq 1, \tag{2.10}$$

where $n(t)$ denotes the number of cells arriving during the tth time slot, and Δ denotes the single-cell transmission time over the link. The residual bandwidth is obtained from the difference between the link capacity and the maximum instantaneous rate observed over a sufficiently long period. For a new connection whose peak rate is R, arriving at a link whose capacity is C, the following admission criteria is used:

$$\frac{R}{C} < 1 - max_{t' \in (t-T_m,\, t)} \lambda(t'),$$

where T_m denotes the monitoring period.

The filter coefficient α is chosen so that the instantaneous link utilization can be estimated accurately. The power spectral function of Eq. 2.10 is given by Eq. 2.11:

$$S(\omega) = \frac{\alpha^2}{1 + (1-\alpha)^2 - 2(1-\alpha)cos(\omega\Delta)}. \tag{2.11}$$

The burst level behavior of traffic rate is captured by eliminating all frequency components higher than the peak cell rate. To eliminate those frequency components higher than the connection's peak rate, α is determined so that the corresponding $S(\omega_0)$ is lower than a small positive number ε, where ω_0 denotes the frequency (in radians) corresponding to the peak cell rate. Then the smoothing coefficient is obtained using

$$\alpha = \frac{-2(1-K) + \sqrt{4(1-K)^2 + 8(\varepsilon^{-1}-1)(1-K)}}{2(\varepsilon^{-1}-1)}$$

where $K = cos(\omega_0\Delta)$.

The monitoring period T_m is also an important parameter in their MBAC. They derive T_m by approximating the 99% cumulative value of the duration of an underload state. An underload state is defined as a state that the instantaneous link utilization does not exceed target load λ_{target}. As an example they present the method of determining T_m for homogeneous exponential on-off sources.

2.2.7 Virtual Buffer Approach

Courcoubetis *et al.* propose an admission control method using virtual buffer techniques [69]. They show that for Markovian fluid traffic model, the cell loss probability $\phi(N, B, c)$, which is expressed as a function of the number of traffic sources N, buffer size B and link capacity c has the following property for large B:

$$\phi(N(1+\varepsilon), B, c) \approx \phi(N, B, c/(1+\varepsilon)). \tag{2.12}$$

According to Eq. 2.17, when a fraction of ε more calls are added to the network, the cell loss probability is equal to the cell loss probability where there are N traffic sources in the network and the link capacity is $c/(1+\varepsilon)$. To estimate $\phi(N, B, c/(1+\varepsilon))$, virtual buffers are employed. They show that for a virtual buffer with buffer size $\frac{B}{k}$, the following relationship exists:

$$\phi\left(N, \frac{B}{k}, \frac{c}{1+\varepsilon}\right) = A\left(\frac{B}{k}\right)^{-\xi} exp\left(-\frac{B}{k} I\left(N, \frac{c}{1+\varepsilon}\right)\right), \tag{2.13}$$

where A and ξ are unknown numbers, I is an unknown number determined by N and $\frac{c}{1+\varepsilon}$. To determine the three unknown numbers A, ξ and I, three virtual buffers with the same input traffic as the real buffer and an output rate of $\frac{c}{1+\varepsilon}$ are employed. When the three unknown numbers are determined, Eq. 2.13 is used to obtain an estimate of $\phi(N, B, c/(1+\varepsilon))$ by making $k = 1$. If $\phi\left(N, B, \frac{c}{1+\varepsilon}\right)$ is smaller than the cell loss probability objective, then according to Eq. 2.12, an additional $N\varepsilon$ number of calls can be admitted into the network without violating the QoS objective. In this approach they assume that all connections on the link are of the same type.

Zhu *et al.* propose an in-service QoS monitoring and estimation method (ISME) which is used to obtain a real-time estimate of the CLR [67]. It is shown in the literature that for general Markovian traffic, the cell loss ratio clr has an asymptotic relationship with the buffer size B [31], [39] when $B \to \infty$:

$$log(clr) \sim -\alpha - \beta \times B \tag{2.14}$$

where α, β are positive constants determined by the traffic process. In contrast to the Markovian model, buffer size and $log(clr)$ for the fractional Brownian motion model, which is long-range dependent, results in the following generalized

relationship:

$$log(clr) \sim -\delta B^{\gamma}, \tag{2.15}$$

where δ and γ are positive constants determined by the traffic process. Eq. 2.15 can also be written in the following linear form:

$$log\left(-log\left(clr\right)\right) \sim log\left(\delta\right) + \gamma log(B). \tag{2.16}$$

Based on the asymptotic relationship between buffer size and CLR, Zhu *et al.* propose to use CLR values observed at three virtual buffers, with much reduced buffer sizes and the same input and output traffic as the real buffer, to estimate the CLR of the real buffer. When n observed samples arrive, linear least square regressions are performed separately based on both Eq. 2.14 and 2.16. Both regression results are stored. R^2 tests are carried out to choose the appropriate equations for regression. They keep two running counters C_M and C_L. If the R^2 result associated with the Markovian model is greater than that associated with the long-range dependent model, then C_M is incremented by one, otherwise C_L is incremented by one. The above procedure is performed until N monitoring periods have finished. If C_M is greater than C_L, the Markovian regression model is chosen, otherwise the long-range dependent regression model is chosen. They then extrapolate to get the N sample CLR estimates at the real buffer size based on the chosen regression model. Moreover they model the CLR estimation process as a Gaussian process with mean $\overline{L}$ and variance σ^2. $\overline{L}$ is the mean of the estimated clr and σ^2 is the variance. This Gaussian model is used for QoS violation detection. Since the virtual buffers have much reduced buffer sizes, therefore much higher cell loss ratio, the CLR values at virtual buffers require much less monitoring period. Therefore virtual buffer technique effectively reduces the monitoring period required.

The in-service QoS monitoring method of Zhu *et al.* and spare capacity estimation method of Courcoubetis *et al.* are combined and used for spare link capacity estimation by Li [70]. He shows that for Markovian traffic, the cell loss probability $\phi\left(N, B, c\right)$, which is expressed as a function of the number of traffic

sources N, buffer size B and link capacity c has the following scaling property:

$$\phi\left(N(1+\varepsilon), B, c\right) = \phi\left(N, B, c/(1+\varepsilon)\right) \tag{2.17}$$

where ε is a small positive number. For traffic which can be modeled by fractional Brownian process, the relationship becomes

$$\phi\left(N(1+\varepsilon), B, c\right) = e^{N log\left(\phi\left(N, \frac{B}{1+\varepsilon}, \frac{c}{1+\varepsilon}\right)\right)}. \tag{2.18}$$

The fundamental concept of his scheme is to employ two sets of several virtual buffers: one for in-service QoS monitoring and the other for spare capacity estimation. The buffer sizes in both sets of virtual buffers are much smaller than the real buffer size. The in-service QoS monitoring set has the same link capacity as the physical link, while the spare capacity estimating set has $1/(1+\varepsilon)$ times the link capacity of the physical link. Using the cell loss probability (CLP) information at the in-service QoS monitoring virtual buffer set, an estimate of CLP at the real buffer is obtained. At the same time an estimate of CLP for the spare capacity calculation can be obtained by the way of the CLP information at virtual buffer set for the spare capacity estimation. If the CLP from the in-service monitoring part indicates no service violation, then the spare capacity estimation part is employed to estimate the spare capacity available. The in-service monitoring part also determines whether the Markovian model or the fractional Brownian model shall be employed. The spare capacity estimation part then uses the chosen model to determine whether or not an additional $N\varepsilon$ number of calls can be admitted.

In another CAC based on virtual buffer techniques, Bensaou *et al.* employ Fuzzy system control [71]. They employ virtual buffers with the same buffer size as the real buffer, and different reduced link capacities. The observed virtual cell loss ratios are fed into a fuzzy logic estimator to obtain an estimate of the spare link capacity. They also present the application of fuzzy system control to in-service QoS monitoring. For in-service QoS monitoring, virtual buffers with different reduced buffer sizes, and the same link capacity as that of the real buffer are used. The virtual cell loss ratios from these in-service QoS monitoring virtual buffers are fed into a fuzzy logic controller which uses Gaussian member functions to obtain an estimate of the cell loss ratio at the real buffer.

2.2.8 Comparison

The proposed measurement-based CAC schemes, although embracing similar goals, differ in four important ways [58]:

- First, some CAC schemes are *principled*, in that they are based on solid mathematical foundations such as large deviations theory, queueing theory, etc., and others are *ad hoc*, in that they lack a theoretical underpinning.
- Second, the specific equations used in making admission decisions are quite different. Some are based on the QoS estimation while others are based on the estimation of the spare bandwidth or admissible region.
- Third, some MBACs have a control parameter that varies the level of the achieved QoS and utilization (by making the scheme more or less aggressive) while others do not have this parameter. This control parameter brings some flexibility as well as complexity into the CAC scheme. This parameter is calibrated in some schemes and serves as an accurate estimate of the resulting performance, while others leave the parameter uncalibrated and assume that the network operator will learn appropriate parameter settings over time.
- At last, the measurement processes used to produce an estimate of the network load are very different. They range from a simple point sample estimate, to estimates based on both the mean and variance of the measured load, to traffic distribution or queue length distribution estimates. Thus the space of measurement-based admission control algorithm is both heavily and broadly populated.

Despite a lot of MBAC schemes proposed, little work has been carried on their performance comparison except for [58]. Lee *et al.* do some illuminative work on performance comparison of MBACs in [58]. They compare the performance of some MBACs and insightfully summarize some common characteristics of these MBACs. The MBACs included in their performance comparison have a control

parameter that varies the level of the achieved QoS and utilization. By varying this parameter they obtain a loss/load curve. By employing this loss/load curve for performance comparison, they arrive at a plausible conclusion that all MBACs have the same performance. However the use of loss/load curve for performance comparison is not justified in their work. Considering that the loss/load curve may be decided by the characteristics of traffic sources used in their simulation, NOT by MBACs, and that in real applications the actual process is reverse (given QoS then decide the control parameter), their method of performance comparison is plausible. In summary, there is no benchmark for performance comparison of MBACs, more work needs to be done in this regard.

2.3 Self-Similarity in Network Traffic

Since the discovery of the existence of self-similarity in network traffic, it has attracted a lot of attentions all over the world [39], [59], [15], [72], [73], [74], [75], [76], [77], [78], [79], [80], [81], [82], [83], [84], [85], [86], [87]. Intuitively, the presence of long-range dependence in a time series indicates that while long-term correlations are individually small, their cumulative effect is not negligible and produces scenarios which are drastically different from those experienced with traditional short-range dependent models such as Markovian processes. The discovery of self-similarity in network traffic challenges the basis of traditional CAC schemes which rely on Markovian traffic source model for performance analysis.

2.3.1 Existence of Self-Similarity - Evidence and Possible Causes

Leland *et al.* present their analysis on Ethernet LAN traffic in [72]. Specifically, they investigate the statistical characteristics of network traffic data collected between August 1989 and February 1992 on several Ethernet LAN's at the Bellcore Mirristown Research and Engineering Center. Based on the experimental analysis on the measured traffic data, they conclude that:

- Ethernet LAN traffic is statistically self-similar.

- The degree of self-similarity measured in terms of the Hurst parameter H is typically a function of the overall utilization of the Ethernet and can be used for measuring the "burstiness" of the traffic (namely, the burstier the traffic the higher the H).
- Major components of Ethernet LAN traffic such as external LAN traffic or external TCP traffic share the same self-similar characteristics as the overall LAN traffic.
- The traffic models currently considered in the literature are not able to capture the self-similarity property. This can be clearly distinguished from their measured data.

They consider two mathematical models for modeling self-similar traffic: fractional Gaussian process and fractional autoregressive integrated moving-average processes. These two models are able to yield elegant representation of the self-similarity phenomenon, but do not provide any physical explanation of self-similarity. However, while these models are able to capture the second order characteristics of network traffic, i.e. self-similarity, the first order characteristics, i.e. probability density distribution, which is also very important for performance analysis, is neglected [15].

Willinger *et al.* attempt to provide a physical explanation for the occurrence of self-similarity in local area network traffic [77]. Their mathematical analysis indicates that the superposition of many on-off sources with strictly alternating on and off periods and whose on periods or off periods exhibits the *Noah effect* (i.e. have high variability or infinite variance) produces aggregate network traffic that exhibits the *Joseph effect* (i.e. is self-similar or long-range dependent) [85]. This result reduces the self-similarity phenomenon for the aggregate LAN traffic to properties of the individual traffic components that make up the aggregate stream. Self-similarity in the network traffic can therefore be explained by traffic rate behavior at individual source level. They also present extensive statistical analysis of Ethernet LAN traffic traces with high time resolution, which confirms that the data at the level of individual sources or source-destination pairs

are consistent with the Noah effect. The proposed physical explanation based on the Noah effect suggests the essential difference between self-similar and traditional traffic modeling in the parameter settings of the well known on-off source models. Traditional traffic modeling assumes finite variance distributions for the on and off periods, while self-similar modeling is based on the assumption of the Noah effect, i.e. it requires infinite variance distribution.

In another research Beran *et al.* analyze 20 large sets of actual VBR video data, generated by a variety of different codecs and representing a wide range of different scenes [73]. Three methods are used to test the existence of self-similarity in VBR video traffic: variance-time analysis, R/S-analysis and Periodogram based analysis. In addition to these three methods, further convincing evidence is gained by performing such simple procedures as re-shuffling and differencing of the original data. Intuitively, if the original time series exhibits long-range dependence, then any re-shuffling should break up the long-range dependencies, and any re-shuffled version of the original time series should resemble white noise. The motivation for differencing a time series with an apparent degree of long-range dependence close to 1 is to ensure that certain types of non-stationarities of the original time series can be safely ruled out as possible explanations for the observed long-range dependence [88], [89]. They conclude that long-range dependence is an inherent feature of VBR video traffic, i.e. a feature that is independent of scene (e.g. video phone, video conference, motion picture video) and codec. They advocate modeling VBR video traffic using self-similar traffic source model.

Paxson *et al.* investigate a number of wide area TCP arrival processes [74]. They study FTP and TELNET session and connection arrivals, FTP data connection arrivals within FTP sessions and TELNET packet arrivals. They show that TELNET connection arrivals and FTP connection arrivals are well-modeled as Poisson process with fixed hourly rates. However, the packet arrivals within a connection can not be captured by the commonly used Poisson packet arrival model. The commonly used exponentially distributed packet arrival model seriously underestimates the burstiness of both TELNET and FTP connections. They conclude that wide area traffic is much burstier than that predicted by Poisson models over

many time scales. Poisson-based modeling of wide area traffic should be abandoned for all but user session arrivals and self-similar traffic model should be used for wide area traffic modeling.

Crovella *et al.* use World Wide Web as an object of study [79]. They present evidence that traffic due to World Wide Web transfers shows characteristics that are consistent with self-similarity. More importantly, they trace the genesis of Web traffic self-similarity along two threads: first, they show that transmission time may be heavily tailed, primarily due to the distribution of available file sizes on the Web. Second, they show that silent time also may be heavily tailed, primarily due to the influence of user "think time". Using the results of Willinger *et al.* [77], the heavily tailed distributions of the transmission and silent time, which corresponds to the on and off periods of an on-off source, result in self-similarity in the aggregate network traffic. Moreover, they attribute the self-similarity in World Wide Web traffic primarily to the distribution of file sizes in the Web (which determine the on time) because they find that the on time distribution is heavier than off time distribution. Their results suggest that the self-similarity of Web traffic is not a machine-induced artifact, it is probably due to the basic characteristics of information organization and retrieval. Therefore changes in protocol processing and document display are not likely to fundamentally remove the self-similarity of the Web traffic.

In summary, the reviewed research proves convincingly that self-similarity phenomenon exists widely in network traffic. The possible causes of self-similarity in the aggregate network traffic are the properties of individual traffic sources that make up the network traffic.

2.3.2 Definition and Inference of Self-Similarity

Leland *et al.* define the self-similarity [72] and Tsybakov *et al.* discuss the definitions of self-similarity in [72] and give simpler criteria for it [75].

Let $X = \{X_t\} = (X_1, X_2, \dots)$ be a semi-infinite segment of a covariance-stationary stochastic process of discrete argument t (time), where $t \in I_1 = \{1, 2, \dots\}$. Denote $\mu = EX_t$ and $x_t = X_t - \mu$. The process X and $x = \{x_t\}$

have the autocorrelation function and the variance denoted by

$$r(k) = \frac{E\left[x_t x_{t+k}\right]}{Ex_t^2},$$

and

$$var\ (X) = var\ (x) = \sigma^2 = Ex_t^2.$$

Both $r(k)$ and $\sigma^2 < \infty$, as well as $\mu < \infty$, are independent of t because of the stationary property of X_t and x_t.

Assume X has an autocorrelation function of the form

$$r(k) \sim k^{-\beta} L(k), \qquad k \to \infty, \tag{2.19}$$

where $0 < \beta < 1$ and L is a slowly varying function as $k \to \infty$, i.e.

$$\lim_{t\to\infty} L(tx)/L(t) = 1 \quad for\ all\ x > 0.$$

For each $m \in I_1$, let

$$X_t^{(m)} = \frac{1}{m}\left(X_{tm-m+1} + \cdots + X_{tm}\right), \ \ t \geq 1.$$

The process $X^{(m)} = \left\{X_t^{(m)}\right\} = \left\{X_1^{(m)}, X_2^{(m)}, \ldots\right\}$ is the semi-infinite segment of a second-order stationary process with discrete argument $t \in I_1$. Let $r^{(m)}(k)$ denote the autocorrelation function of process $X^{(m)}$, the process X is called exactly second-order self-similar with self-similarity parameter (Hurst parameter) $H = 1 - \beta/2$ if its autocorrelation function $r(k)$ satisfies:

$$\begin{aligned} r(k) &= \frac{1}{2}\left[(k+1)^{2-\beta} - 2k^{2-\beta} + (k-1)^{2-\beta}\right] \\ &\triangleq g(k). \end{aligned}$$

Process X is called asymptotically second-order self-similar with self-similarity parameter (Hurst parameter) $H = 1 - \beta/2$ if for all large enough values of k:

$$r^{(m)}(k) \to r(k), \quad as\ m \to \infty,$$

or for all $k \in I_1$,

$$\lim_{m\to\infty} r^{(m)}(k) = g(k).$$

Mathematically, self-similarity manifests itself in a number of ways:

- The variance of the sample mean decreases more slowly than the reciprocal of the sample size (slowly decaying variances), i.e. $var\left(X^{(m)}\right) \sim a_1 m^{-\beta}$, as $m \to \infty$, with $0 < \beta < 1$ and a_1 is a finite positive constant.

- The autocorrelations decay hyperbolically rather than exponentially fast, implying a non-summable autocorrelation function $\Sigma_k r(k) = \infty$ (long-range dependence), i.e. $r(k)$ satisfies the relation in Eq. 2.19.

- The spectral density f obeys a power-law near the origin ($1/f$ - noise), i.e. $f(\lambda) \sim a_2 \lambda^{-\gamma}$, as $\lambda \to \infty$, with $0 < \gamma < 1$ and $\gamma = 1 - \beta$.

Accordingly there are mainly three methods to test the existence of self-similarity. These methods are used in [72], [79]. A summary of the relative accuracy of these methods on synthetic datasets is presented in [90].

- Variance-time plot: The variance-time plot relies on the slowly decaying variance of a self-similar series. The variance of $X^{(m)}$ is plotted against m on a log-log plot. A straight line with slope $-\beta$ greater than -1 is indicative of self-similarity, and the parameter H is given by $H = 1 - \beta/2$.

- R/S plot: R/S plot uses the fact that for a self-similar dataset, the rescaled range or R/S statistic grows according to a power law with exponent H as a function of the number of points included, denoted by n. Therefore, the plot of R/S against n on a log-log plot has a slope which is an estimate of H.

- Periodogram method: The Periodogram method uses the slope of the power spectrum of the series as frequency approaches zero. On a log-log plot, the Periodogram slope is a straight line with slope $\beta - 1 = 1 - 2H$ close to the origin.

While the preceding three graphical methods are widely used for self-similarity test, they do not provide confidence intervals and may be biased for large H [90]. Another method, called the *Whittle estimator*, does provide a confidence interval, but has the drawback that the form of the underlying stochastic process must be

supplied. The two forms that are most commonly used are fractional Gaussian process with Hurst parameter $1/2 < H < 1$ and fractional autoregressive integrated moving-average process (ARIMA) [91], [92]. These two models differ in their assumptions about the short-range dependences in the datasets. Fractional Gaussian process assumes no short-range dependence, while fractional ARIMA can assume a fixed degree of short-range dependence. Another method in the literature is the stochastic re-shuffling method [93], [73]. The stochastic re-shuffling method divides the sampled time series into small segments with a length of time T. These small segments are then re-shuffled randomly. Random re-shuffling makes the sampled points that are more than T away independent, while the autocorrelations that are less than T are preserved. The stochastic re-shuffling method can only be used to detect the existence of self-similarity, but it can not generate an estimate of self-similarity parameter.

2.3.3 Relevance of Self-Similarity - *The Pros and Cons*

Norros studies the tail behavior of a fluid queue fed with a fraction Brownian traffic process with self-similarity parameter $H \in \left[\frac{1}{2}, 1\right)$ [81]. He obtains a lower bound for the tail distribution of the queue length. He shows that the tail behavior of the fluid queue in the fractional Brownian model is in the best case Weibullian:

$$P\left(Q > x\right) \sim exp\left(-\gamma x^{\beta}\right),$$

where γ and β are constants decided by the traffic process. Krishnan shows that the implication of Norro's result is that when a sufficiently large number of sources are multiplexed, high-H sources require less bandwidth than low-H sources [94]. He calls this the *crossover effect*. Garrett *et al.* [95] develop a fractional $ARIMA(0, d, 0)$ model for a monochrome intra-frame coding of the movie *Star Wars*. They show that neither the Hurst parameter nor the marginal distribution, by themselves, can be relied upon to produce a good model. Erramilli *et al.* perform a number of simulation experiments with actual traces of Ethernet LAN traffic and certain randomly "shuffled" versions of the actual traces [78]. They

demonstrate empirically that, beyond its statistical significance in traffic measurements, long-range dependence has considerable impact on queueing performance, and is a dominant characteristic for a number of traffic engineering problems, such as dimensioning of buffers and determining usable capacity.

However, there is considerable debate about the impact of long-range dependence on bandwidth allocation and network performance, and whether or not self-similar model should be used for traffic modeling [39], [76], [15], [96], [97]. They are not challenging the existence of self-similarity in network traffic, however, they take a more practical view at the problems. They point out that the long-range dependence is not a crucial property in determining the behavior of real buffers with *finite buffer size*. Since the objective of any traffic modeling is for performance analysis, and Markovian traffic model is accurate enough to predict the performance of real buffers fed with real traffic sources, Markovian traffic model should be used instead of self-similar traffic model.

Elwalid *et al.* find that buffer overflow probability decreases exponentially with buffer size [39], i.e.

$$log(clr) \approx A + r \times B, \tag{2.20}$$

where A and r are constants and B denotes the buffer size. They test their hypothesis using simulation, where real video teleconference sequence coded by different algorithms are used as traffic sources. Their results show that Eq. 2.20 accurately captures the relationship between buffer overflow probability and buffer size, and $DAR(1)$ (discrete autoregressive process with order 1) traffic source model, which takes into account only short-range dependence, is accurate enough for admission control and bandwidth allocation of video teleconferences. In particular, the video conference sequences used for their simulation exhibit long-range dependence [73]. Therefore, their results effectively counter the assertion in [72] that when traffic is long-range dependent "overall packet loss decreases very slowly with increasing buffer capacity".

Heyman *et al.* employ a generic buffer model to explore the effect of long-range dependence [76]. The buffer has capacity B, and receives input at deterministic times. Let X_i be the number of arrivals at time T_i, $i = 1, 2, \ldots$. Let d

be the number of traffic that is processed during $[T_i,\ T_{i+1})$, referred to as the ith interval, and let V_i be the buffer content at the end of the ith interval. Then,

$$V_i = min\left\{(V_{i-1} + X_i - d)^+, B\right\}.$$

They show that the *resetting effect* of the buffer when buffer becomes empty and the *truncating effect of finite buffers* when buffer become full diminish the effect of long-range dependence. Since VBR video traffic has stringent QoS requirements, the traffic intensity for these services will not be large. Thus the resetting effect and truncating effect should be strong in practical regions. Their numerical examples show that Markovian chain model can accurately predict the QoS of real video teleconference and video sequences, which are long-range dependent. Based on it, they conclude that long-range dependence is not a crucial property in determining the buffer behavior of VBR video sources.

Ryu *et al.* investigate the practical implications of long-range dependence by studying the behavior of buffers fed with VBR video sources over a range of desirable cell loss rates and buffer sizes [97]. Using results based on large deviations theory, they introduce the notion of *Critical Time Scale* (CTS). For a given buffer size, link capacity, and the marginal distribution of frame size, the CTS of a VBR video source is defined as the number of frames whose correlations contribute to the cell loss rate. They show that whether the model is Markovian or has the long-range dependence property, its CTS is finite, it attains a small value for a small buffer, and is a non-decreasing function of the buffer size. In other words, under realistic scenarios of ATM buffer dimensioning, the number of frame correlations which affect buffer overflow probability is finite and small even in the presence of the long-range dependence property.

Ryu *et al.* uses simulation to validate their claims. They use the superposition of FBNDP (Fractional-Binomial-Noise-Driven Poisson Process) and $DAR(1)$ (discrete autoregressive process of order 1) as their long-range dependent VBR video traffic model, in which the long-term correlation and the short-term correlation can be effectively controlled. Their simulation shows that the buffer overflow probability of the long-range dependent traffic can be accurately captured by $DAR(p)$

process (discrete autoregressive process of order p), which is short-range dependent, *under the practical ranges of buffer size and cell loss ratio.* Their numerical results show that:

- even in the presence of long-range dependence, long-term correlations do not have significant impact on cell loss rate; and
- short-term correlations have dominant effect on cell loss rate, and therefore, well-designed Markov traffic models are effective for predicting QoS of long-range dependent VBR video traffic.

They conclude that it is unnecessary to capture the long-term correlation of a real-time VBR video source under realistic ATM buffer dimensioning scenarios as far as the cell loss rate and maximum buffer delays are concerned.

Grossglauser *et al.* find the existence of the CTS independently [96]. They argue that most of recent modeling work has failed to consider the impact of two important parameters, namely the finite range of time scales of interest in performance evaluation and prediction problems, and the first-order statistics such as the marginal distribution of the process. They introduce a modulated fluid traffic model in which the correlation function of the fluid rate matches that of an asymptotically second-order self-similar process with given Hurst parameter up to an arbitrary cutoff time lag, then drops to zero. Numerical experiments are performed to evaluate the performance of a single server queue fed with the above fluid input process. They find that the amount of correlation that needs to be taken into account for performance evaluation depends not only on the correlation structure of the source traffic, but also on time scales specific to the system under study. For example, the time scale associated with a queueing system is a function of the maximum buffer size. For finite buffer queues, they find that the impact of the correlation in the arrivals process on loss becomes nil beyond a time scale referred to as the correlation horizon. This means, in particular, that for performance-modeling purposes, any model among the panoply of available models can be chosen as long as the chosen model captures the correlation structure of the source traffic up to the correlation horizon.

In [59], Grossglauser *et al.* study a robust measurement-based admission control with emphasis on the impact of estimation errors, measurement memory, call-level dynamics and separation of time scales. Their work [59], [15] identifies a *critical time-scale* $\widetilde{T_h}$ such that aggregate traffic fluctuation slower than $\widetilde{T_h}$ can be tracked by the admission controller and compensated for by connection admissions and departures. Fluctuations faster than $\widetilde{T_h}$ have to be absorbed by reserving spare bandwidth on the link. Using Gaussian aggregate traffic model and heavy traffic approximations, the critical time scale is shown to scale as $T_h/\sqrt{n}$, where T_h is the average flow duration and n is the size of the link in terms of the number of flows it can carry. The main insight that can be gained from their work is that call level dynamics, i.e. connection admissions and departures can diminish the impact of long-range dependence on the performance of a MBAC.

In summary, the debate on the impact of long-range dependence on bandwidth allocation and network performance, especially its impact on the performance of a MBAC, and whether or not self-similar model should be used for traffic modeling, has never stopped since the discovery of the existence of long-range dependence in network traffic. This is due to the complexity of the problem. There are too many factors to be considered, which include, traffic characteristics, statistical multiplexing, call level dynamics, resetting and truncating effects of finite size buffers, network buffer size, network utilization level, QoS requirements of traffic sources, etc. Considering that, in this thesis, we primarily rely on simulation rather than theoretical analysis to analyze the impact of long-range dependence. Long-range dependent real traffic traces will be used in the simulation to test the impact of long-range dependence on the proposed CAC scheme.

2.4 Summary

In this chapter we reviewed current research on connection admission control, traffic modeling and performance analysis. Considering the vast number of publications in the field, this literature review by no means can be exhaustive. Other research noteworthy in this field includes, but not limited to: the work by Yamada

et al. [98], which presents a traffic measurement method for accurate measurement of traffic characteristics like mean traffic rate and autocorrelations.

Consistent work by Li and his group on bandwidth allocation, traffic modeling and network traffic measurements, sheds some insights into the problems [99], [100], [17], [101], [28], [102], [103][1]. Their work focuses on the impact of frequency-domain traffic characteristics on bandwidth allocation, traffic modeling, buffer engineering and traffic measurements. Their contributions can be summarized as:

1. Traffic measurements and network buffer engineering: they decompose the traffic into three frequency domains (angular frequency): low-frequency traffic in $0 < |\omega| \leq \omega_L$, high-frequency traffic in $|\omega| \geq \omega_H$ and mid-frequency components in $\omega_L < |\omega| < \omega_H$, where $\omega_L = \frac{0.01\pi}{d_{max}}$, $\omega_H = \frac{2\pi}{d_{max}}$ and d_{max} is the network buffer size measured in time. They show that the use of buffers is most effective for high frequency traffic and ineffective for low frequency traffic. Accordingly low-frequency traffic should be assigned its peak bandwidth and high-frequency traffic should be assigned its average bandwidth. Since the link bandwidth allocation of low- and high-frequency traffic requires no measurement of second-order statistics, the timescale of interest for traffic measurement must be identified in the range $[d_{max}, 200d_{max}]$. Their study also suggests that higher-order (3rd order and above) traffic statistics are generally unimportant to queueing solutions, and can therefore be ignored in traffic modeling and measurements.

2. Traffic modeling: they develop a circulant modulated rate process to match two primary statistical functions: traffic rate distribution and autocorrelation.

Lee *et al.* use variable sized measurement windows in their MBAC [22], [25].

The proposed CAC schemes, although varies in details, can be described by the two open-loop architectures shown in Fig. 2.1 and Fig. 2.2.

[1]More publications of theirs can be found at their webpage http://www.ece.utexas.edu/~sanqi . Here we only talk about their contributions in the field of interest for this thesis.

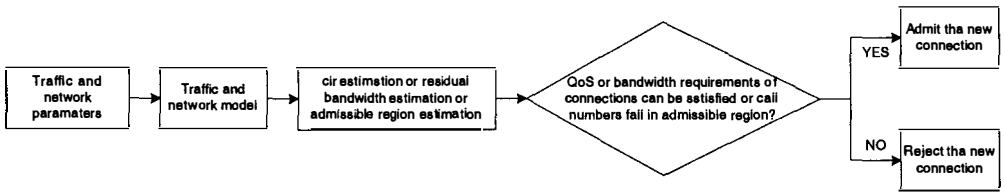


Figure 2.1: Architecture of traditional CAC scheme

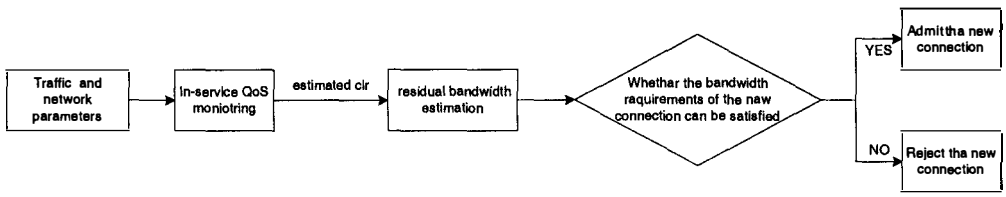


Figure 2.2: Architecture of CAC scheme using in-service monitoring

CAC schemes that fit into the architecture of Fig. 2.1 obtain traffic parameters from either traffic measurements or traffic descriptors. These parameters are then fed into the chosen traffic and network model. Based on the performance analysis of traffic sources, which is described by traffic model parameters, in the network model, either an estimate of QoS parameter when the new connection request is admitted, or an estimate of the residual bandwidth which is unused by existing network connections, or an admissible region, is obtained. If the QoS parameter when the new connection is admitted is less than QoS objective, or the residual bandwidth is greater than the bandwidth requirement of the new connection request, or the numbers of connections fall within the admissible region, the new connection is admitted. Otherwise, the new connection is rejected.

There are several other measurement-based admission control schemes which are based on in-service QoS monitoring [69], [70], [71]. They adopt the architecture shown in Fig. 2.2. These CAC schemes obtain an estimate of the CLR of existing traffic sources in the network through an in-service QoS monitoring scheme. This CLR estimate is then used to estimate the residual bandwidth. If the residual bandwidth is greater than the bandwidth requirement of the new connection request the new connection is admitted. Otherwise it is rejected. There are fundamental drawbacks in these open-loop architectures. The performance

of a CAC scheme based on an open-loop architecture relies on accurate traffic and network models, and accurate model parameters. There is no feedback on its performance that can be used to adapt the CAC scheme to achieve the optimum performance of the CAC scheme in the open-loop architecture. On the other hand, the complexity and heterogeneity of network traffic make it very difficult, if at all possible, to obtain accurate traffic models. In this thesis we shall present a novel CAC scheme based on a closed-loop architecture that is able to overcome the inherent drawbacks of the open-loop architecture.

Chapter 3

Loss Performance Analysis

In this chapter, the loss performance of heterogeneous traffic sources in the network will be analyzed. As introduced in section 1.2.4, our analysis follows the time-scale decomposition approach of Hui [14], and our loss performance analysis rests naturally on the burst time scale, or sometimes referred to as rate-variation scale in the literature. At the burst time scale, the finer cell scale granularity is ignored and the input process is characterized by its instantaneous rate. Consequently, fluid flow model appears as a natural modeling tool. At the burst time scale, problems arise from possible excesses of the total input rate over the output rate.

3.1 Fluid Flow Model

In the fluid flow model all traffic sources send their traffic to a network buffer with buffer size B, while at the same time the buffer is drained continuously at a rate C which is equal to the link capacity. The incoming traffic is lost when the buffer becomes full. Fig. 3.1 illustrates the fluid flow model.

Denote the sum of the instantaneous rates of all traffic sources on the link at time t by R_t. Stability condition then requires that:

$$E(R_t) < C. \tag{3.1}$$

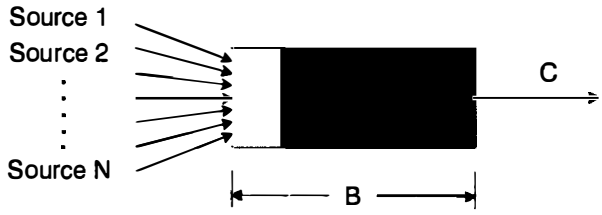


Figure 3.1: Fluid flow model

Assuming that the fluid queue has always been operational (that is, since $t = -\infty$), we define a backward accumulation process W_s as:

$$W_s \triangleq \int_{-s}^{0} (R_t - C)dt. \tag{3.2}$$

The steady state queue length distribution is then given by [104]:

$$P(Q > x) = P(sup_{s \geq 0} W_s > x). \tag{3.3}$$

Therefore the cell loss probability (or buffer overflow probability) can be obtained:

$$P(Q > B) = P(sup_{s \geq 0} W_s > B).$$

This cell loss probability can be used to approximate a more meaningful QoS parameter, cell loss ratio.

Although cell loss probability and cell loss ratio closely match in value in most cases, the difference between these two terms should be understood. Cell loss probability refers to the probability that the buffer content of an infinite buffer queue exceeds a certain threshold, i.e. the buffer size B. Cell loss probability is an effective mathematical tool to investigate the cell loss of traffic sources because its value closely matches cell loss ratio. Cell loss ratio is defined as the ratio of the number of cells lost to the total number of cells offered to the network. In theoretical analysis, cell loss ratio is much more difficult to estimate than cell loss probability. Therefore people choose to use cell loss probability to approximate cell loss ratio. This approximation is successful in most cases, but not in all cases.

Recent studies show that the finite range of time scale of practical interest should be considered in performance evaluation and prediction problems [96], [97]. And the truncating effect of a finite buffer when the buffer becomes full diminishes the effect of long-range dependence [76] (refer to section 2.3.3 for details). These results suggest that the fact that in real applications buffer size is limited, possibly small, should be considered in performance evaluation. However, cell loss probability comes from evaluation of infinite buffer queues and fails to account for limited buffer size. Thus it is possible that some discrepancies may be found between practice and theoretical analysis which is based on analysis of cell loss probability.

Depending on the multiplexing option adopted in the system design, two approaches can be distinguished in the context of fluid flow model. Using *Rate Envelop Multiplexing*, buffers are provided only for cell scale queueing and the system appears as a bufferless system at the burst scale. In the *Rate Sharing Multiplexing*, buffers are provided to absorb at least part of the burst scale traffic fluctuations and the system appears as a buffered system at the burst scale. Both approaches are adopted in performance analysis of network traffic. This study adopts the rate envelop multiplexing approach, which is also referred to as bufferless fluid flow model (bffm).

3.2 Bufferless Fluid Flow Model

Bufferless fluid flow model assumes that there is no buffer in the system at the burst scale, therefore cell loss occurs if and only if the sum of the instantaneous rates of all traffic sources exceeds link capacity C. Fig. 3.2 illustrates the cell loss in the bffm.

Let us denote the sum of the instantaneous rates of all traffic sources at time t by R_t. Then the cell loss rate of the aggregate traffic can be evaluated as:

$$\text{Cell Loss Rate} = E\left(R_t - C\right)^+ . \tag{3.4}$$

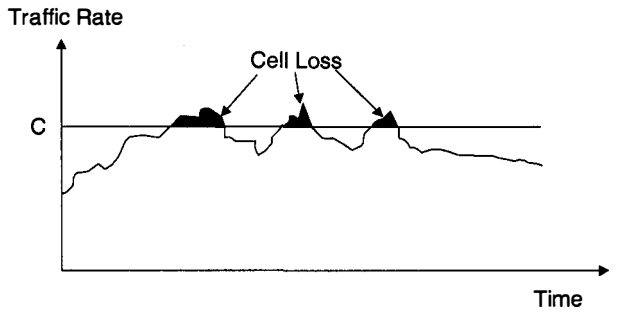


Figure 3.2: Cell loss in bufferless fluid flow model

The cell loss ratio of the aggregate traffic is given by:

$$clr = \frac{E\left(R_t - C\right)^+}{E\left(R_t\right)}. \tag{3.5}$$

Eq. 3.5 has a simple form. However, to obtain an estimate of cell loss ratio using Eq. 3.5, we need to know the probability density function (pdf) of the aggregate traffic, which in turn relies on convolution of the pdfs of instantaneous traffic rates of individual traffic sources making up the aggregate traffic. In real applications, the convolution is difficult to compute and the pdf of traffic rate of individual traffic source is difficult to obtain. To simplify the theoretical analysis as well as computation, we introduce the cell loss rate function (clrf) to characterize loss performance of heterogeneous traffic sources.

3.3 Cell Loss Rate Function

Let us define a function $F : R \rightarrow R^+$ as:

$$F(y) \triangleq E\left[(X - y)^+\right] \tag{3.6}$$

where R is a set of all real values and R^+ is a set of all non-negative real values. We call F the cell loss rate function of a non-negative random variable X. Our study shows that clrf is an effective tool for investigating cell loss in the bffm. In this thesis, since we are discussing traffic rates of traffic sources, all random

variables are non-negative random variables. The clrf has many attractive features which facilitate the analysis of cell loss in the bffm. For example,

$$F(C) = E\,(X - C)^+$$

denotes the cell loss rate of a traffic source X on a link with link capacity C. Traffic sources with similar clrf can be regarded as equivalent in the cell loss analysis. From the definition of clrf it can be shown that:

$$F(y) = F(0) - y \quad \text{for } y < 0, \tag{3.7}$$

and

$$F(0) = E(X). \tag{3.8}$$

Therefore, the cell loss ratio of X on a link with link capacity C can be evaluated as:

$$clr = \frac{F(C)}{F(0)}. \tag{3.9}$$

It is our novel contribution to use clrf to investigate the loss performance of heterogeneous traffic sources. By employing the clrf for loss performance analysis, both theoretical analysis and computation are significantly simplified.

First, we shall introduce two important properties of the clrf which are used in our theoretical analysis.

Property 1 *Let f and g denote the traffic density distribution of independent traffic sources X_1 and X_2 respectively. Then the cell loss rate function of $X_1 + X_2$ equals $F * g$, where F is the clrf of X_1, and $*$ denotes convolution.*

Proof: Construct a function $h : R \rightarrow R^+$ such that:

$$h(x) = \begin{cases} -x & x < 0 \\ 0 & x \geq 0 \end{cases}.$$

It can be shown that the clrf of X_1 is given by:

$$\begin{aligned} F(y) &= E\left[(X_1 - y)^+\right] \\ &= \int_y^\infty (x-y)\, f(x)dx \\ &= \int_y^\infty h\,(y-x)\, f(x)dx \\ &= \int_{-\infty}^\infty h(y-x) f(x)dx \\ &= h * f(y). \end{aligned}$$

Thus the clrf of $X_1 + X_2$, denoted by FG, is:

$$\begin{aligned} FG(y) &= E\left[(X_1 + X_2 - y)^+\right] \\ &= \int_y^\infty (x-y)\,(f * g(x))\, dx \\ &= h * (f * g(y)) \\ &= (h * f) * g(y) \\ &= F * g(y). \end{aligned}$$

■

Property 2 *Let f, g and q denote the traffic density distribution of independent traffic sources X_1, X_2 and X_3 respectively. Denote the cell loss rate functions of X_1, X_2, $X_1 + X_2$ and $X_2 + X_3$ by F, G, FG and GQ respectively. If for the cell loss rate functions F and G, $F(y) \geq G(y)$ for all $y \in R$, then $FQ(y) \geq GQ(y)$ for all $y \in R$.*

Proof: Note that clrf and the traffic density distribution function is always non-negative. The proof of this property is straightforward. ■

These two properties enable us to decompose the complex analysis of the multiplexing of several traffic sources into the simpler analysis of individual sources.

3.4 Stochastic Ordering Theory

To analyze the clrf of a traffic source X, we turn to the stochastic ordering theory [105], [106].

Definition 1 *Given two random variables X and Y, we say that X is smaller than Y with respect to the* increasing convex ordering, *written as $X <_{icx} Y$, if for the distribution functions of X and Y, denoted by F_d and G_d,*

$$\int_{-\infty}^{+\infty} \phi(x) dF_d(x) \leq \int_{-\infty}^{+\infty} \phi(x) dG_d(x) \tag{3.10}$$

holds for all increasing convex function ϕ, for which the integral exists [105, page 10].

Let us recall the definition of a convex function. Denote by I an arbitrary interval of R (possible infinite), a function $f : I \rightarrow R$ is convex if

$$f(\alpha \times x + (1-\alpha) y) \leq \alpha f(x) + (1-\alpha) f(y)$$

for all x, $y \in I$, $\alpha \in [0, 1]$.

The relation $<_{icx}$ is a *partial order* on the set $\mathcal{P}$ of distributions with finite means, therefore the increasing convex ordering has the following three properties which are inherent properties of a partial order.

Property 3 *(Reflexivity) $X <_{icx} X$.*

That is, a random variable X is smaller than itself with respect to the increasing convex ordering.

Property 4 *(Transitivity) If $X <_{icx} Y$ and $Y <_{icx} Z$, then $X <_{icx} Z$.*

Property 5 *(Antisymmetry) If $X <_{icx} Y$ and $Y <_{icx} X$, then $X = Y$, where $X = Y$ implies that X and Y are independent identically distributed (i.i.d.) random variables.*

It is following property of the increasing convex ordering that enables us to use it to analyze the clrf.

Lemma 1 *$X <_{icx} Y$ if and only if*

$$E\left[(X-x)^+\right] \leq E\left[(Y-x)^+\right] \tag{3.11}$$

for any $x \in R$.

See Theorem A in section 1.3 of [105] for a proof. Condition 3.11 can be rewritten in the following equivalent forms:

1. $E\left[\max\left(x,\ X\right)\right] \leq E\left[\max\left(x,\ Y\right)\right]$, where $x \in R$.
2. $\int_x^\infty \overline{F}_d(t)dt \leq \int_x^\infty \overline{G}_d(t)dt, x \in R$, where $\overline{F}_d(t) = 1 - F_d(t), \overline{G}_d(t) = 1 - G_d(t)$, and F_d and G_d are the distribution function of X and Y respectively.

From Lemma 1, $X <_{icx} Y$ if and only if the clrf of X is smaller than or equal to the clrf of Y for any real value. We shall now introduce the following property of the increasing convex ordering:

Proposition 1 *If $X_1, \ldots, X_n$ are independent and $Y_1, \ldots, Y_n$ are independent, and $X_i <_{icx} Y_i$, $i = 1, \ldots, n$, then*

$$g(X_1, \ldots, X_n) <_{icx} g(Y_1, \ldots, Y_n)$$

for all increasing convex function g.

See Proposition 8.5.4 of [106] for a proof.

According to Proposition 1, if X_i and Y_i, $i = 1, \ldots, n$, are all independent random variables, then $X_i <_{icx} Y_i$ implies that $\sum_{i=1}^n X_i <_{icx} \sum_{i=1}^n Y_i$ since $g(X_1, \ldots, X_n) = \sum_{i=1}^n X_i$ is an increasing convex function in each X_i [107]. Using Property 2 and Lemma 1, it can also be shown that, if X, Y and Z are independent random variables, then $Y <_{icx} Z$ implies that $X + Y <_{icx} X + Z$. These two properties are important for our theoretical analysis.

An important *special case* of the increasing convex ordering is when X and Y have the same mean value, i.e. $E(X) = E(Y)$. In this case, we say that X is smaller than Y with respect to the *convex ordering*, written as $X <_{cx} Y$, because the characterizing inequality 3.10 holds for all convex functions [106, Corollary 8.5.2].

Definition 2 *Given two random variables X and Y, we say that X is smaller than Y with respect to the* convex ordering, *written as $X <_{cx} Y$, if for the distribution function of X and Y, denoted by F_d and G_d,*

$$\int_{-\infty}^{+\infty} \phi(x) dF_d(x) \leq \int_{-\infty}^{+\infty} \phi(x) dG_d(x) \tag{3.12}$$

holds for all convex function ϕ, for which the integral exists [105, page 10].

It is the convex ordering which will be widely used for our loss performance analysis. Since convex ordering is a special case of the increasing convex ordering, it can be easily shown that all properties of the increasing convex ordering also apply to the convex ordering. For two independent traffic sources X and Y, $X <_{cx} Y$ implies that not only the cell loss generated by X is less than or equal to that of Y. Moreover, according to Eq. 3.8 and Eq. 3.9, the cell loss ratio of X is also less than or equal to that of Y.

One property of the convex ordering is the following:

Property 6 *If $X <_{cx} Y$, then $\sigma(X) \leq \sigma(Y)$ provided that $\sigma(X)$ and $\sigma(Y)$ exist, where $\sigma(\bullet)$ denotes the standard deviation of random variable $\bullet$.*

Proof: Since the function x^2 is a convex function, $E(X^2) \leq E(Y^2)$. From the definition of convex ordering $E(X) = E(Y)$, thus it can be shown that

$$E(X^2) - [E(X)]^2 \leq E(Y^2) - [E(Y)]^2 .$$

■

This property of the convex ordering will be used in the traffic measurement analysis in Chapter 4.

3.5 Traffic Source Model

For simplicity, let us consider on-off traffic sources. According to the probability density distribution of on and off periods, on-off source models can be classified into exponential on-off source model, periodic on-off source model, Pareto on-off

source model, etc. They are widely used for loss performance analysis in CAC schemes [16], [52], [65], [21], [108], [109], [110]. On-off source models have been successfully used to characterize the on/off nature of an individual source or source element, like packetized voice and video [109], [111], [112]. They provide the worst case analysis of traffic sources in terms of cell loss. Recent studies indicate that on-off source models are also suitable for modeling self-similar traffic [77], [85].

Many current modeling techniques model each traffic source or source element by an ON-OFF source. These techniques fail to apply in high speed networks, simply because of the exploded input state space when a large number of diverse sources are multiplexed on each link. Take a simple example of 3 heterogeneous traffic types multiplexed on a link. Assume that each traffic type has 100 connections. The overall state space will reach an overwhelming size of about $1,000,000$. Efforts have been made to reduce the input state space. Rasmussen *et al.* [108] propose a cell loss upper bound for heterogeneous ON-OFF sources. They conjecture that, for N heterogeneous ON-OFF sources which have peak cell rates $pcr_1, \cdots, pcr_N$, and the sum of their mean cell rates is S, then the case of N homogeneous ON-OFF sources, each source has peak cell rate $pcr = max\{pcr_1, \cdots, pcr_N\}$ and mean cell rate $mcr = S/N$, will constitute the upper bound in terms of cell loss. Using this argument, the state space of our earlier example reduces to 300, resulting in great computational savings. However, if sources with large bandwidth demands and sources with small bandwidth demands are multiplexed, the upper bound will be far from the actual cell loss. Hwang *et al.* [99] propose a method of input state space reduction. Their research shows that the queueing performance is mainly determined by low frequency components of traffic sources [100], [113]. In their approach 2-state Markov chains are built to statistically match with the power spectrum function and probability density function (pdf) of the aggregate traffic. The state space of each traffic type can be reduced to 30 using his method. Hence the state space of our example reduces to $27,000$. This is still too large to implement practically.

Lee *et al.* [21] propose an algorithm which is suitable for real-time estimation of cell loss of the multiplexing of heterogeneous on-off sources. However their approach can only be applied to traffic descriptor-based CAC, hence tight characterization of mean cell rate (mcr) in traffic descriptors is necessary in their approach. In real networks, it is very difficult for all traffic sources to tightly characterize their mcr, so their approach only has limited use. Moreover, choosing an appropriate quantization unit for traffic rate remains a problem in their approach. Small quantization unit will result in huge computation efforts and large quantization unit will result in too conservative cell loss estimates.

Here we try to solve these problems with a different approach. Using clrf defined in section 3.3 and the stochastic ordering theory, we develop a cell loss upper bound suitable for real-time cell loss ratio evaluation for heterogeneous on-off sources. The upper bound is then applied to design a measurement-based CAC.

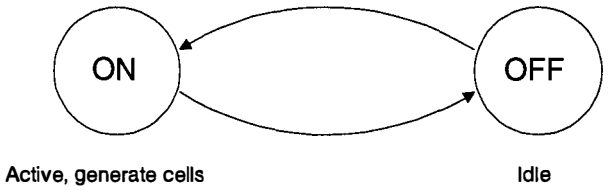
An on-off source generates cells at a peak cell rate (pcr) denoted by pcr in active periods. In idle periods no cells are generated. Denote the mean cell rate of the on-off source by mcr. The activity parameter of an on-off source is defined as the ratio of mean cell rate to peak cell rate:

$$p \triangleq \frac{mcr}{pcr}. \tag{3.13}$$

The probability that an on-off source is active or idle is given by p or $1-p$ respectively. Fig. 3.3 illustrates the source model and traffic model of an ON-OFF source.

Assume there are n independent heterogeneous on-off sources $X_1, \ldots, X_n$ on the link, where $X_{i(i=1,\ldots,n)}$ has peak cell rate pcr_i, mean cell rate mcr_i and activity parameter $p_i = mcr_i/pcr_i$. The probability mass function (pmf) of $X_{i(i=1,\ldots,n)}$ can be expressed as:

$$f^{(p_i)}_{1,(pcr_i)}(x) = \begin{cases} p_i & x = pcr_i \\ 1-p_i & x = 0 \\ 0 & otherwise \end{cases} \tag{3.14}$$



(a) Source model

(b) Traffic model

Figure 3.3: On-off source model

and the pmf of $\sum_{i=1}^{n} X_i$ can be expressed as:

$$f^{(p_1,\cdots,p_n)}_{n,(pcr_1,\cdots,pcr_n)}(x) = f^{(p_1)}_{1,(pcr_1)} * \cdots * f^{(p_n)}_{1,(pcr_n)}(x). \tag{3.15}$$

In this chapter we use subscript n, subscript (pcr) and superscript (p) to denote the number of on-off sources, peak cell rates of sources and their activity parameters respectively when we need to emphasize the dependence of a function on these parameters. For example, in Eq. 3.15 $f^{(p_1,\cdots,p_N)}_{n,(pcr_1,\cdots,pcr_N)}(x)$ denotes the pmf of n independent heterogeneous ON-OFF sources with activity parameters $p_1, \cdots, p_N$ and peak cell rates $pcr_1, \cdots, pcr_N$ respectively. $f^{(p)}_{n,(pcr)}(x)$ denotes the pmf of n independent homogeneous ON-OFF sources and each source has activity parameter p and peak cell rate pcr.

Since the instantaneous traffic rate of an on-off source or the multiplexing of several on-off sources, denoted by X, is a discrete random variable, definition in Eq. 3.6 becomes

$$\begin{aligned} F(y) &\triangleq E\left[(X-y)^+\right] \\ &\triangleq \sum_x (x-y)^+ f(x) \end{aligned} \tag{3.16}$$

where $f(x)$ is the probability mass function of discrete random variable X.

We shall now introduce an important property of on-off sources.

Proposition 2 *Assume X is an arbitrary traffic source such that $E(X) = mcr$ and $\|X\|_\infty = pcr$, where $\|X\|_\infty = inf\{x : Pr\{X > x\} = 0\}$. Denote by Y an on-off source with mean cell rate mcr and peak cell rate pcr. Then $X <_{cv} Y$.*

Proof: Let F_d and G_d denote the distribution function of X and Y. Since Y is an on-off source, G_d is given by:

$$G_d(x) = \begin{cases} 1 & x \geq pcr \\ 1-p & 0 \leq x < pcr \\ 0 & x < 0 \end{cases}$$

where p is the activity parameter of Y. Noting that

$$E(X) = E(Y) = mcr,$$

from Lemma 1 and the property of convex ordering, it suffices to show that for any $y \geq 0$:

$$\int_y^\infty \overline{F}_d(x)dx \leq \int_y^\infty \overline{G}_d(x)dx.$$

Define $\alpha = inf\ \{y :\ F_d(y) \geq 1-p\}$.

- If $y \geq \alpha$, then for any $x \in [y,\ pcr)$

  $$F_d(x) \geq 1-p = G_d(x),$$

  and for $x \in [pcr,\ \infty)$,

  $$F_d(x) = G_d(x) = 1.$$

  So for any $x \in [y,\ \infty)$, $\overline{F}_d(x) \leq \overline{G}_d(x)$. Therefore it can be shown that

  $$\int_y^\infty \overline{F}_d(x)dx \leq \int_y^\infty \overline{G}_d(x)dx.$$

- If $0 \leq y < \alpha$, then for any $x \in [0, y]$, $F_d(x) < 1 - p = G_d(x)$. Hence

$$\int_0^y F_d(x)dx \leq \int_0^y G_d(x)dx.$$

Therefore,

$$\begin{aligned}
\int_y^\infty \overline{F}_d(x)dx &= \int_0^\infty \overline{F}_d(x)dx - \int_0^y \overline{F}_d(x)dx \\
&= mcr - \int_0^y (1 - F(x))\, dx \\
&= mcr - y + \int_0^y F_d(x)dx \\
&\leq mcr - y + \int_0^y G_d(x)dx \\
&= \int_y^\infty \overline{G}_d(x)dx,
\end{aligned}$$

where in the above equation we have used the fact that

$$\int_0^\infty \overline{F}_d(x)dx = \int_0^\infty \overline{G}_d(x)dx = mcr.$$

■

This proof uses the same idea as that used in Appendix B of [107] which shows that $X <_{icx} Y$.

Proposition 2 implies that among traffic sources with the same mean cell rate and peak cell rate, on-off source constitutes the worst-case in loss performance analysis.

3.6 Loss Performance Analysis of Heterogeneous On-Off Sources

In this section we shall investigate the loss performance of heterogeneous traffic sources using on-off source model.

3.6.1 Loss Performance Analysis of Individual On-Off Source

First we shall introduce a lemma on the multiplexing of two independent on-off sources. This lemma is used in many parts of this thesis. The comparison of the multiplexing of two independent on-off sources constitutes one of the basic cases for loss performance analysis of on-off sources, from which we can extend to the comparison of the multiplexing of many on-off sources.

Lemma 2 *Let X_1, X_2 be two independent on-off sources with peak cell rates $\alpha_1 \times pcr$ and $(1-\alpha_1) \times pcr$; and activity parameters p_1 and p_2 respectively, where $0.5 \leq \alpha_1 < 1$. Let Y_1 and Y_2 be two independent on-off sources with peak cell rates $\alpha_2 \times pcr$ and $(1-\alpha_2) \times pcr$, and activity parameters q_1 and q_2 respectively, where $0.5 \leq \alpha_2 < 1$. Moreover, $E(X_1+X_2) = E(Y_1+Y_2)$ and $\alpha_1 \leq \alpha_2$. Then $X_1+X_2 <_{cx} Y_1+Y_2$ if and only if $p_1p_2 \leq q_1q_2$ and $p_1+p_2-p_1p_2 \geq q_1+q_2-q_1q_2$.*

Proof: Using Lemma 1 and the property of convex ordering, it suffices to show that the clrf of X_1+X_2, denoted by $F(y)$, is smaller than or equal to that of Y_1+Y_2, denoted by $G(y)$, for any $y \in R$.

For $y < 0$:

$$F(y) = G(y) = E(X_1+X_2) - y.$$

For $y > pcr$:

$$F(y) = G(y) = 0.$$

We shall now consider $F(y)$ and $G(y)$ in the region $0 \leq y \leq pcr$. $F(y)$ and $G(y)$ are calculated respectively as:

$$F(y) = \begin{cases} (pcr-y)p_1p_2; & \alpha_1 pcr < y \leq pcr \\ pcr\,[p_1p_2 + \alpha_1 p_1(1-p_2)] - p_1 y; & \\ & (1-\alpha_1)pcr < y \leq \alpha_1 pcr \\ pcr\,[\alpha_1 p_1 + (1-\alpha_1)p_2] - (p_1+p_2-p_1p_2)y; & \\ & 0 \leq y \leq (1-\alpha_1)pcr \end{cases} \tag{3.17}$$

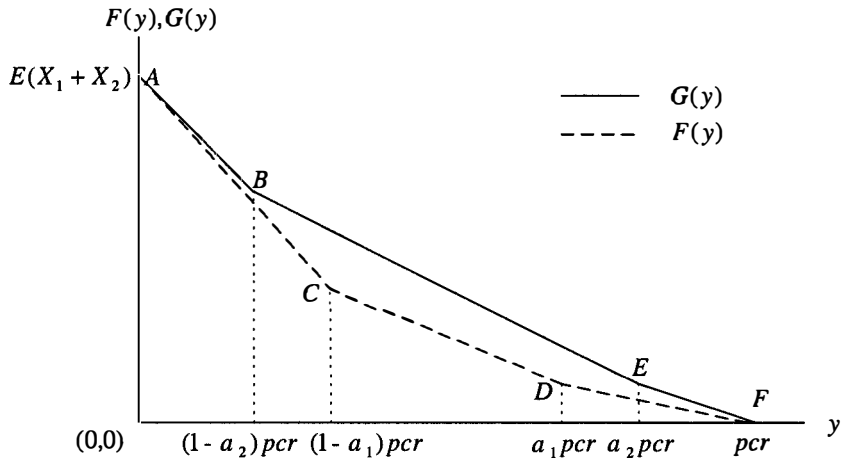


Figure 3.4: Illustration of $F(y)$ and $G(y)$

$$G(y) = \begin{cases} (pcr - y)q_1q_2; & \alpha_2 pcr < y \leq pcr \\ pcr\,[q_1q_2 + \alpha_2 q_1(1-q_2)] - q_1 y; & \\ & (1-\alpha_2)pcr < y \leq \alpha_2 pcr \\ pcr\,[\alpha_2 q_1 + (1-\alpha_2)q_2] - (q_1 + q_2 - q_1q_2)y; & \\ & 0 \leq y \leq (1-\alpha_2)pcr \end{cases} . \tag{3.18}$$

This lemma can be proved by solving the inequality $F(y) \leq G(y)$ directly. However, here we use a simpler method. $F(y)$ and $G(y)$ are continuous functions and linear in segments. From conditions of this lemma, it can be shown that $F(0) = G(0) = E(X_1 + X_2)$, $F(pcr) = G(pcr) = 0$ and $\alpha_1 \leq \alpha_2$. An illustration of $F(y)$ and $G(y)$ is shown in Fig. 3.4.

It is clearly shown in Fig. 3.4 that $F(y) \leq G(y)$ if and only if the following condition is satisfied:

$$\begin{cases} k_{AB} \leq k_{AC} \\ k_{EF} \leq k_{DF} \end{cases}$$

where the symbol $k_\bullet$ means the slope of segment $\bullet$.

Using Eq. 3.17 and Eq. 3.18, it can be shown that

$$\begin{cases} k_{AB} &= -(q_1 + q_2 - q_1 q_2) \\ k_{AC} &= -(p_1 + p_2 - p_1 p_2) \\ k_{EF} &= -q_1 q_2 \\ k_{DF} &= -p_1 p_2 \end{cases}.$$

Therefore we are able to conclude that $X_1 + X_2 <_{cx} Y_1 + Y_2$, if and only if

$$\begin{cases} -(q_1 + q_2 - q_1 q_2) \leq -(p_1 + p_2 - p_1 p_2) \\ -q_1 q_2 \leq -p_1 p_2 \end{cases}.$$

■

We shall now introduce another theorem on on-off sources.

Theorem 1 *Let X and Y be two independent on-off sources with the same mean cell rate denoted by mcr. X and Y have peak cell rates pcr_X and pcr_Y respectively. If $pcr_X \leq pcr_Y$, then $X <_{cx} Y$.*

Proof: Since $E(X) = E(Y)$, it suffices to show that the clrf of Y is greater than or equal to that of X. The activity parameters of X and Y are $\frac{mcr}{pcr_X}$ and $\frac{mcr}{pcr_Y}$ respectively. We shall now calculate the clrf of X and Y. The clrf of X is

$$F(y) = \begin{cases} mcr - y & y \leqslant 0 \\ mcr\left(1 - \frac{y}{pcr_X}\right) & 0 < y \leqslant pcr_X \\ 0 & y > pcr_X \end{cases},$$

and the clrf of Y is

$$G(y) = \begin{cases} mcr - y & y \leqslant 0 \\ mcr\left(1 - \frac{y}{pcr_Y}\right) & 0 < m \leqslant pcr_Y \\ 0 & m > pcr_Y \end{cases}.$$

It is then easy to show that for any $y \in R$

$$F(y) \leq G(y). \tag{3.19}$$

■

A result to the same effect of Theorem 1 is proved using another approach in [21]. Here it is proved using the clrf. Comparing the proof here and that in [21], the advantage of using clrf for cell loss analysis in the bffm is revealed. Loss performance analysis using clrf is much simpler than that using other methods.

Theorem 1 was used by Lee *et al.* [21] to design a traffic-descriptor based CAC scheme. In their CAC scheme, they choose a traffic rate unit u, and all traffic rates are normalized with regard to u. However, not all peak cell rates of traffic sources are going to be integer multiples of u. The case when pcr is not an integer will result in additional complexity in computation. Theorem 1 shows that the cell loss of an on-off source with peak cell rate pcr and mean cell rate mcr is upper bounded by a quantized on-off source whose peak cell rate is $\lceil pcr \rceil$ and mean cell rate is mcr. Therefore Theorem 1 enables us to design a CAC scheme based on the quantized on-off sources which are easier to handle in computation. On the basis of Theorem 1, Lee *et al.* designed a CAC scheme capable of real-time CLR estimation for the multiplexing of heterogeneous on-off sources (refer to section 2.1.3 for details of their CAC scheme).

Theorem 1 will be used for loss performance analysis and CAC scheme design in this thesis.

3.6.2 A Cell Loss Upper Bound for Bernoulli Sources

Based on Lemma 2, we shall now introduce an important theorem about heterogeneous Bernoulli sources. Bernoulli sources are on-off sources with the same peak cell rate.

Theorem 2 *Let $X_1, \ldots, X_n$ be n independent heterogeneous Bernoulli sources with the same peak cell rate, and their activity parameters are given by $p_1, \cdots, p_n$ respectively. Their activity parameters are subject to $p_1 + \cdots + p_n = P$. Let $Y_1, \ldots, Y_n$ represent n independent homogeneous Bernoulli sources, where Y_i has the same peak cell rate as $X_{i\,(i=1,\ldots,n)}$ and an activity parameter $p = P/n$. Then $\sum_{i=1}^{n} X_i <_{cx} \sum_{i=1}^{n} Y_i$.*

Proof: It is easy to show that

$$E\left(\sum_{i=1}^{n} X_i\right) = E\left(\sum_{i=1}^{n} Y_i\right).$$

Then from Lemma 1, it suffices to show that the clrf of $\sum_{i=1}^{n} Y_i$ is greater than or equal to that of $\sum_{i=1}^{n} X_i$ for any real value.

The proof is by induction on n.

1. First let us consider the case where $n = 2$. We must show that the clrf of $X_1 + X_2$ is less than or equal to that of $Y_1 + Y_2$, where X_1, X_2 are two independent heterogeneous Bernoulli sources with activity parameters p_1, p_2, and the same peak cell rate pcr; and Y_1, Y_2 are two homogeneous Bernoulli sources with activity parameter $p = (p_1 + p_2)/2$ and peak cell rate pcr. It can be shown that:

$$p^2 = \frac{(p_1 + p_2)^2}{4} \geq p_1 p_2.$$

Then, from $2p = p_1 + p_2$, we are able to conclude that:

$$2p - p^2 \leq p_1 + p_2 - p_1 p_2.$$

Using Lemma 2 ($\alpha_1 = \alpha_2 = 0.5$), it can be shown that:

$$X_1 + X_2 <_{cx} Y_1 + Y_2.$$

That is, the clrf of X_1 and X_2 is less than or equal to that of $Y_1 + Y_2$.

2. Let $F_n^{(p_1,\cdots,p_n)}(y)$ represent the clrf of n heterogeneous Bernoulli sources. The ith Bernoulli source X_i has an activity parameter p_i and a peak cell rate pcr. Let $F_n^{(p)}(y)$ represent the clrf of n homogeneous Bernoulli sources where each Bernoulli source has an activity parameter $p = (p_1 + \cdots + p_n)/n$ and a peak cell rate pcr. Suppose $F_n^{(p_1,\cdots,p_n)}(y) \leq F_n^{(p)}(y)$ holds for the case when $n = k$, i.e.

$$F_k^{(p_1,\cdots,p_n)}(y) \leq F_k^{(p)}(y) \quad \text{for any } y \in R$$

where $p = (p_1 + \cdots + p_k)/k$. Let us consider the case when $n = k+1$. It can be shown that:

$$
\begin{aligned}
& F_{k+1}^{(p_1,\cdots,p_{k+1})}(y) \\
= \; & F_k^{(p_1,\cdots,p_k)} * f_1^{(p_{k+1})}(y) \\
\leqslant \; & F_k^{\left(\frac{p_1+\cdots+p_k}{k}\right)} * f_1^{(p_{k+1})}(y) \\
= \; & \left[F_{k-1}^{\left(\frac{p_1+\cdots+p_k}{k}\right)} * f_1^{\left(\frac{p_1+\cdots+p_k}{k}\right)}\right] * f_1^{(p_{k+1})}(y) \\
= \; & \left[F_{k-1}^{\left(\frac{p_1+\cdots+p_k}{k}\right)} * f_1^{(p_{k+1})}\right] * f_1^{\left(\frac{p_1+\cdots+p_k}{k}\right)}(y) \\
\leqslant \; & F_k^{\left(\frac{(k-1)\frac{p_1+\cdots+p_n}{k}+p_{k+1}}{k}\right)} * f_1^{\left(\frac{p_1+\cdots+p_k}{k}\right)}(y) \\
& \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots
\end{aligned}
$$

where in the above equations, $f_1^{(p)}$ denotes the pmf of an on-off source with peak cell rate pcr and activity parameter p. Property 1 is used in the derivations.

Define a sequence a_i so that the above procedure can be expressed as:

$$
\begin{aligned}
& F_{k+1}^{(p_1,\cdots,p_{k+1})}(y) \\
\leqslant \; & F_k^{(a_1)} * f_1^{(a_0)}(y) \\
& \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\
\leqslant \; & F_k^{(a_i)} * f_1^{(a_{i-1})}(y) \\
& \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots .
\end{aligned}
\tag{3.20}
$$

Eq. 3.20 implies that the clrf of $k+1$ heterogeneous on-off sources with activity parameter $p_1, \ldots, p_n$ is smaller than or equal to that of $k+1$ heterogeneous on-off sources, among which k on-off sources have activity parameters a_1 and the other on-off source has activity parameter a_0. The clrf of the $k+1$ heterogeneous on-off sources, among which k on-off sources have activity parameters a_1 and the other on-off source has activity parameter a_0, is smaller than or equal to the clrf of $k+1$ heterogeneous on-off

sources, among which k on-off sources have activity parameter a_2 and the other on-off source has activity parameter a_1. .....

It can be shown that:

$$a_i = \frac{(i-1)a_{i-1}+a_{i-2}}{i}$$

$$a_0 = p_{k+1} \quad \text{and} \quad a_1 = \frac{p_1+\cdots+p_k}{k}$$

Solving for a_i, when $i \geqslant 2$ it can be obtained that:

$$\begin{aligned} a_i &= a_1 - \frac{1-\left(-\frac{1}{k}\right)^{i-1}}{k+1}(a_1 - a_0) \\ &= \frac{p_1+\cdots+p_k}{k} \quad \frac{1-\left(-\frac{1}{k}\right)^{i-1}}{k+1}\left(\frac{p_1+\cdots+p_k}{k} - p_{k+1}\right). \end{aligned}$$

Therefore it can be shown that:

$$\lim_{i\to\infty} a_i = \lim_{i\to\infty} a_{i-1} = \frac{p_1+\cdots+p_{k+1}}{k+1} \quad . \tag{3.21}$$

This implies that when the process in Eq. 3.20 goes on and on, the activity parameters of the k homogeneous Bernoulli sources and the single Bernoulli source in the above equations will converge to $\frac{p_1+\cdots+p_{k+1}}{k+1}$. So it can be concluded that:

$$\begin{aligned} & F_{k+1}^{(p_1,\cdots,p_{k+1})}(y) \\ \leqslant\ & F_k^{\left(\frac{p_1+\cdots+p_{k+1}}{k+1}\right)} * f_1^{\left(\frac{p_1+\cdots+p_{k+1}}{k+1}\right)}(y) \\ =\ & F_{k+1}^{\left(\frac{p_1+\cdots+p_{k+1}}{k+1}\right)}(y). \end{aligned}$$

Therefore, from the assumption that $F_n^{(p)}(y) \geqslant F_n^{(p_1,\cdots,p_n)}(y)$ holds for the case when $n = k$, we derive the conclusion that the inequality also holds for the case when $n = k+1$.

Combining step 1 and step 2 we conclude that:

$$F_n^{(p)}(y) \geqslant F_n^{(p_1,\cdots,p_n)}(y) \quad \text{for any } y \in R$$

holds for any n.

■

Theorem 2 states that homogeneous Bernoulli sources generate more cell loss than that of heterogeneous Bernoulli sources. Theorem 2 was first proposed as a conjecture by Rasmussen *et al.* [108]. Based on Theorem 2, they proposed that the cell loss of n heterogeneous on-off sources, whose maximum peak cell rate is pcr, is upper bounded by that of n homogeneous on-off sources with peak cell rate pcr, where the sum of mean cell rates remains the same. This conjecture can be easily proved now using Theorem 2 and Theorem 1. In addition to [108], Theorem 2 is also used for loss performance analysis in many other literature [16], [114], [115].

In real networks, many traffic sources of the same type have the same peak cell rate. However, because of their specific application circumstances, they have different mean cell rates. Theorem 2 is very useful for analyzing this kind of traffic sources.

3.6.3 A Cell Loss Upper Bound for Heterogeneous Traffic Sources

On the basis of the theorems, lemmas and properties shown previously, we shall now introduce an important theorem about independent heterogeneous on-off sources. First we introduce a theorem which will be used in the proof of Theorem 4.

Theorem 3 *Let X_1, X_2 be two independent on-off sources with peak cell rates $\alpha_1 \times pcr$, $(1-\alpha_1) \times pcr$, and activity parameters p_1, p_2 respectively, where $0.5 \leq \alpha_1 < 1$. Let α_2 be any value satisfying $\alpha_1 \leq \alpha_2 < 1$. Then there exist two independent on-off sources Y_1, Y_2 with peak cell rates $\alpha_2 \times pcr$, $(1-\alpha_2) \times pcr$, and $E(Y_1+Y_2) = E(X_1+X_2)$, such that $X_1 + X_2 <_{cv} Y_1 + Y_2$.*

Proof: We show the existence of such Y_1 and Y_2 by constructing an instance of such on-off sources. Without loss of generality, we assume pcr to be 1. Then the mean cell rate of $X_1 + X_2$ is:

$$mcr = E(X_1 + X_2) = \alpha_1 \times p_1 + (1-\alpha_1) \times p_2$$

Since p_1 and p_2 are the activity parameters of on-off sources, it naturally follows that $0 \leq p_1 \leq 1$ and $0 \leq p_2 \leq 1$. We shall show that the two independent

on-off sources Y_1 and Y_2 with activity parameters q_1 and q_2 given by:

$$\begin{cases} q_1 = \frac{mcr+\sqrt{mcr^2-4\alpha_2(1-\alpha_2)p_1p_2}}{2\alpha_2} \\ q_2 = \frac{mcr-\sqrt{mcr^2-4\alpha_2(1-\alpha_2)p_1p_2}}{2(1-\alpha_2)} \end{cases} \tag{3.22}$$

is an instance of the desired on-off sources.

First, we shall show that the activity parameters q_1 and q_2 given in Eq. 3.22 have meaningful values, i.e.

$$mcr^2 - 4\alpha_2(1-\alpha_2)p_1p_2 \geq 0 \tag{3.23}$$

$$0 \leq q_1 \leq 1 \tag{3.24}$$

$$0 \leq q_2 \leq 1 \quad . \tag{3.25}$$

Since the function $g(\alpha) = \alpha(1-\alpha)$ is a decreasing function when $\alpha \geq 0.5$, it can be shown that

$$\alpha_2(1-\alpha_2) \leq \alpha_1(1-\alpha_1).$$

Therefore,

$$\begin{aligned} & mcr^2 - 4\alpha_2(1-\alpha_2)p_1p_2 \\ = \; & [\alpha_1p_1 + (1-\alpha_1)p_2]^2 - 4\alpha_2(1-\alpha_2)p_1p_2 \\ = \; & \alpha_1^2p_1^2 + 2\alpha_1(1-\alpha_1)p_1p_2 + (1-\alpha_1)^2p_2^2 - 4\alpha_2(1-\alpha_2)p_1p_2 \\ \geq \; & \alpha_1^2p_1^2 - 2\alpha_1(1-\alpha_1)p_1p_2 + (1-\alpha_1)^2p_2^2 \\ = \; & [\alpha_1p_1 - (1-\alpha_1)p_2]^2 \geq 0. \end{aligned} \tag{3.26}$$

Thus inequality 3.23 is proved. It is trivial to show that $0 \leq q_1$ and $0 \leq q_2$. Now let us consider the inequality $q_1 \leq 1$.

Define a function $f(x)$ as follows:

$$f(x) = 4\alpha_2^2 - 4\alpha_2\left[\alpha_1x + (1-\alpha_1)p_2\right] + 4\alpha_2(1-\alpha_2)p_2x,$$

then

$$\begin{aligned} \frac{df(x)}{dx} &= -4\alpha_1\alpha_2 + 4\alpha_2(1-\alpha_2)p_2 \\ &= 4\alpha_2\left[(1-\alpha_2)p_2 - \alpha_1\right]. \end{aligned}$$

Since

$$(1-\alpha_2)p_2 - \alpha_1 \leq 1 - \alpha_2 - \alpha_1 \leq 0,$$

$f(x)$ is a decreasing function. Therefore,

$$\begin{aligned} f(p_1) &\geq f(1) \\ &= 4\alpha_2^2 - 4\alpha_2\left[\alpha_1 + (1-\alpha_2)p_2\right] + 4\alpha_2(1-\alpha_2)p_2 \\ &= 4\alpha_2(\alpha_2 - \alpha_1) \geq 0. \end{aligned}$$

So, we have shown that

$$f(p_1) = 4\alpha_2^2 - 4\alpha_2\left[\alpha_1 p_1 + (1-\alpha_1)p_2\right] + 4\alpha_2(1-\alpha_2)p_1p_2 \geq 0.$$

Then from the following derivation, we are able to conclude that $q_1 \leq 1$:

$$\begin{aligned} q_1 &= \frac{mcr + \sqrt{mcr^2 - 4\alpha_2(1-\alpha_2)p_1p_2}}{2\alpha_2} \leq 1 \\ &\Leftrightarrow \sqrt{mcr^2 - 4\alpha_2(1-\alpha_2)p_1p_2} \leq 2\alpha_2 - mcr \qquad (3.27) \\ &\Leftrightarrow mcr^2 - 4\alpha_2(1-\alpha_2)p_1p_2 \leq 4\alpha_2^2 - 4\alpha_2 \times mcr + mcr^2 \\ &\Leftrightarrow 4\alpha_2^2 - 4\alpha_2 \times mcr + 4\alpha_2(1-\alpha_2)p_1p_2 \geq 0 \\ &\Leftrightarrow 4\alpha_2^2 - 4\alpha_2\left[\alpha_1 p_1 + (1-\alpha_1)p_2\right] + 4\alpha_2(1-\alpha_2)p_1p_2 \geq 0 \end{aligned}$$

Note that in inequality 3.27, $2\alpha_2 \geq pcr = 1 \geq mcr$, so

$$2\alpha_2 - mcr \geq 0.$$

Let us now consider the inequality $q_2 \leq 1$. It can be shown that:

$$\begin{aligned} & q_2 \leq 1 \\ \Leftrightarrow\ & \frac{mcr - \sqrt{mcr^2 - 4\alpha_2(1-\alpha_2)p_1p_2}}{2(1-\alpha_2)} \leq 1 \\ \Leftrightarrow\ & mcr \leq \sqrt{mcr^2 - 4\alpha_2(1-\alpha_2)p_1p_2} + 2(1-\alpha_2) \\ \Leftrightarrow\ & mcr^2 \leq \left[\sqrt{mcr^2 - 4\alpha_2(1-\alpha_2)p_1p_2} + 2(1-\alpha_2)\right]^2 \\ \Leftrightarrow\ & \sqrt{mcr^2 - 4\alpha_2(1-\alpha_2)p_1p_2} \geq \alpha_2 p_1 p_2 - (1-\alpha_2) \qquad (3.28) \end{aligned}$$

When $p_1p_2 \leq \frac{1-\alpha_2}{\alpha_2}$:

$$\alpha_2 p_1 p_2 - (1-\alpha_2) \leq 0.$$

We have shown in inequality 3.26 that

$$mcr^2 - 4\alpha_2(1-\alpha_2)p_1p_2 \geq 0$$

therefore inequality 3.28 is true.

Now let us consider the case when $p_1p_2 > \frac{1-\alpha_2}{\alpha_2}$, it can be shown that

$$\begin{aligned}
& \sqrt{mcr^2 - 4\alpha_2(1-\alpha_2)p_1p_2} \geq \alpha_2 p_1 p_2 - (1-\alpha_2) \\
\Leftrightarrow\ & mcr^2 - 4\alpha_2(1-\alpha_2)p_1p_2 \geq \\
& \alpha_2^2 p_1^2 p_2^2 + (1-\alpha_2)^2 - 2\alpha_2(1-\alpha_2)p_1p_2 \\
\Leftrightarrow\ & mcr^2 \geq \alpha_2^2 p_1^2 p_2^2 + (1-\alpha_2)^2 + 2\alpha_2(1-\alpha_2)p_1p_2 \\
\Leftrightarrow\ & mcr^2 \geq [\alpha_2 p_1 p_2 + (1-\alpha_2)]^2 \\
\Leftrightarrow\ & \alpha_1 p_1 + (1-\alpha_1)p_2 - \alpha_2 p_1 p_2 - (1-\alpha_2) \geq 0.
\end{aligned}$$

Define a function

$$h(x) = x \times p_1 + (1-x)p_2 - \alpha_2 p_1 p_2 - (1-\alpha_2).$$

It can be shown that:

$$\begin{aligned}
h(\alpha_2) &= \alpha_2 p_1 + (1-\alpha_2)p_2 - \alpha_2 p_1 p_2 - (1-\alpha_2) \\
&= (1-p_2)\left[\alpha_2 p_1 - (1-\alpha_2)\right].
\end{aligned}$$

Also, It can be shown that:

$$p_1p_2 > \frac{1-\alpha_2}{\alpha_2} \Rightarrow p_1 > \frac{1-\alpha_2}{\alpha_2},$$

therefore $h(\alpha_2) \geq 0$. It can also be shown that:

$$\begin{aligned}
h(\frac{1}{2}) &= \frac{1}{2}(p_1+p_2) - \alpha_2 p_1 p_2 - (1-\alpha_2) \\
&\geq \sqrt{p_1p_2} - \alpha_2 p_1 p_2 - (1-\alpha_2) \\
&= \frac{1}{4\alpha_2} - (1-\alpha_2) - \left(\sqrt{\alpha_2} \times \sqrt{p_1p_2} - \frac{1}{2\sqrt{\alpha_2}}\right)^2.
\end{aligned}$$

When $p_1p_2 = 1$, $h(\frac{1}{2}) = 0$; when $p_1p_2 = \frac{1-\alpha_2}{\alpha_2}$,

$$h(\frac{1}{2}) = \frac{\sqrt{1-\alpha_2}}{\sqrt{\alpha_2}}\left(\sqrt{\alpha_2} - \sqrt{1-\alpha_2}\right)^2 \geq 0.$$

Therefore, for $\frac{1-\alpha_2}{\alpha_2} < p_1p_2 \leq 1$, it can be concluded that $h(\frac{1}{2}) \geq 0$.

Since $\frac{1}{2} \leq \alpha_1 \leq \alpha_2$ and $h(x)$ is a linear function, it can be concluded that $h(\alpha_1) \geq 0$. Hence, when $p_1p_2 > \frac{1-\alpha_2}{\alpha_2}$, the inequality of Eq. 3.28 is also true. So we conclude that $q_2 \leq 1$, and the activity parameters q_1 and q_2 given in Eq. 3.22 are valid values.

Next we shall show that for on-off sources Y_1 and Y_2 with peak cell rates α_2 and $(1-\alpha_2)$, and activity parameters q_1 and q_2,

$$X_1 + X_2 <_{cx} Y_1 + Y_2.$$

Since

$$E(Y_1 + Y_2) = \alpha_2 q_1 + (1-\alpha_2)q_2 = mcr = E(X_1 + X_2),$$

using Lemma 2, it suffices to show that:

$$\begin{cases} p_1p_2 \leq q_1q_2 \\ p_1 + p_2 - q_1q_2 \geq q_1 + q_2 - q_1q_2 \end{cases} \tag{3.29}$$

First we show that,

$$\begin{aligned} q_1q_2 &= \frac{mcr + \sqrt{mcr^2 - 4\alpha_2(1-\alpha_2)p_1p_2}}{2\alpha_2} \\ &\quad \times \frac{mcr - \sqrt{mcr^2 - 4\alpha_2(1-\alpha_2)p_1p_2}}{2(1-\alpha_2)} \\ &= \frac{mcr^2 - \left[\sqrt{mcr^2 - 4\alpha_2(1-\alpha_2)p_1p_2}\right]^2}{4\alpha_2(1-\alpha_2)} \\ &= p_1p_2 \end{aligned}$$

Then we shall show that $p_1 + p_2 \geq q_1 + q_2$. It can be shown that:

$$
\begin{aligned}
& q_1 + q_2 \leq p_1 + p_2 \\
\Leftrightarrow\ & \frac{mcr + \sqrt{mcr^2 - 4\alpha_2(1-\alpha_2)p_1p_2}}{2\alpha_2} \\
& + \frac{mcr - \sqrt{mcr^2 - 4\alpha_2(1-\alpha_2)p_1p_2}}{2(1-\alpha_2)} \leq p_1 + p_2 \\
\Leftrightarrow\ & \frac{mcr + (1-2\alpha_2)\sqrt{mcr^2 - 4\alpha_2(1-\alpha_2)p_1p_2}}{2\alpha_2(1-\alpha_2)} \leq p_1 + p_2 \\
\Leftrightarrow\ & (2\alpha_2 - 1)\sqrt{mcr^2 - 4\alpha_2(1-\alpha_2)p_1p_2} \\
& \geq mcr - 2\alpha_2(1-\alpha_2)(p_1+p_2) \\
\Leftarrow\ & (2\alpha_2 - 1)^2 \left[mcr^2 - 4\alpha_2(1-\alpha_2)p_1p_2\right] \\
& \geq \left[mcr - 2\alpha_2(1-\alpha_2)(p_1+p_2)\right]^2 \\
\Leftrightarrow\ & \left(4\alpha_2^2 - 4\alpha_2 + 1\right)\left[mcr^2 - 4\alpha_2(1-\alpha_2)p_1p_2\right] \\
& \geq mcr^2 - 4\alpha_2(1-\alpha_2)(p_1+p_2)mcr + \left[2\alpha_2(1-\alpha_2)(p_1+p_2)\right]^2 \\
\Leftrightarrow\ & 4\alpha_2(1-\alpha_2)\left[mcr^2 - (2\alpha_2-1)^2 p_1p_2\right] \\
& \geq 4\alpha_2(1-\alpha_2)\left[(p_1+p_2)mcr + \alpha_2(1-\alpha_2)(p_1+p_2)^2\right] \\
\Leftrightarrow\ & mcr^2 - (2\alpha_2-1)^2 p_1p_2 \geq (p_1+p_2)mcr + \alpha_2(1-\alpha_2)(p_1+p_2)^2 \\
\Leftrightarrow\ & -mcr^2 - p_1p_2 - \alpha_2(1-\alpha_2)(p_1-p_2)^2 + (p_1+p_2)mcr \geq 0
\end{aligned}
$$

Here we are able to show that:

$$
\begin{aligned}
& -mcr^2 - p_1p_2 - \alpha_2(1-\alpha_2)(p_1-p_2)^2 + (p_1+p_2)mcr \\
= & -\left[\alpha_1 p_1 + (1-\alpha_1)p_2\right]^2 - p_1p_2 - \alpha_2(1-\alpha_2)(p_1-p_2)^2 \\
& +(p_1+p_2)\left[\alpha_1 p_1 + (1-\alpha_1)p_2\right] \\
= & (p_1-p_2)^2(\alpha_2-\alpha_1)(\alpha_2+\alpha_1-1) \geq 0
\end{aligned}
$$

Therefore $q_1 + q_2 \leq p_1 + p_2$, and Eq. 3.29 is satisfied.

We have shown that the two on-off sources Y_1 and Y_2 with the given activity parameters have $X_1 + X_2 <_{cx} Y_1 + Y_2$. Thus this theorem is proved by constructing an instance of such on-off sources. ■

This theorem is used in proving the following theorem which is about loss performance of heterogeneous on-off sources:

Theorem 4 *Let $X_1, \ldots, X_n$ be n independent heterogeneous on-off sources with peak cell rates $pcr_1, \ldots, pcr_n$ and mean cell rates $mcr_1, \ldots, mcr_n$ respectively. Let $Y_1, \ldots, Y_m$ represent m independent homogeneous on-off sources with peak cell rate pcr and mean cell rate mcr, where $pcr \geq max\{pcr_1, \ldots, pcr_n\}$, $m = \left\lceil \frac{\sum_{i=1}^{n} pcr_i}{pcr} \right\rceil$ and $mcr = \frac{\sum_{i=1}^{n} mcr_i}{m}$. Then*

$$\sum_{i=1}^{n} X_i <_{cv} \sum_{i=1}^{m} Y_i \tag{3.30}$$

Proof: We prove this theorem using the induction method.

1. Let us consider the case when $n = 1$, then

$$m = \left\lceil \frac{pcr_1}{pcr} \right\rceil = 1.$$

We must show that $X_1 <_{cx} Y_1$, where X_1 is an on-off source with peak cell rate pcr_1 and mean cell rate mcr_1, and Y_1 is an on-off source with peak cell rate $pcr \geq pcr_1$ and mean cell rate mcr_1. This can be proved easily using Theorem 1.

2. Suppose inequality 3.30 holds for the case when $n = k$, we must show that it also holds for the case when $n = k + 1$.

First we point out that since $max\{pcr_1, \ldots, pcr_k\} \leq max\{pcr_1, \ldots, pcr_{k+1}\}$, from our supposition that inequality 3.30 holds for $n = k$ and $pcr \geq max\{pcr_1, \ldots, pcr_k\}$, it naturally follows that inequality 3.30 also holds for $n = k$ and $pcr \geq max\{pcr_1, \ldots, pcr_{k+1}\}$. Now let us consider the two on-off source X_k and X_{k+1}, we shall consider the following two cases:

(a) When $pcr_k + pcr_{k+1} \leq pcr$, using Proposition 2, it can be shown that

$$X_k + X_{k+1} <_{cv} Z$$

where Z is an on-off source with peak cell rate $pcr_Z = pcr_k + pcr_{k+1}$

and mean cell rate $mcr_Z = mcr_k + mcr_{k+1}$. Then applying our supposition for k on-off sources $X_1, \ldots, X_{k-1}, Z$, we can show that inequality 3.30 holds for the $k+1$ on-off sources $X_1, \ldots, X_{k+1}$, i.e.

$$\begin{aligned} & X_1 + \cdots + X_{k-1} + X_k + X_{k+1} \\ <_{cv} \; & X_1 + \cdots + X_{k-1} + Z \\ <_{cv} \; & Y_1 + \cdots + Y_{m_{k+1}} \end{aligned}$$

where

$$m_{k+1} = \left\lceil \frac{\sum_{i=1}^{k-1} pcr_i + pcr_Z}{pcr} \right\rceil = \left\lceil \frac{\sum_{i=1}^{k+1} pcr_i}{pcr} \right\rceil,$$

and $Y_{i(i=1,\ldots,m_{k+1})}$ is an independent on-off source with peak cell rate pcr and mean cell rate $mcr = \frac{\sum_{i=1}^{k+1} mcr_i}{m_{k+1}}$.

(b) When $pcr_k + pcr_{k+1} > pcr$, using Theorem 3, we are able to find two independent on-off source Z_1 and Z_2. Z_1 has peak cell rate $pcr_{Z_1} = pcr_k + pcr_{k+1} - pcr$ and mean cell rate mcr_{Z_1}. Z_2 has peak cell rate pcr and mean cell rate mcr_{Z_2}, where $mcr_{Z_2} = mcr_k + mcr_{k+1} - mcr_{Z_1}$. Z_1 and Z_2 satisfy that:

$$X_k + X_{k+1} <_{cv} Z_1 + Z_2.$$

Note that here $pcr \geq max\{pcr_k, pcr_{k+1}\}$ is equivalent to the condition in Theorem 3 that $\alpha_1 \leq \alpha_2$.

From $pcr \geq max\{pcr_1, \ldots, pcr_{k+1}\}$, it is guaranteed that $pcr \geq pcr_{Z_1}$. Then applying our supposition for the k independent on-off sources $X_1, \ldots, X_{k-1}, Z_1$, it can be obtained that:

$$\begin{aligned} & X_1 + \cdots + X_{k-1} + Z_1 \\ <_{cv} \; & \lambda_1 + \cdots + \lambda_{m'_k} \end{aligned}$$

where

$$\begin{aligned} m'_k &= \left\lceil \frac{\sum_{i=1}^{k-1} pcr_i + pcr_{Z_1}}{pcr} \right\rceil \\ &= \left\lceil \frac{\sum_{i=1}^{k+1} pcr_i - pcr}{pcr} \right\rceil \\ &= \left\lceil \frac{\sum_{i=1}^{k+1} pcr_i}{pcr} \right\rceil - 1 \end{aligned}$$

$\lambda_{i(i=1,\ldots,m'_k)}$ is an independent on-off source with peak cell rate pcr and mean cell rate

$$mcr_\lambda = \frac{\sum_{i=1}^{k-1} pcr_i + mcr_{Z_1}}{m'_k}.$$

Here we note that $\lambda_{i(i=1,\ldots,m'_k)}$ and Z_2 have the same peak cell rate pcr. Therefore using Theorem 2, it can be shown that

$$\begin{aligned} & \lambda_1 + \cdots + \lambda_{m'_k} + Z_2 \\ <_{cv} \; & Y_1 + \cdots + Y_{m_{k+1}} \end{aligned}$$

where

$$m_{k+1} = m'_k + 1 = \left\lceil \frac{\sum_{i=1}^{k+1} pcr_i}{pcr} \right\rceil$$

and $Y_{i(i=1,\ldots,m_{k+1})}$ is an independent on-off source with peak cell rate pcr and mean cell rate:

$$\begin{aligned} mcr &= \frac{\sum_{i-1}^{k-1} pcr_i + mcr_{Z_1} + mcr_{Z_2}}{m_{k+1}} \\ &= \frac{\sum_{i=1}^{k+1} pcr_i}{m_{k+1}} \end{aligned}$$

Therefore, from the supposition that inequality 3.30 holds for $n = k$, we arrive at the conclusion that it should also hold for $n = k + 1$.

Combining 1 and 2, we conclude that 3.30 holds for all n.

■

Theorem 4 is the core of our loss performance analysis. After proving Theorem 4, all other theorems and lemmas shown in this section become special cases of this theorem.

Theorem 4 provides a cell loss upper bound for heterogeneous on-off sources. It states that the cell loss of heterogeneous on-off sources is less than or equal to that of corresponding homogeneous on-off sources given in the theorem. The sum of mean cell rates of the homogeneous sources remains the same as that of heterogeneous sources, and the sum of peak cell rates of the homogeneous sources is also substantially the same as that of the heterogeneous sources. It is not difficult, using the clrf and the stochastic ordering theory, to prove that our upper bound is tighter than the upper bound proposed by Rassussen *et al.* [108].

The proposed upper bound can also be explained intuitively as follows: substituting n independent heterogeneous on-off sources with $\left\lceil \frac{\sum_{i=1}^{n} pcr_i}{pcr} \right\rceil$ independent homogeneous on-off sources, the sum of mean cell rates does not change, i.e. the traffic load is unchanged. However, with the decrease in the number of multiplexed on-off sources the aggregate traffic becomes more bursty. Therefore for the same utilization cell loss will increase.

Using Theorem 4 and Proposition 2, we easily obtain the following theorem with more general applicability than Theorem 4.

Theorem 5 *Let $X_1, \ldots, X_n$ be n independent heterogeneous traffic sources with peak cell rates $pcr_1, \ldots, pcr_n$ and mean cell rates $mcr_1, \ldots, mcr_n$ respectively. $X_1, \ldots, X_n$ may have any distribution. Let $Y_1, \ldots, Y_m$ represent m independent homogeneous on-off sources with peak cell rate pcr and mean cell rate mcr, where $pcr \geq max\{pcr_1, \ldots, pcr_n\}$, $m = \left\lceil \frac{\sum_{i=1}^{n} pcr_i}{pcr} \right\rceil$ and $mcr = \frac{\sum_{i=1}^{n} mcr_i}{m}$. Then*

$$\sum_{i=1}^{n} X_i <_{cv} \sum_{i=1}^{m} Y_i \tag{3.31}$$

3.7 Summary

Bufferless fluid flow model is widely used in the literature for loss performance analysis. In this chapter we proposed an efficient and effective means of investigating loss performance of heterogeneous traffic sources in the bffm. We defined the cell loss rate function and used it to characterize the cell loss of traffic sources in the bffm. The clrf significantly simplifies theoretical analysis as well as computation.

Stochastic ordering theory was used to analyze the clrf. The introduction of the stochastic ordering theory not only simplifies the theoretical analysis but also makes it possible to extend applications of theoretical analysis presented in this chapter to a broader area. More specifically, the loss performance analysis presented in this chapter not only applies to cell loss in the bffm. In fact it can be applied to any model where cell loss function is a convex function of the traffic rate.

On-off traffic source models were chosen to analyze the loss performance of heterogeneous traffic sources. On-off source models are widely used for loss performance analysis in CAC schemes. They have been successfully used to characterize the on/off nature of an individual source or source element, and they provide the worst case analysis of cell loss. The loss performance of heterogeneous on-off sources was widely discussed. A set of theorems were presented which have great significance in both loss performance analysis and connection admission scheme design. At last we presented a cell loss upper bound for heterogeneous on-off sources. The proposed upper bound is tighter than those found in the literature. Moreover we showed that the proposed upper bound for heterogeneous on-off sources can be extended to traffic sources with any traffic rate distribution.

It is worth noting that the loss performance analysis presented in this chapter is a steady-state analysis. It implies that the proposed upper bound constitutes the worst case in terms of average loss performance. *Theoretically speaking*, the average loss performance constraints may not be meaningful, if the aggregate traffic exhibits long-range dependence. Specifically, if the aggregate traffic exhibits long-range dependence, although average performance may be deemed to be fine,

there may be rare periods of time in which performance is consistently poor.

However, as pointed out in section 2.3.3, the actual impact of long-range dependent traffic on network performance is still a myth. Therefore, we advocate to investigate the impact of long-range dependence on CAC using real long-range dependent traffic sources. CAC using the proposed upper bound may not be applicable for long-range dependent traffic *provided that* long-range dependence does have significance in performance evaluation. In the next chapter, we shall investigate this problem.

Chapter 4

A Measurement-based Connection Admission Control Scheme

In Chapter 3 we presented theoretical analysis on the loss performance of heterogeneous traffic sources. The results obtained in Chapter 3 will be applied in this chapter to design a measurement-based connection admission control scheme.

To date, many connection admission control schemes have been proposed [26], [27], [28], [23], [29], [30], [16], [22]. These schemes can be classified into two basic categories: traffic descriptor-based CAC and measurement-based CAC. Traffic descriptor-based CAC uses the *a priori* traffic characterizations provided by sources at connection setup phase to calculate the probabilistic behavior of all existing connections in addition to the incoming one. It achieves high network utilization when traffic descriptions required by the CAC scheme are tight. Measurement-based CAC uses the *a priori* traffic characterizations only for the incoming connection and uses measurements to characterize existing connections. Under measurement-based CAC scheme, network utilization does not suffer significantly if traffic descriptions are not tight. However, because of the fact that source behavior is not static in general, it is difficult for measurement-based CAC to obtain traffic characteristics accurately from on-line measurements. Measurement-based CAC is able to deliver significant gain in utilization when there is a high degree of statistical multiplexing [26].

Considering the difficulty for both traffic descriptor-based and measurement-based CAC to obtain accurate traffic characteristics, the performance of a CAC scheme should not be measured only by the utilization achieved under ideal circumstances where traffic sources are all tightly characterized. Also, one must consider whether enough accurate traffic parameters can be obtained from sources and/or network practically, and, the robustness of the CAC scheme against the inaccuracies in those traffic characteristics. In addition to high network utilization an ideal CAC scheme should satisfy the following requirements [116], [20]:

- Simplicity: the scheme must be both economically implementable and fast. The traffic characteristics required by the CAC scheme should be easily and reliably obtained from the traffic sources and/or network.
- Flexibility: the scheme must not only be able to satisfy the current needs of network services but also be able to adapt to new services which are likely to evolve.
- Robustness: the scheme must be able to handle imperfect assumptions.

These principles will be used to govern the design of our CAC scheme.

A cell loss upper bound was proposed in Theorem 5 which can be applied to design a CAC scheme. Two traffic parameters are required in the upper bound to evaluate the cell loss ratio of heterogeneous traffic sources, i.e., the peak cell rate and the mean cell rate of the aggregate traffic. Both these traffic parameters can be obtained either from traffic descriptors of traffic sources or network traffic measurements. Therefore the proposed upper bound can be applied either to a traffic-descriptor based CAC and a measurement-based CAC.

When the upper bound is applied to design a traffic-descriptor based CAC, the traffic source is required to specify its mean cell rate and peak cell rate at the connection setup phase. To a very large extent, peak cell rate of a traffic source is determined by the nature of the traffic source or the traffic rate of the access line (e.g. the data rate of a telephone wire). For example, the pcr of an ordinary telephone voice traffic can only be $64Kbps$, $32Kbps$ or $16Kbps$, which

is determined by the encoding scheme. This pcr is also restricted by the data rate of the access line. Moreover, as introduced in section 1.2.2, in real networks, in addition to CAC there is usages parameter control, which is defined as the set of actions taken by the network to monitor and control traffic to protect network resources from malicious as well as unintentional misbehavior, which can affect the QoS of other already established connections. One of the most well-known UPC functions is leaky bucket algorithm. Peak cell rate can be easily monitored and regulated by UPC. Therefore, we consider that it is reasonable to assume that traffic source is able to specify its pcr accurately. However it becomes difficult for traffic source to predict its mcr accurately. Again taking telephone voice traffic for example, in additions to the factors mentioned above which affect pcr, the mcr of a voice traffic is also affected by some unpredictable factors, e.g. speed of a human's speech, occasion of the conversion, etc. Mean cell rate is also more difficult for UPC to monitor and regulate. So, it is difficult for a traffic to accurately specify its mean cell rate. Thus it is desirable to use network measurements to alleviate the burden on traffic sources to tightly specify their traffic parameters. Therefore we choose to design a measurement-based connection admission control scheme.

4.1 CAC Scheme

Based on the theoretical analysis in Chapter 3, we shall develop a measurement-based connection admission control scheme in this section. The principles introduced above is used to govern the CAC scheme design. Robustness, flexibility and simplicity become major concerns in our CAC design.

Let us select a traffic rate unit u such that u is greater than or equal to the maximum pcr of all traffic sources on the link. From here on in this chapter all traffic rates, as well as link capacity C, are normalized with respect to u unless otherwise specified. Without loss of generality we assume that link capacity C is an integer.

Assume there are n independent heterogeneous traffic sources, denoted by $X_1, \ldots, X_n$, on the link. The declared peak cell rate of traffic source $X_{i\,(i=1,\ldots,n)}$

is pcr_i. We keep a list of the declared peak cell rates of all connections on the link and denote the sum of the peak cell rates by PCR, i.e.

$$PCR = \sum_{i=1}^{n} pcr_i.$$

Denote the true value of the mean cell rate of $\sum_{i=1}^{n} X_i$ by r_T. Let $Y_1, \ldots, Y_m$ denote m independent homogeneous on-off sources where each on-off source has peak cell rate 1 (remember here that traffic rates are normalized by u) and activity parameter p. $\sum_{i=1}^{n} X_i$ and $\sum_{i=1}^{n} Y_i$ have the same mean cell rate. The parameters m and p are given by:

$$m = \lceil PCR \rceil, \tag{4.1}$$

and

$$p = \frac{r_T}{m}.$$

From Theorem 5,

$$\sum_{i=1}^{n} X_i <_{cv} \sum_{i=1}^{m} Y_i,$$

or in other words, the clrf of $\sum_{i=1}^{n} X_i$ is less than or equal to that of $\sum_{i=1}^{m} Y_i$, and the cell loss ratio of $\sum_{i=1}^{m} X_i$ is also less than or equal to that of $\sum_{i=1}^{n} Y_i$.

Since $Y_1, \ldots, Y_m$ are independent homogeneous on-off sources, the pmf of $\sum_{i=1}^{m} Y_i$ is given by the following binomial distribution:

$$f(x) = \begin{cases} \binom{m}{k} p^k (1-p)^{m-k} & x = k \\ 0 & else \end{cases}. \tag{4.2}$$

Denote the clrf of $\sum_{i=1}^{m} Y_i$ by F. From the property of clrf introduced in section 3.3, it can be shown that:

$$F(0) = E\left(\sum_{i=1}^{m} Y_i\right) = m \times p.$$

For any positive integer k,

$$\begin{aligned}
F(k) &= \sum_{x} (x-k)^{+} f(x) \\
&= \sum_{i=k+1}^{m} (i-k) f(i) \\
&= \sum_{i=k+1}^{m} (i-k+1) f(i) - \sum_{i=k+1}^{m} f(i) \\
&= \sum_{i=k}^{m} (i-k+1) f(i) - \sum_{i=k}^{m} f(i) \\
&= F(k-1) - 1 + \sum_{i=0}^{k-1} f(i).
\end{aligned}$$

Therefore the calculation of the clrf of $\sum_{i=1}^{m} Y_i$ can be simplified as:

$$F(k) = \begin{cases} m \times p & ; \quad k = 0 \\ F(k-1) - 1 + \sum_{i=0}^{k-1} f(i) & ; \quad k \geqslant 1 \end{cases} . \tag{4.3}$$

The cell loss ratio of $\sum_{i=1}^{m} Y_i$ is computed using the following equation:

$$clr = \frac{F(C)}{F(0)}.$$

According to Theorem 5 and the property of the convex ordering, cell loss ratio of $\sum_{i=1}^{m} Y_i$ is greater than or equal to that of $\sum_{i=1}^{n} X_i$.

4.1.1 Estimation of The Activity Parameter p

There are two critical parameters which are required for the estimation of the CLR, i.e. m and p. Parameter m is calculated from the sum of the declared peak cell rates of connections using Eq. 4.1. Parameter p is calculated from the sum of the mean cell rates of connections. Realizing that it is difficult for traffic sources to tightly characterize their mean cell rates, we obtain the sum of mean cell rates from on-line measurements. This is of course the mean cell rate of the link. We denote the measured mcr of the link by r. The activity parameter p is

then estimated from the measured mean cell rate of the aggregate traffic. We shall now describe the method of estimating the activity parameter p.

Since we obtain the mean cell rate of $\sum_{i=1}^{n} X_i$ from on-line measurements, p can be directly estimated as:

$$\widehat{p} = \frac{r}{m}. \tag{4.4}$$

Increasing the measurement window size, T_m, will increase the accuracy of the measured mean cell rate r as well as the accuracy of the estimation. Yamada *et al.* [98] introduce a method for choosing the measurement window size. Estimation of p from Eq. 4.4 is simple, however for accurate estimation a large measurement window size is required. Here we use another approach.

In ATM networks, traffic can only arrive in integer multiples of an ATM cell. Therefore we first choose a sampling period T_s such that the impact of such ATM cell level granularity on traffic rate measurements taken over T_s can be ignored. In our analysis, T_s is chosen to be 100 cell unit time. One cell unit time is the time required to transmit an ATM cell on the link. Denote the mean and the variance of the traffic rate sample r_1 measured over T_s by S_T and σ_T^2 , and the mean and the variance of the traffic rate sample r_K measured over a sampling period of $K \times T_s$ by S_K and σ_K^2 respectively. Assuming that the aggregate traffic is stationary, then it can be shown that S_K, σ_K, S_T and σ_T are related by

$$S_K = S_T,$$

and

$$\sigma_K^2 = \sigma_T^2 \left[\frac{1}{K} + \frac{1}{K^2} \sum_{i=1}^{K} (K-i)\rho_i \right],$$

where ρ_i is the autocorrelation between traffic rate samples taken over $[0, T_s]$ and over $[(i-1)T_s, i \times T_s]$. σ_K is a non-increasing function of K. In the above analysis we ignored the impact of call level dynamics, i.e., it is assumed that no call is admitted into the network or depart from the network during the measurement window. The impact of call level dynamics is discussed later at the end of this section.

When the number of connections on the link is large enough, according to the central limit theorem the aggregate traffic can be accurately modeled by Gaussian distribution. Therefore we assume that the aggregate traffic is Gaussian. If the following equation is chosen as an estimate of the mean cell rate:

$$\widehat{r} = r_K + \varepsilon \times \sigma_K$$

where ε is a constant. In order to satisfy the estimation objective

$$P\left(\widehat{r} \geq S_K\right) \geq 0.95, \tag{4.5}$$

we have to choose $\varepsilon = 1.65$. If the aggregate traffic is not Gaussian, a larger ε can be obtained using the Chebyshev's inequality. In the rest of this chapter, we only consider the case when the aggregate traffic is Gaussian traffic.

An estimate of σ_K can either be obtained directly from on-line traffic measurements, or can be obtained from on-line estimation of σ_T and autocorrelation function ρ. However, on-line estimation of these parameters is not an easy task, so we take another approach. Using Property 6 and Theorem 5, it can be shown that the variance of $\sum_{i=1}^{m} Y_i$ is the maximum of the variance of the aggregate traffic $\sum_{i=1}^{n} X_i$. The variance of $\sum_{i=1}^{m} Y_i$ is given by $m\frac{S_T}{m}\left(1 - \frac{S_T}{m}\right)$. Thus instead of measuring the variance directly, we estimate σ_T as follows:

$$\widehat{\sigma_T} = \sqrt{m\frac{r_K}{m}\left(1 - \frac{r_K}{m}\right)}. \tag{4.6}$$

When a large enough K is chosen such that traffic fluctuations with time scale larger than $K \times T_s$ are small, the estimated σ_T is larger than its true value despite possible fluctuations in r_K. $\widehat{\sigma_K}^2$ is then obtained as:

$$\widehat{\sigma_K}^2 = \frac{\widehat{\sigma_T}^2}{K^\delta} \tag{4.7}$$

where δ is a constant between 0 and 1.

δ can be obtained by inspecting the variance-time plot obtained from traffic measurements [79], [72]. In contrast to [79] and [72] which study the self-similarity in network traffic, our interest is mainly in the variance-time curve in a

relatively small time region from $10T_s$ to $1000T_s$. In this region even for short-range dependent traffic, a δ much smaller than 1 may be observed. In our simulation shown later, δ is a very stable value. This is because the traffic mix in the simulation is almost time-invariant. As a result, the autocorrelation in network traffic does not change significantly. However, our analysis on traffic measurement data from real ATM networks[1] shows that the value of δ will change slowly with time within a day. In that case, the smallest value of observed δ should be used in Eq. 4.7. As an example, Fig. 4.1 shows the variance time plot of the aggregate traffic in the saturation scenario presented later. In Fig. 4.1, traffic rate is normalized by the traffic rate unit u chosen for the saturation scenario. By fitting a line to the variance time plot using least square fit, we obtain a line with a slope of -0.3698. Therefore, for simulation using exponential on-off sources shown later, δ is chosen to be 0.35. For simulation using Motion-JPEG encoded video sources, using the same method δ is chosen to be 0.4. The advantage of this method is that the value of δ can be obtained empirically from off-line traffic analysis, therefore on-line estimation of the second order statistics is avoided. Here we would like to comment that this method will not give an accurate estimate of σ_K. However the estimated value of σ_K is generally larger than its true value, which will satisfy the estimation objective in Eq. 4.5.

Then an estimate of p can be obtained: $\widehat{p} = \frac{\widehat{r}}{m}$.

To summarize the above analysis, if the measurement window size is chosen to be $T_m = K \times T_s$, an estimate of p can be obtained as follows:

$$\widehat{p} = \frac{r}{m} + \alpha\sqrt{\frac{\frac{r}{m}(1-\frac{r}{m})}{m}}, \tag{4.8}$$

where

$$\alpha = 1.65 \times K^{-\frac{\delta}{2}}, \tag{4.9}$$

and r is the mean traffic rate measured over T_m.

[1]These traffic traces are collected by Waikatp Applied Network Dynamics group at the University of Auckland since November 1999. The time-stamp of measurement data in the trace has an accuracy of $1\mu s$. More details about the traces can be found at their webpage http://moat.nlanr.net/Traces/Kiwitraces .

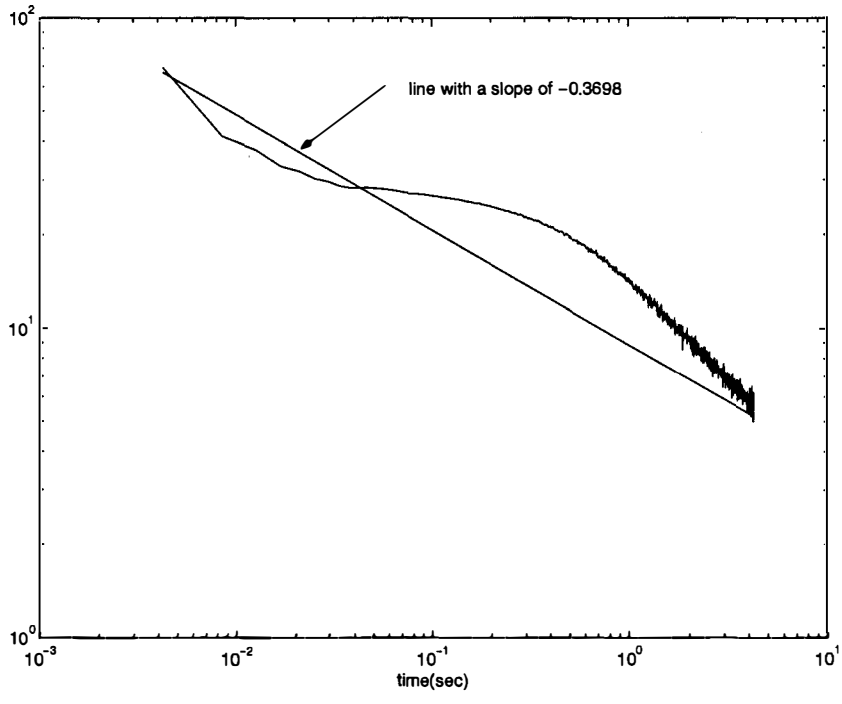


Figure 4.1: Variance time plot of the aggregate traffic rate in the saturation scenario

This approach first appeared in [16]. The introduction of the safety margin α enables us to significantly reduce the required measurement window size while maintaining the robustness of the estimation. Moreover, the safety margin introduces additional benefits. Loss performance analysis presented in Chapter 3 gave a cell loss upper bound of the aggregate traffic based on bufferless fluid flow model. The measurement scheme shown above also gives a robust estimate of p. Therefore the above measurement scheme will give a tight QoS guarantee. However, for a network with large buffers, cell loss ratio will decrease due to large buffer size. The proposed CAC scheme is conservative for network with large buffers since it is based on loss performance analysis from bffm. Therefore in a network with large buffer size safety margin α can be chosen to be smaller than that given in Eq. 4.9, or even zero, to achieve higher network utilization, while still able to satisfy the QoS requirements of connections. Our simulation shows that for a fixed buffer size, controlling the safety margin α can control the cell loss

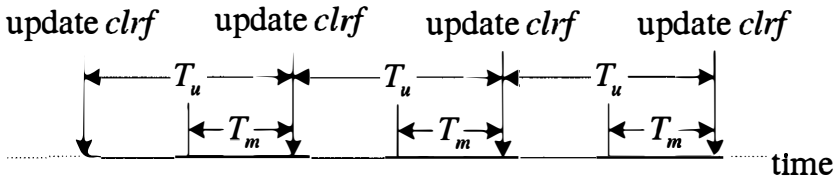


Figure 4.2: Relationship between update interval and measurement interval

ratio, as well as utilization.

Therefore the introduction of α brings some flexibility into the CAC scheme which enables us to efficiently utilize network resources. This characteristic will be investigated in Chapter 6 to design a closed-loop CAC. For the above reasons we use Eq. 4.8 to estimate p in our CAC scheme instead of Eq. 4.4. The clrf of $\sum_{i=1}^{m} Y_i$ is updated using Eq. 4.2 and Eq. 4.3 every T_u second. Fig. 4.2 shows the relationship between the update period T_u and measurement period T_m. The update period T_u is chosen empirically. We suggest choosing a T_u in the range $2T_m \sim 10T_m$, depending on the network state. When call level dynamics is fast, i.e., connections enter and leave the network very often, T_u should be chosen close to $2T_m$. On the other hand, for a network where call level dynamics is slow, T_u should be chosen close to $10T_m$.

Call level dynamics affects the measurements. If a new connection request is admitted in the measurement window, the new connection possibly generates traffic only during part of the measurement window, or does not generate traffic at all. Therefore the new connection will make the measured mean cell rate of the link smaller than its actual value, which will affect the robustness of the CAC scheme with regard to QoS guarantees. To solve this problem, when the measurement window size T_m is much smaller than the connection setup time, we delay the admission of new connection during the measurement window. The admission will be delayed till the end of the measurement window. In the worst case, this will introduce a delay of T_m to the connection setup time. This method is used in our simulation. Alternatively, if the delay caused by T_m to the connection setup time becomes a concern, one can add the sum of declared mean cell rates, if available in the traffic descriptor, or the sum of declared peak cell rates of the connections

admitted in the measurement window, divided by m, to the estimated p in Eq. 4.8. In this case, we do not need to delay the admission of new connections in the measurement window.

It is possible that during the measurement window some existing connections are released, thus affecting the measured mean cell rate. Departing connections contribute to the measured mean cell rate of the link. However, PCR is the sum of peak cell rates of connections on the link at the instant of updating clrf, not including peak cell rates of the departing connections. So if there are some connections that are released during the measurement interval they will make the estimated p larger, which in turn makes the CAC scheme more conservative, assuring that QoS guarantees are not affected. Another alternative is to update the clrf only when no existing connections depart during the measurement interval. The reason that we do not adopt this approach in our CAC scheme is that it may result in the clrf not being updated for a long time, thus affecting the performance of the CAC scheme.

In our CAC scheme, we do not update the clrf for departing connections. The changes in traffic parameters due to departing connections are caught up by updating the clrf at the end of each measurement period.

4.1.2 Cell Loss Ratio Estimation

We shall now show the method of estimating the cell loss ratio when a new connection request arrives. Suppose there are M connections, denoted by $Z_1, \ldots, Z_M$, admitted into the network since the last measurement interval. Denote by $X_1, \ldots, X_n$ the connections existing in the network at the instant when the last measurement interval is finished, and denote by F the clrf of the corresponding upper bound of $X_1, \ldots, X_n, \sum_{i=1}^{m} Y_i$, which is calculated using Eq. 4.1, Eq. 4.2, Eq. 4.3 and Eq. 4.8. When a new connection request, denoted by Z_{M+1}, arrives, if $PCR + \sum_{i=1}^{M+1} pcr_{Z_i} \leq C$, where $pcr_{Z_{i(i=1,\ldots,M+1)}}$ is the declared peak cell rate of Z_i, the new connection can be admitted directly and no cell loss will occur. Otherwise, there are three methods to estimate the cell loss ratio if the new connection is accepted and to determine whether the new connection should be accepted:

1. Assume the new connection Z_{M+1} generates cells constantly at peak cell rate, as well as $Z_1, \ldots, Z_M$. Then $\sum_{i=1}^{M+1} Z_i$ can be regarded as an on-off source which generates cells at peak cell rate $\sum_{i=1}^{M+1} pcr_{Z_i}$ with probability 1, which is, with respect to the convex ordering, smaller than an on-off source λ with peak cell rate $pcr_\lambda = \left\lceil \sum_{i=1}^{M+1} pcr_{Z_i} \right\rceil$ and activity parameter

$$p_\lambda = \frac{\sum_{i=1}^{M+1} pcr_{Z_i}}{\left\lceil \sum_{i=1}^{M+1} pcr_{Z_i} \right\rceil}.$$

From Property 1, the clrf of $\sum_{i=1}^{m} Y_i + \lambda$, denoted by FG can be calculated as:

$$FG(y) = F * g(y),$$

where g is the pmf of λ. Thus if the new connection Z_{M+1} is admitted, the cell loss ratio of the aggregate traffic is upper bounded by:

$$clr = \frac{FG(C)}{FG(0)} = \frac{F(C - pcr_\lambda) \times p_\lambda + F(C) \times (1 - p_\lambda)}{F(0) + \sum_{i=1}^{M+1} pcr_{Z_i}}. \tag{4.10}$$

If the estimated cell loss ratio upper bound is less than the cell loss ratio objective then the new connection is admitted; otherwise it is rejected.

2. Denote the declared mean cell rates of connections $Z_1, \ldots, Z_M, Z_{M+1}$ by $mcr_{Z_1}, \ldots, mcr_{Z_M}, mcr_{Z_{M+1}}$. Since $\sum_{i=1}^{M+1} Z_i$ is, with respect to the convex ordering, smaller than an on-off source λ with peak cell rate $pcr_\lambda = \left\lceil \sum_{i=1}^{M+1} pcr_{Z_i} \right\rceil$ and activity parameter

$$p_\lambda = \frac{\sum_{i=1}^{M+1} mcr_i}{\left\lceil \sum_{i=1}^{M+1} pcr_{Z_i} \right\rceil},$$

then the cell loss ratio, if the new connection Z_{M+1} is admitted, is upper bounded by:

$$clr = \frac{(1 - p_\lambda)\, F(C) + p_\lambda F\,(C - pcr_\lambda)}{F(0) + \sum_{i=1}^{M+1} mcr_i}. \tag{4.11}$$

If the estimated cell loss ratio upper bound is less than the cell loss ratio objective then the connection will be accepted; otherwise the connection is rejected.

3. In the previous two methods we do not need to update the clrf after the admission of a new connection. This results in great computational savings, but as a penalty, the CLR estimations are conservative. We shall now introduce another method. Denote the updated clrf after the admission of Z_M by F_M. The procedure of updating clrf is described later. Denote the peak cell rate and mean cell rate of the new connection request Z_{M+1} by $pcr_{Z_{M+1}}$ and $mcr_{Z_{M+1}}$. Z_{M+1} is smaller than, with respect to the convex ordering, an on-off source λ with a peak cell rate 1 and mean cell rate $mcr_{Z_{M+1}}$. The cell loss ratio, if the new connection is admitted, is estimated as:

$$clr = \frac{(1 - mcr_{Z_{M+1}})F_M(C) + mcr_{Z_{M+1}}F_M(C-1)}{F_M(0) + mcr_{Z_{M+1}}}. \tag{4.12}$$

If the estimated cell loss ratio is less than the cell loss ratio objective then the connection is admitted; otherwise the connection is rejected. If the connection is admitted, clrf will be updated:

$$F_{M+1}(k) = \begin{cases} F_M(0) + mcr_{Z_{M+1}}; & k = 0 \\ (1 - mcr_{Z_{M+1}})F_M(k) + mcr_{Z_{M+1}}F_M(k-1); & k \geqslant 1 \end{cases} \tag{4.13}$$

For the special case of $M = 0$, the clrf $F_0(y)$ is actually $F(y)$, the clrf of $\sum_{i=1}^{m} Y_i$.

Alternatively, one can also take $\widehat{r} + \sum_{i=1}^{M+1} mcr_{Z_i}$ as an estimate of the sum of mean cell rates of all connections in the network if the new connection request Z_{M+1} is admitted; and calculate the clrf $F_{M+1}(y)$ using Eq. 4.1, Eq. 4.2, Eq. 4.3 and Eq. 4.8. Accordingly, in Eq. 4.1, PCR now means the sum of peak cell rates of all connections in the network if connection request Z_{M+1} is admitted.

Updating of the clrf using Eq. 4.13 is computationally much more efficient. However, since Eq. 4.13 actually takes the peak cell rate of $Z_1, \ldots, Z_M, Z_{M+1}$ as 1, it will generate more conservative results.

Comparing the three methods, the advantage of method 1 is obvious as the traffic source only needs to specify its peak cell rate. However, since it assumes the new connection generates cells constantly at peak cell rate, it results in more call

rejections and lower utilization, especially when the traffic source is very bursty and/or call level dynamics is fast. This is verified by our simulation. However when the traffic source is smooth and call level dynamics is slow, method 1 is more attractive due to its simplicity.

In methods 2 and 3, the traffic source is required to specify its mean cell rate. Different from traffic descriptor-based CAC, the impact of inaccuracy of user-specified mean cell rate is limited to one clrf update interval. Hence performance of the CAC scheme will not suffer significantly if the traffic source can not tightly specify its mean cell rate. In method 3, clrf is updated every time when a new connection is admitted, therefore it needs more computation, but method 3 is expected to achieve higher utilization when call level dynamics is fast.

Noting that in the estimation of cell loss ratio only the computation of $F(y)$ and $f(x)$ from 0 to C is needed, we do not need to calculate all values of $F(y)$ and $f(x)$ from 0 to $\lceil PCR \rceil$.

4.2 Cell Loss Ratio of Individual Connections

In the previous sections, we developed a CAC scheme which is able to guarantee the CLR requirement of the aggregate traffic. By appropriately choosing the CLR objective of the aggregate traffic in the CAC scheme, the CLR requirements of individual connections can also be satisfied.

Assume that there are n independent connections $X_1, \ldots, X_n$ on the link. $X_1, \ldots, X_n$ may have any distribution. The cell loss ratio of connection $X_{k(k=1,\ldots,n)}$ can be calculated as:

$$clr_{X_k} = \frac{E\left[\left(\sum_{i=1}^{n} X_i - C\right)^+ \frac{X_k}{\sum_{i=1}^{n} X_i}\right]}{E\left(X_k\right)}.$$

Let $Y_1, \ldots Y_n$ be n independent on-off sources where on-off source $Y_{k(k=1,\ldots,n)}$ has the same pcr and mcr as X_k. It is shown in [24, Theorem 1] that the worst case cell loss ratio for connection X_k is given by

$$clr_{X_k} \leq clr_k \triangleq \frac{E\left[\left(\sum_{i=1}^{n} Y_i - C\right)^+ Y_k\right]}{C \times E\left(Y_k\right)}.$$

Denote the pcr and activity parameter of Y_k by pcr_k and p_k respectively. Using the fact that Y_k is an on-off source, it can be shown that

$$\begin{aligned} clr_k &= \frac{E\left[\left(\sum_{i=1}^{n} Y_i - C\right)^+ Y_k\right]}{C \times E\left(Y_k\right)} \\ &= \frac{E\left[\left(\sum_{i=1, i\neq k}^{n} Y_i + pcr_k - C\right)^+\right] pcr_k \times p_k}{C \times E\left(Y_k\right)} \\ &= \frac{E\left[\left(\sum_{i=1, i\neq k}^{n} Y_i + pcr_k - C\right)^+\right]}{C} \\ &= \frac{F(C - pcr_k)}{C}, \end{aligned} \tag{4.14}$$

where the function F is the clrf of $\sum_{i=1, i\neq k}^{n} Y_i$. It can be shown that the CLR of the aggregate traffic $\sum_{i=1}^{n} Y_i$, denoted by clr_Y, is related to clr_k and clr_{X_k} by:

$$clr_Y = \frac{p_k \times F\left(C - pcr_k\right) + \left(1 - p_k\right) \times F\left(C\right)}{E\left(\sum_{i=1}^{n} Y_i\right)} \tag{4.15}$$

$$= \frac{p_k \times C \times clr_k + \left(1 - p_k\right) \times F\left(C\right)}{E\left(\sum_{i=1}^{n} Y_i\right)} \tag{4.16}$$

$$\geq \frac{p_k \times C \times clr_{X_k} + \left(1 - p_k\right) \times F\left(C\right)}{E\left(\sum_{i=1}^{n} Y_i\right)}. \tag{4.17}$$

From Eq. 4.17, denote the CLR objective of X_k by $clr_{k,obj}$, if

$$clr_Y \leq \frac{p_k \times C \times clr_{k,obj} + \left(1 - p_k\right) \times F\left(C\right)}{E\left(\sum_{i=1}^{n} Y_i\right)} \tag{4.18}$$

then the CLR requirement of connection X_k can be satisfied. We can further remove the term $\left(1 - p_k\right) \times F\left(C\right)$ from Eq. 4.18, i.e., if

$$clr_Y \leq \frac{p_k \times C}{E\left(\sum_{i=1}^{n} X_i\right)} \times clr_{k,obj} \tag{4.19}$$

then the CLR QoS requirement of connection X_k can be satisfied. In inequality 4.19, the fact that

$$E\left(\sum_{i=1}^{n} X_i\right) = E\left(\sum_{i=1}^{n} Y_i\right)$$

is used. Eq. 4.19 gives more conservative CLR objective for the aggregate traffic than that given by Eq. 4.18, however Eq. 4.19 is much easier to implement practically. Since $F(C) < C \times clr_{k,obj}$, when p_k is not a very small value, the CLR objective for the aggregate traffic given by Eq. 4.19 is close to that given by Eq. 4.18.

Therefore, if the CLR objective of the aggregate traffic is chosen to be

$$\frac{p_k \times C}{E\left(\sum_{i=1}^{n} X_i\right)} \times clr_{k,obj},$$

the proposed CAC scheme is able to guarantee the CLR requirement of connection X_k. And if the CLR objective of the aggregate traffic is chosen to be

$$\inf_{1 \le k \le n} \left(\frac{p_k \times C}{E\left(\sum_{i=1}^{n} X_i\right)} \times clr_{k,obj} \right), \tag{4.20}$$

the CLR QoS requirements of all connections can be satisfied.

In summary, when the traffic source is modeled by an on-off source, if the cell loss ratio of the aggregate traffic as evaluated from on-off source model is below the threshold given in Eq. 4.20, the CLR requirements of all connections on the link will be satisfied. In real applications, the term $\frac{C}{E\left(\sum_{i=1}^{n} X_i\right)}$ in Eq. 4.20 can be approximated by 1 to simplify the implementation. p_k is the ratio of the peak rate to the mean rate of X_k.

In the rest of this thesis, we limit our discussion to the loss performance of the aggregate traffic.

4.3 Simulation Study

In this section, we study the performance of our CAC scheme using method 3 as introduced in the last section. Eq. 4.13 is used to update the value of clrf in method 3. The aim of simulation study is to evaluate the performance of our CAC scheme with respect to link utilization and its effectiveness in terms of its ability to guarantee the QoS required by the connections.

The simulation is carried out using OPNET. The following parameters are used for our simulation unless otherwise specified: cell loss ratio objective is set

to be 10^{-4}; switching speed of the ATM switch is set to be infinity, hence every incoming cell is placed immediately in the output buffer; the output buffer size is set to be 20 cells to absorb cell level traffic fluctuations [14], [117]. These parameters are only used as a test on the performance of our CAC scheme. The link utilization and cell loss ratio are observed in our simulation. Link utilization is calculated as the ratio of instantaneous link traffic rate to link capacity. Cell loss ratio is calculated as the ratio of the total observed cell loss to the total cells offered to the link in a moving window with size T_c. More specifically, cell loss ratio at time t is the ratio of the number of cell loss occurred in the interval $(t - T_c, t]$ to the total number of cells offered to the link for transmission in the same interval, where T_c equals $500s$. The utilization data is collected every $0.1s$, so it is actually the average utilization over a $0.1s$ interval. Cell loss ratio data is collected every $1s$. In each scenario, there are several types of traffic sources multiplexed on the link. The connection arrival process of each type of traffic is a Poisson process with a mean of λ calls per second. The connection holding time for all traffic types is exponentially distributed with a mean of 100 seconds.

4.3.1 Simulation Model

In this section exponential on-off source model is used in the simulation. The duration of the on and off periods are independent and exponentially distributed with means β and γ respectively. During on periods, cells are generated at peak cell rate. During off periods no cells are generated. We define the burstiness of a traffic source as:

$$Burstiness = \frac{pcr}{mcr} = \frac{\beta + \gamma}{\beta}. \tag{4.21}$$

Furthermore, the following parameters are used for our simulation: the link capacity is set to be $10Mb/s$, and the measurement window size is chosen to be 0.08 second. In the simulation one cell unit time is

$$\frac{53 \times 8}{10 \times 10^{+6}} = 42.4\mu s.$$

As introduced in section 4.1.1, δ is chosen to be 0.35. Therefore using Eq. 4.9, the safety margin can be determined:

$$\alpha = 1.65 \times \left(\frac{0.08}{42.4 \times 10^{-6} \times 100} \right)^{\frac{-0.35}{2}} = 1.0.$$

Clrf update period is chosen to be 0.2 second. In the simulation, three types of traffic sources are multiplexed on the link. Traffic rate unit u is set to be $100kb/s$, which is the maximum pcr of the three traffic types.

Two scenarios are considered in this section. In the first scenario, referred to as the saturation scenario, the call arrival rate is chosen to be very high. The high call arrival rate means that the system is continually receiving new connection requests. Thus the CAC scheme is expected to achieve the maximum utilization in the saturation scenario. This scenario is used to establish the performance of our CAC scheme with regard to QoS guarantees, because if calls are offered at a very high rate, the rate at which calls are admitted in error becomes very large too [16]. In the second scenario, referred to as the moderate scenario, the call arrival rate of each traffic type is carefully chosen to make the *call blocking ratio* fall between $0 - 0.03$. The *call blocking ratio* is defined as the ratio of the number of calls rejected to the total number of call arrivals. Since in real networks, it is unlikely that calls can arrive at such high rates as those in the saturation scenario, the utilization achieved in the moderate scenario can better represent the utilization that can be achieved by our CAC scheme in real networks.

4.3.2 Saturation Scenario

The parameters of the three traffic types of the saturation scenario are listed in Table 4.1.

The mean on time of the three traffic types shown in Table 4.1 implies that traffic type 1 has a mean burst length of 100 cells, traffic type 2 has a mean burst length of 50 cells and traffic type 3 has a mean burst length of 20 cells. The mean burst length is several times larger than the buffer size. This is used as a trial to establish the performance of the CAC scheme using on-line measurement.

Table 4.1: Parameters of the three traffic types in saturation scenario

	$\lambda\,(s^{-1})$	$pcr\,(kb/s)$	$burstiness$	$\beta\,(s)$	$\gamma\,(s)$
type 1	10	100	10	0.424	3.816
type 2	50	50	5	0.424	1.696
type 3	100	10	2	0.424	0.424

The simulation was run for $10,000$ seconds. An average utilization of 0.76 is achieved by our CAC scheme. Fig. 4.3 shows the observed utilization during the period $3,000$ to $4,000$ seconds. Fig. 4.4 shows the observed cell loss ratio. Fig. 4.5 shows the number of each connection type on the link. Fig. 4.6 shows the admissible region for the three types of calls as well as the number of calls actually admitted by the CAC scheme. The numbers of admitted calls are close to the boundary of the admissible region whilst still within the admissible region. Therefore, the CAC scheme is robust with regard to QoS guarantee, and is capable of achieving a high network utilization. This is also verified by the observed CLR shown in Fig. 4.4. The observed CLR is within the same order of the CLR objective.

Fig. 4.5 shows that the number of connections varies within a large range, and Fig. 4.3 shows that the aggregate traffic is very bursty. However the fast call level dynamics and burstiness of the aggregate traffic do not affect the robustness of the CAC scheme with regard to QoS guarantee. This indicates that the proposed measurement scheme is able to satisfy the requirements of the CAC scheme on robust QoS guarantees even in the presence of very bursty traffic.

According to Gibbens *et al.* [32] and Guerin *et al.* [33], the effective bandwidth of an on-off source is given by:

$$\widehat{c} = \frac{\eta\beta(1-p) \times pcr - B + \sqrt{[\eta\beta(1-p)pcr - B]^2 + 4B\eta\beta p(1-p)pcr}}{2\eta B(1-p)}.$$

where $\eta = ln(1/clr)$, clr is the cell loss ratio objective, B is the buffer size and β is the mean on time. Define the statistical multiplexing gain as the ratio of the bandwidth required to support the same number of connections in the simulation using equivalent bandwidth approach to the link capacity, which implies the

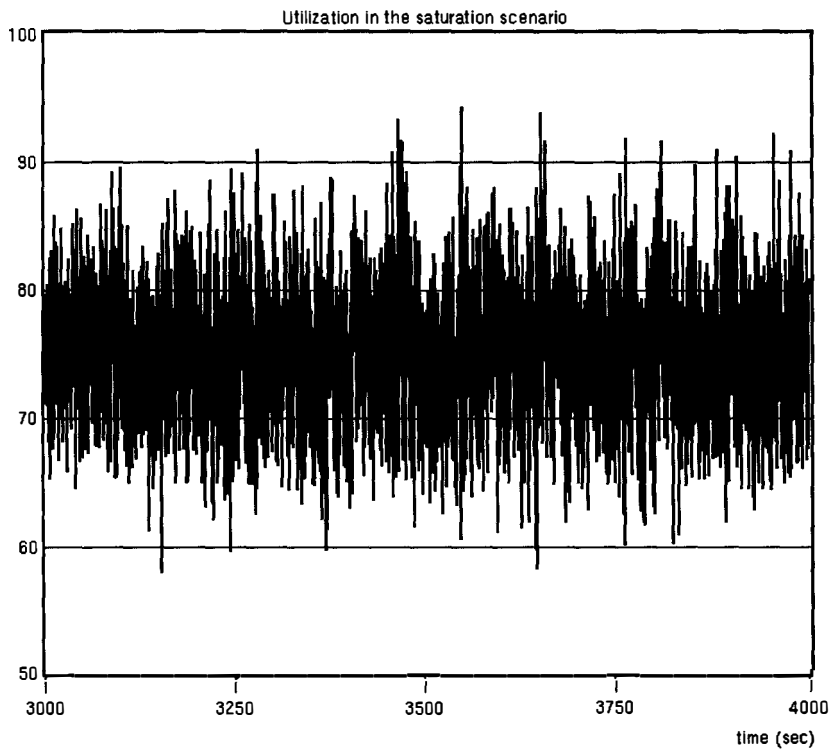


Figure 4.3: Utilization achieved in the saturation scenario

bandwidth required by our CAC scheme. Fig. 4.7 shows the statistical multiplexing gain achieved in the saturation scenario. An average statistical multiplexing gain of 2.7 is achieved.

The saturation scenario is used as the worst case in our simulation study to test the performance of the CAC scheme with regard to robustness on QoS guarantee. Simulation results indicate that the proposed measurement-based CAC scheme is able to provide robust QoS guarantee whilst achieving much higher utilization than that using equivalent bandwidth approach.

4.3.3 Moderate Scenario

In this scenario, we study the case when calls of each type arrive at a moderate rate. The parameters of the three traffic types of the moderate scenario are listed in Table 4.2. Again, the simulation was run for 10, 000 seconds. Fig. 4.8 and Fig.

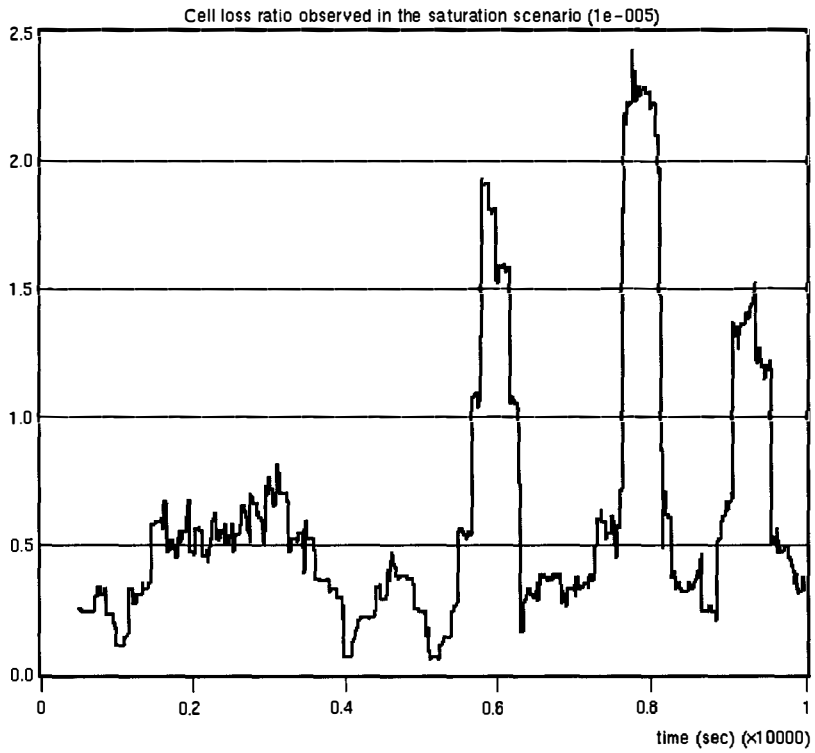


Figure 4.4: Cell loss ratio observed in the saturation scenario

4.9 show the observed cell loss ratio and the number of each traffic type on the link. Fig. 4.10 shows the call blocking ratio of each traffic type on the link. It is shown in Fig. 4.10 that the call blocking ratio of each traffic type is controlled within the desired region of $0 - 0.03$.

Fig. 4.11 shows the statistical multiplexing gain achieved in the moderate scenario. An average statistical multiplexing gain of 2.1 is achieved in the moderate scenario, with an average utilization of 0.68. Compared with the saturation scenario, network utilization decreases by 0.08 and statistical multiplexing gain decreases by 0.6. Cell loss ratio in the moderate scenario also decreases. This is a natural consequence of decrease in utilization. In our measurement-based CAC scheme, the estimated mean traffic rate is larger than its true value in most cases. While such arrangement can ensure robust QoS guarantee, as a penalty, it will inevitably result in false rejection of connections. In the saturation scenario,

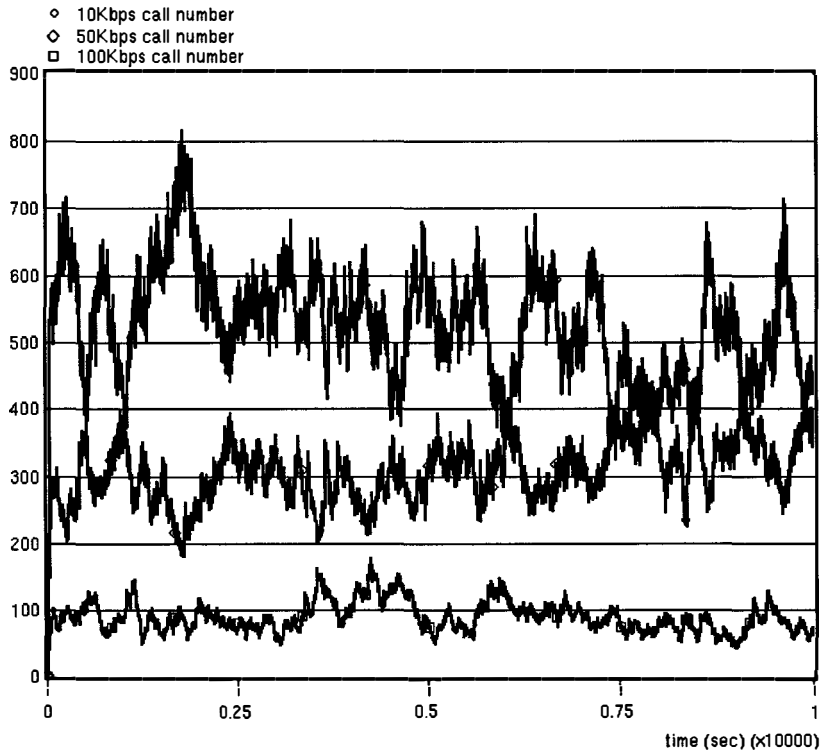


Figure 4.5: The number of the three traffic types on the link observed in the saturation scenario

the false rejections are compensated by high call arrival rates, thus the utilization is unaffected. However, in the moderate scenario, the decrease in utilization is inevitable. The safety margin α can be chosen to be smaller to achieve better utilization, but as a penalty, the robustness of the CAC scheme with regard to QoS guarantees will be decreased in the presence of high call arrival rate.

4.4 Robustness of The CAC Scheme

In section 4.3, traffic parameters specified by the traffic sources are tight and accurate. However, this is impossible in real networks. Therefore, it is essential for a CAC scheme to be robust against inaccuracies in the declared traffic parameters. In this section we study the performance of our CAC scheme when the declared

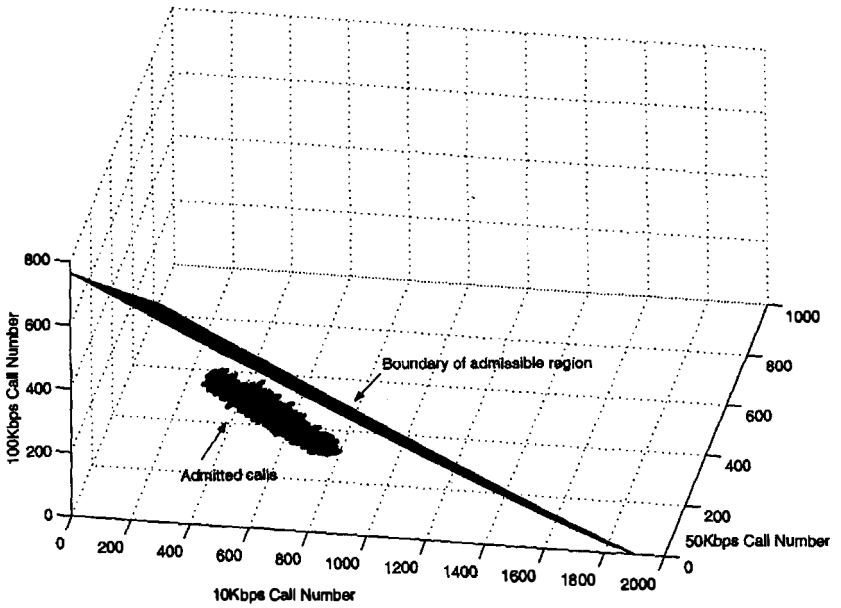


Figure 4.6: Comparison between admissible region and actually admitted calls

Table 4.2: Parameters of the three traffic types in moderate scenario

	$\lambda\,(s^{-1})$	$pcr\,(kb/s)$	$burstiness$	$\beta\,(s)$	$\gamma\,(s)$
type 1	0.202	100	10	0.424	3.816
type 2	0.756	50	5	0.424	1.696
type 3	1.512	10	2	0.424	0.424

traffic parameters are not tight.

Our scheme requires that traffic sources declare their mean cell rates and peak cell rates. The impact of inaccuracies in the declared mean cell rate and peak cell rate is studied separately in this section. The simulation parameters are the same as those in the saturation scenario unless otherwise specified. For ease of comparison, utilization shown in this section is the moving average of observed utilization, average window size is $500s$; cell loss ratio at time t is the ratio of the total number of cell losses occurred in the interval $[0,\ t]$ to the total number of cells offered to the link in the same interval.

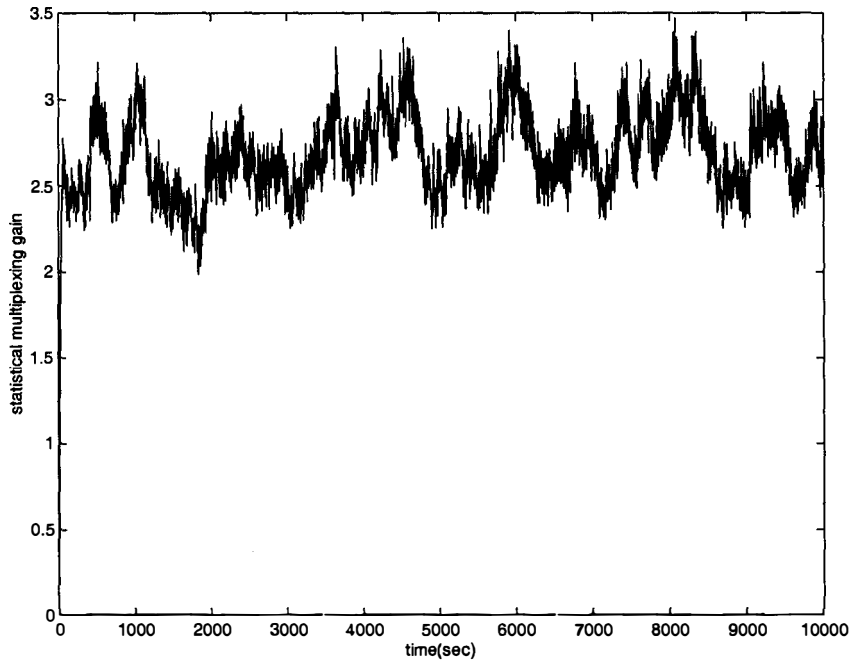


Figure 4.7: Statistical multiplexing gain achieved in the saturation scenario

4.4.1 Impact of Inaccuracies in the Declared Mean Cell Rate

In this subsection we study the impact of inaccuracies in the declared mean cell rates on the performance of the CAC scheme. Two more scenarios are considered in this subsection. In one scenario, referred to as the over-declared mcr scenario, the declared mean cell rates of all three traffic types in the traffic contract are set to be 1.5 times their actual values. In the other scenario, referred to as the under-declared mcr scenario, the declared mean cell rates are set to be 0.5 times their actual values. The performance of the CAC scheme under the saturation scenario, the under-declared mcr scenario and the over-declared mcr scenario is compared.

Fig. 4.12 shows the moving average of the observed utilization in the three scenarios. Fig. 4.13 shows the cell loss ratio. Comparing the utilization and the cell loss ratio in the three scenarios, it is observed that varying the declared mcr in such a large range only results in very small variations in utilization, say, less than 0.01, and slight variations in cell loss ratio. Simulation results show that our CAC scheme is robust against inaccuracies in the declared mcr. This result

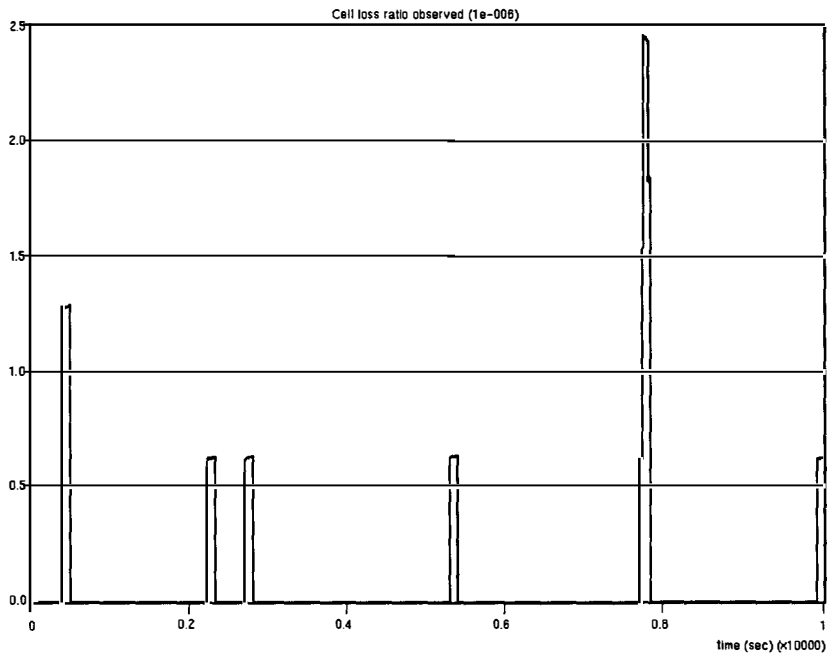


Figure 4.8: Cell loss ratio observed in the moderate scenario

is not unexpected, because inaccuracies in the declared mcr only have "localized effects" on the performance of the CAC scheme, that is, its impact is limited to one clrf update interval following which the declared mcr will be replaced by the measured value.

4.4.2 Impact of Inaccuracies in the Declared Peak Cell Rate

In real networks, it is reasonable to assume that the declared pcr is tight. However, it is still possible that the actual pcr varies within a relatively small range around the declared pcr. In this subsection we study the impact of inaccuracies in the declared pcr on the performance of the CAC scheme. In contrast to the mean cell rate for which inaccuracies in the declared value only have localized effects, the inaccuracies in the declared pcr have long term effects on the performance of our CAC scheme and will persist for the duration of the connection. Hence it is very important that the CAC scheme is robust against inaccuracies in the declared pcr.

There are two new simulation scenarios that we would like to consider here.

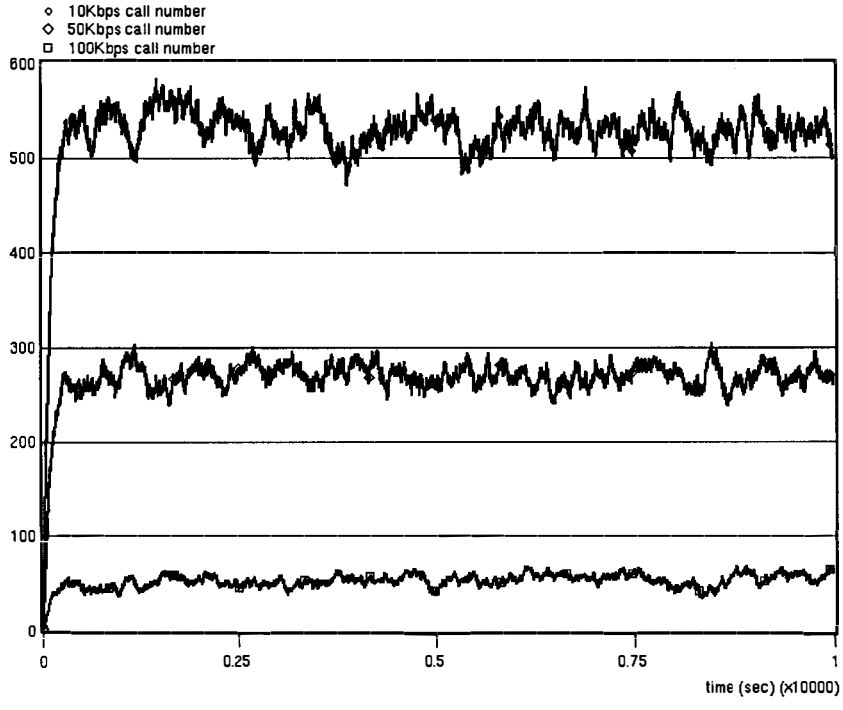


Figure 4.9: Number of the three traffic types on the link observed in the moderate scenario

In the first scenario, referred to as the over-declared pcr scenario, the declared pcrs of all traffic types in the traffic contract are set to be 2 times their actual values. In the second scenario, referred to as the under-declared pcr scenario, the declared pcrs are set to be 0.75 times the actual value. The performance of the CAC scheme in the saturation scenario, the under-declared pcr scenario and the over-declared pcr scenario is compared.

Fig. 4.14 shows the moving average of the observed utilization for the three scenarios. Fig. 4.15 shows the observed cell loss ratio. In the under-declared pcr scenario the declared peak cell rates are decreased to 75 percent of their actual values. This results in an increase of 0.01 in utilization and a slight increase in cell loss ratio. In the over-declared pcr scenario the declared peak cell rates are increased to 2 times the actual values. As a result, utilization decreases by 0.02 and cell loss ratio decreases to half of its value in the saturation scenario. These results are very encouraging, because although the declared pcr is varied in such a

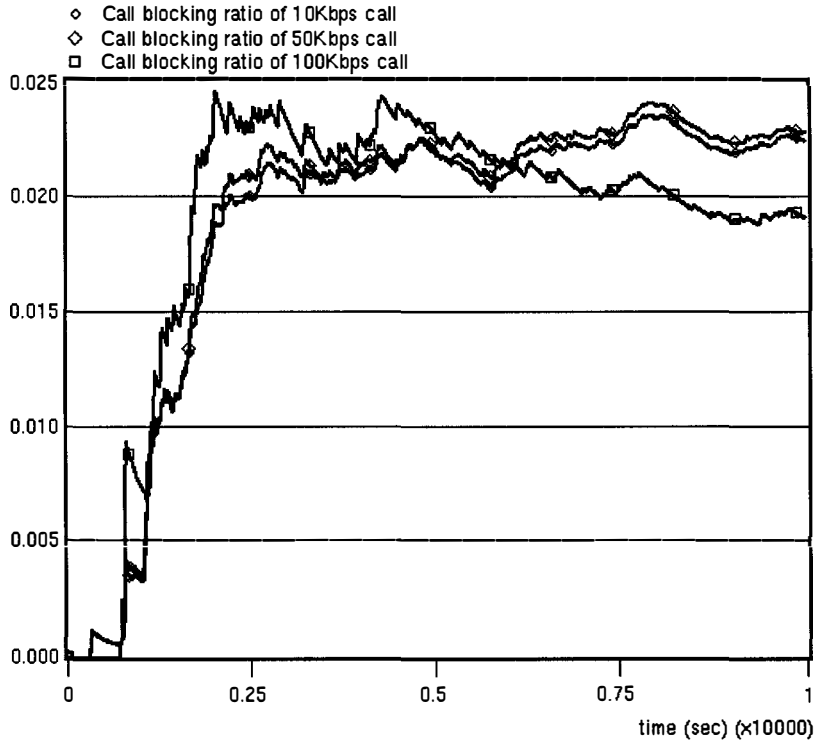


Figure 4.10: Call blocking ratios of the three traffic types in the moderate scenario

large range, the link utilization and cell loss ratio do not suffer significantly. These results indicate that the CAC scheme achieves very good performance even when the declared peak cell rate is very inaccurate.

We explain the robustness of the CAC scheme against inaccuracies in the declared peak cell rate as follows: over-specifying the peak cell rate will usually decrease the network utilization, but on other hand, it also causes the estimated activity parameter p in our CAC scheme to drop by the same percentage, which leads to an increase in utilization. These two effects offset each other. The combination of these two effects makes link utilization less sensitive to the changes in the declared pcr. When traffic sources under-specify their peak cell rates the reverse process occurs. It is this mechanism which makes our CAC scheme robust against inaccuracies in the declared pcr.

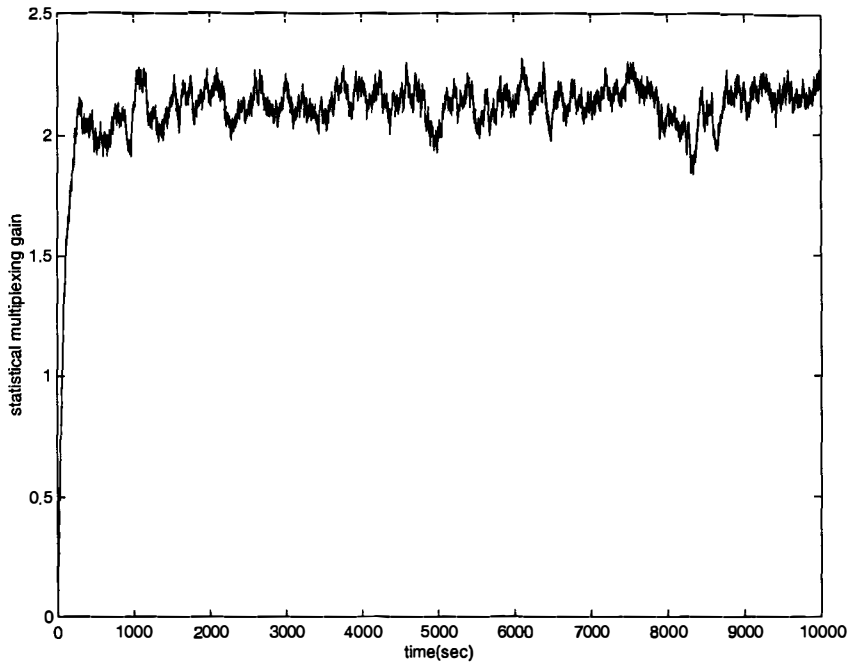


Figure 4.11: Statistical multiplexing gain achieved in the moderate scenario

Simulation results and analysis presented in this section showed that the proposed CAC scheme is robust against inaccuracies in declared traffic parameters. This feature is very attractive for real applications as it relieves the pressure on traffic sources to tightly characterize their traffic parameters. In the simulation, we assume that all traffic sources under-specify or over-specify their traffic parameters. This assumption is only used to establish the performance of the CAC scheme against inaccuracies in declared traffic parameters. In real networks the inaccuracies in declared traffic parameter will possibly offset each other, hence make the sum of the declared traffic parameters more accurate. Therefore, in real applications, the impact of inaccuracies in declared traffic parameters will possibly be even less severe than that presented in this section.

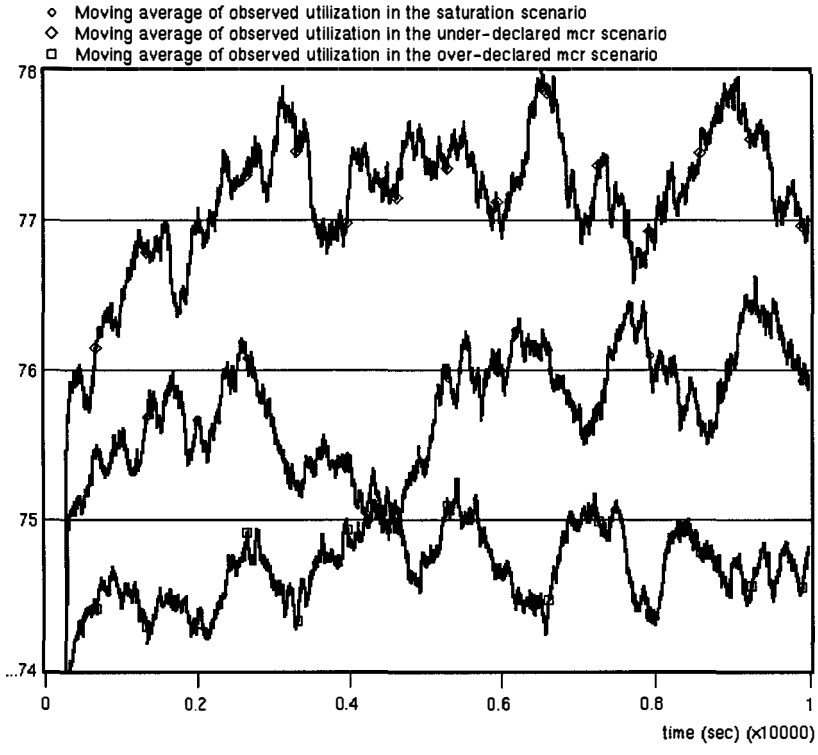


Figure 4.12: Impact of inaccuracies in declared mcr on utilization

4.5 Application of The CAC Scheme to Real Traffic Sources

In previous sections we studied the performance of our CAC scheme using the exponential on-off source model. In this section, we will further study the performance of our CAC scheme using variable bit rate video sources. Eight Motion-JPEG (M-JPEG) encoded movies are used in the simulation. Accordingly, there are eight traffic types. Connections of each type have an exponentially distributed duration with a mean of $100s$, and connection arrival process is a Poisson process with a mean of $1call/s$. When a connection is admitted, the simulation program starts reading the corresponding M-JPEG encoded movie file from the beginning and generates traffic according to the movie file. The statistics of the M-JPEG encoded movies are shown in Table 4.3. The frame rate of the M-JPEG encoded

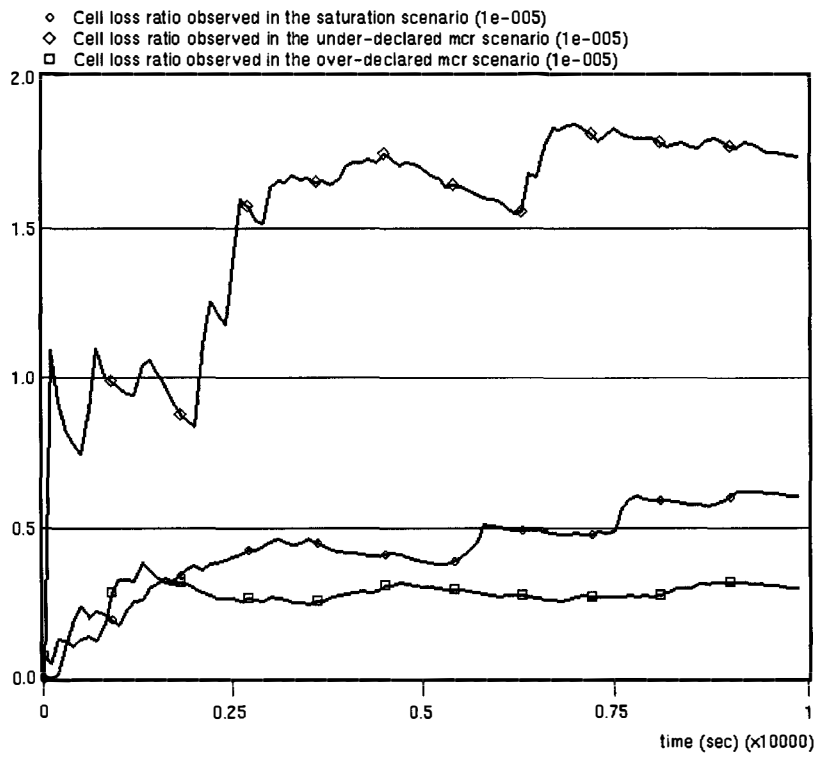


Figure 4.13: Impact of inaccuracies in declared mcr on cell loss ratio

movies is $30 frames/s$. Details about the M-JPEG encoded movies can be found in [118]. Traffic rate unit u is chosen to be $7.288 Mbps$, which is the maximum peak cell rate of the movie sources. $OC3$ link is used in the simulation. The measurement window size is chosen to be 0.10 second. Clrf update period is chosen to be 0.2 second. Using the method introduced in section 4.1.1, safety margin α is chosen to be 0.5.

The simulation was run for $6,000s$. Fig. 4.16 shows the cell loss ratio observed in this scenario. Fig. 4.17 and Fig. 4.18 show the number of calls of each type multiplexed on the link.

An average network utilization of 0.65 is achieved in this scenario while CLR QoS is guaranteed. The utilization achieved is smaller that that achieved using the exponential on-off sources. There are mainly two factors affecting the utilization. First, the peak cell rate of the traffic sources are close to 10 percent of the link

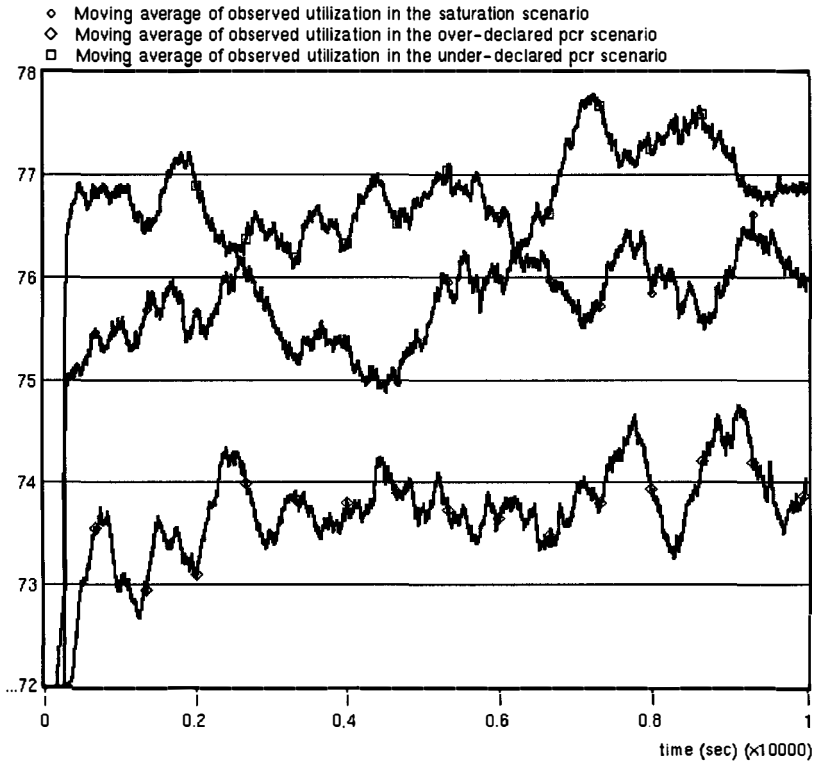


Figure 4.14: Impact of inaccuracies in declared pcr on utilization

rate. As a rule of thumb, when the peak cell rate of the traffic sources are close to 10 percent of the link rate, the statistical multiplexing gain which can be achieved is very small. Second, on-off sources considered in the CAC scheme are the worst case among all traffic sources with the same peak cell rate and mean cell rate. In this scenario, each video source is modeled by an on-off source, which is actually the worst case of the real source. This will also result in lower utilization.

It is worth noting that both the individual variable bit rate (vbr) video sources and the aggregate traffic in this simulation scenario presents significant self-similar behavior. As an example, Fig. 4.19 presents the variance time plot of "the beauty and the beast" vbr video sources used in the simulation (refer to section 2.3.2 for details about variance time plot). This video source shows significant self-similar behavior with a Hurst parameter of about 0.8 [72]. Fig. 4.20 shows the variance time plot of the aggregate traffic rate. The aggregate traffic also presents

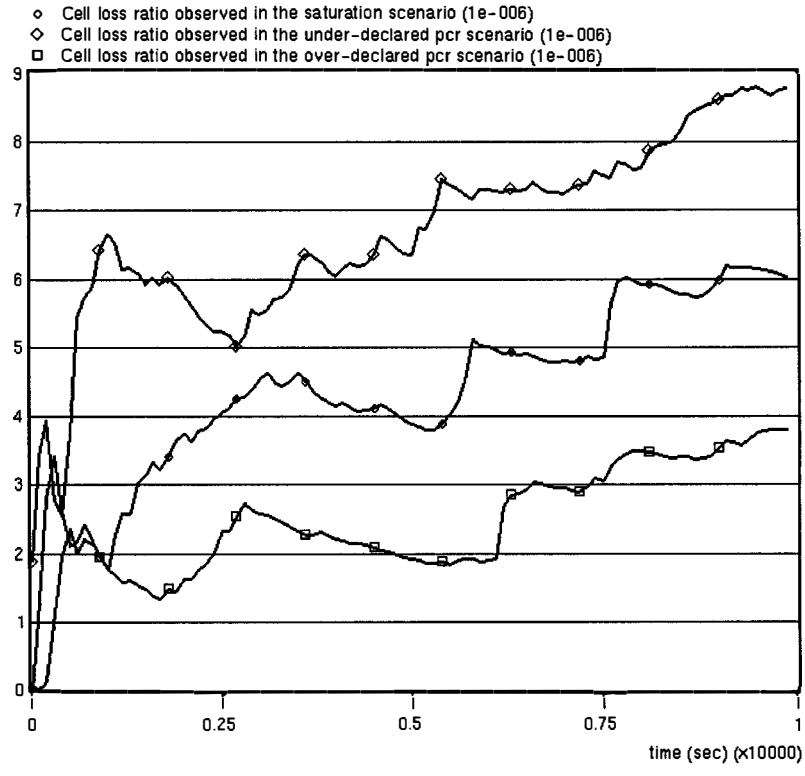


Figure 4.15: Impact of inaccuracies in declared pcr on cell loss ratio

significant self-similar behavior with a *Hurst* parameter of about 0.7. However as shown in Fig. 4.16 the long-range dependence in the network traffic does not threaten the performance of our CAC scheme, our CAC scheme is able to provide robust QoS guarantee even for long-range dependent traffic. Therefore the proposed measurement-based CAC scheme can be applied to self-similar traffic.

Self-similarity, or called long-range dependence, is one of the most difficult problems puzzling the CAC research. As introduced in Chapter 3, the loss performance analysis presented there is a steady-state analysis, thus the proposed upper bound constitutes the worst case in terms of average loss performance. Theoretically speaking, the average loss performance constraints may not be meaningful, if the aggregate traffic exhibits long-range dependence. Specifically, if the aggregate traffic exhibits long-range dependence although average performance may be deemed to be fine, there may be rare periods of time in which performance is

Table 4.3: Traffic rate of the M-JPEG encoded movies ($bytes/frame$)

Type	Name	peak rate	mean rate
1	Sleepless in Seattle	16617	9477.6
2	Crocodile Dundee	19439	10772.9
3	Home Alone, II	22009	11382.8
4	Jurassic Park	23883	11363.0
5	Rookie of the Year	27877	12434.9
6	Speed	29385	12374.4
7	Hot Shots, Part Duex	29933	12766.1
8	Beauty and the Beast	30367	12661.5

consistently poor.

However, there is considerable debate about the impact of long-range dependent traffic on bandwidth allocation and network performance [76], [96], [97], [17], especially the impact of long-range dependent traffic on the performance of a measurement-based CAC scheme [59], [15], because the time scale of interest in real applications is limited (refer to section 2.3.3 for details).

For measurement-based CAC, Grossglauser *et al.* [59], [15] identify a critical time-scale $\widetilde{T_h}$ such that aggregate traffic fluctuations slower than $\widetilde{T_h}$ can be tracked by the admission controller and compensated for by connection admissions and departures. Using Gaussian aggregate traffic model and heavy traffic approximations, the critical time scale is shown to scale as $T_h/\sqrt{n}$, where T_h is the average connection duration and n is the size of the link in terms of number of flows it can carry.

We consider that for measurement-based CAC such a *critical time scale* concept can be extended to all kinds of network traffic, whether or not the aggregate traffic is Gaussian. Let us consider the following example. In our simulation, in which mean call holding time is $100s$, we can assert that traffic fluctuations with time-scales larger than $100s$ can not affect the performance of our measurement-based CAC. Because by the time traffic fluctuation with a time scale larger than $100s$ rising from its lowest point to its highest point, half of the calls have left

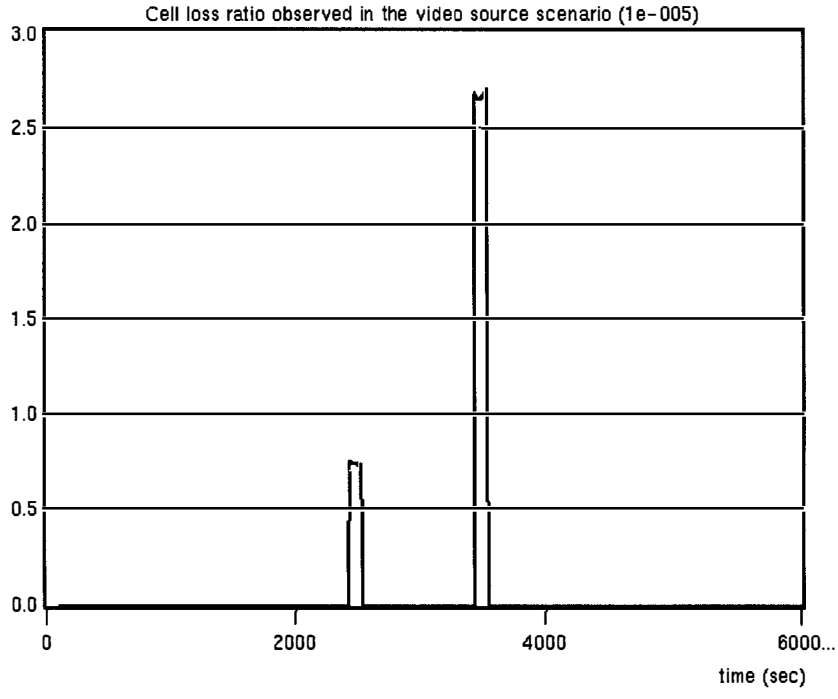
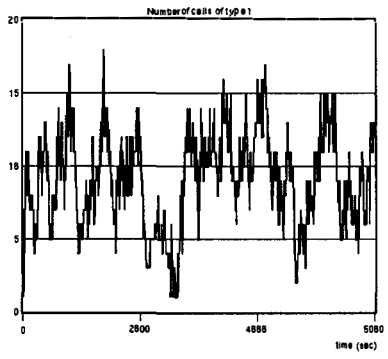


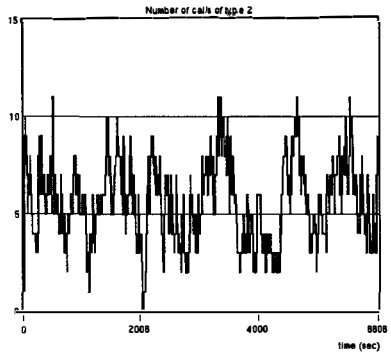
Figure 4.16: Cell loss ratio observed in the video source scenario

the link. The CAC controller has enough time to accommodate the bandwidth requirement of the traffic fluctuation with a time-scale larger than $100s$ by admitting less calls. This is the advantage of measurement-based CAC - it can adapt to the traffic environment. Following this analogy, no matter whether or not the aggregate traffic is Gaussian there must exists a critical time-scale $\widetilde{T_h}$ less than the mean call holding time, traffic fluctuations slower than which have no impact on the performance on a measurement-based CAC. However the exact value of the critical time-scale is still subject to further investigation.

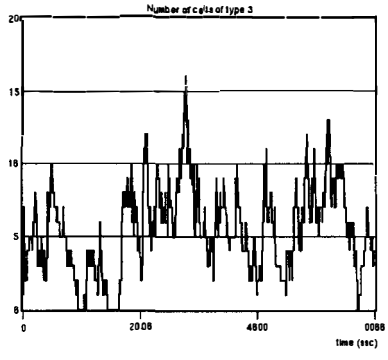
Remark 1 *Such a critical time scale may also exist for traffic-descriptor based CAC. Ryu* et al. *[97] and Grossglauser* et al. *[96] show the existence of critical time scale using numerical examples (refer to section 2.3.3 for details). According to their study, the critical time scale of a traffic-descriptor based CAC is a non-decreasing function of buffer size. Theoretical analysis on loss performance of self-similar traffic in the literature is actually the analysis of cell loss probability*



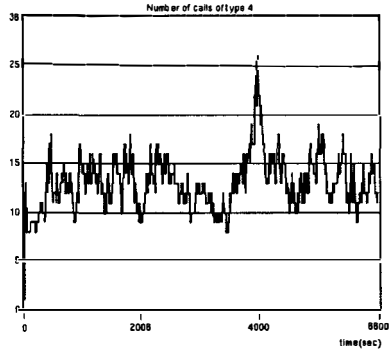
(a) Call number of traffic type 1



(b) Call number of traffic type 2



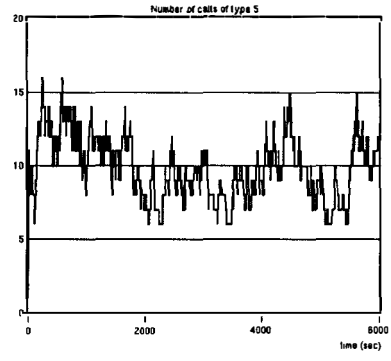
(c) Call number of traffic type 3



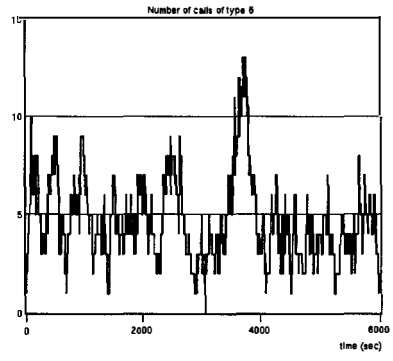
(d) Call number of traffic type 4

Figure 4.17: Number of each type of calls multiplexed on the link in the video source scenario Part I

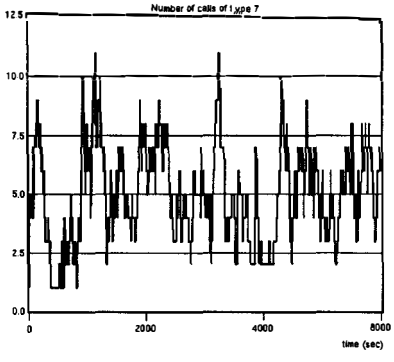
(for example theoretical analysis based on large deviations theory), not cell loss ratio. We have identified these two concepts in section 3.1. Cell loss probability refers to the probability that the buffer content of an infinite buffer queue exceeds a certain threshold, i.e. the buffer size B. It comes from evaluation of infinite buffer queues, and it fails to consider the interaction between the traffic process and the limited buffer size, and the finite range of time scale of practical interest. Recent studies show that the finite range of time scale of practical interest should be considered in performance evaluation and prediction problems [96], [97], and the truncating effect of finite buffers when buffer become full diminishes the effect



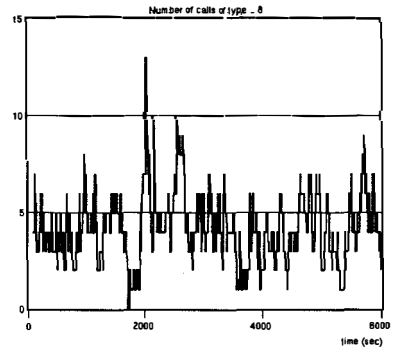
(a) Call number of traffic type 5



(b) Call number of traffic type 6



(c) Call number of traffic type 7



(d) Call number of traffic type 8

Figure 4.18: Number of each types of calls multiplexed on the link in the video source scenario Part II

of long-range dependence [76] (refer to section 2.3.3 for details). These results suggest that the fact that in real applications buffer size is limited, possibly small, should be considered in the performance evaluation of long-range dependent traffic. Failure in considering these effects may compromise some conclusions in the literature about loss performance analysis of self-similar traffic.

Therefore, we consider that there are two possible reasons which make our CAC scheme robust with regard to QoS guarantees for long-range dependent traffic:

1. Our CAC scheme is a measurement-based CAC scheme. So, there exists a

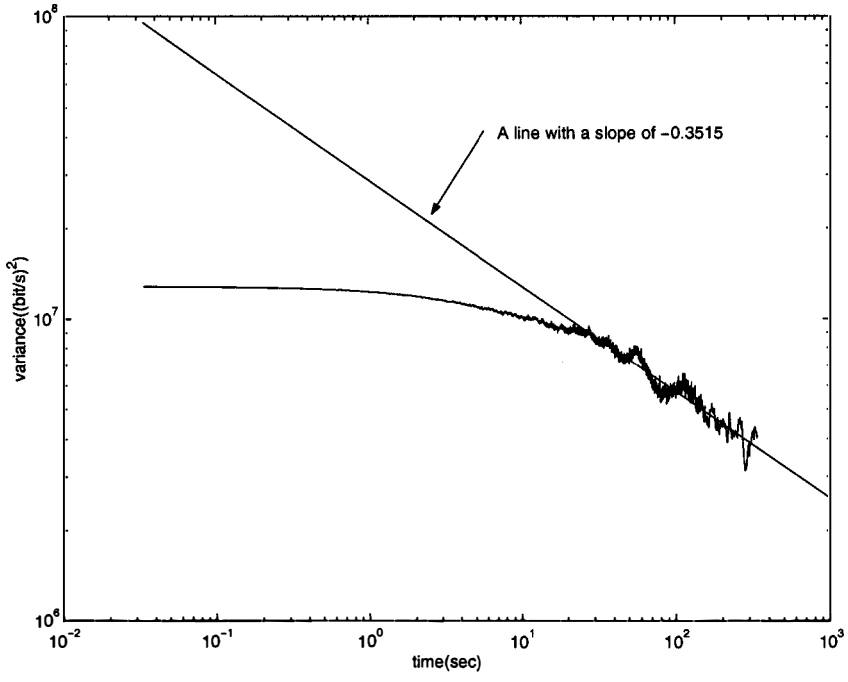


Figure 4.19: Variance time plot of beauty and the beast vbr video source

critical time scale $\widetilde{T_h}$ for our CAC scheme. Only those traffic fluctuations with time scale greater than measurement window size T_m but less than $\widetilde{T_h}$ will threaten the performance of our CAC scheme.

2. Our CAC scheme is based on an upper bound, i.e. it will over-allocate bandwidth. This over-allocated bandwidth helps to effectively absorb traffic fluctuations with time scale greater than the measurement window size T_m but less than $\widetilde{T_h}$.

The impact of the long-range dependent traffic on the performance of our measurement-based CAC is in fact also a problem for almost all measurement-based CAC. Here we attempted to explain the observed phenomenon that long-range dependent traffic does not affect the performance of our CAC scheme. However, further studies are required to clarify this problem.

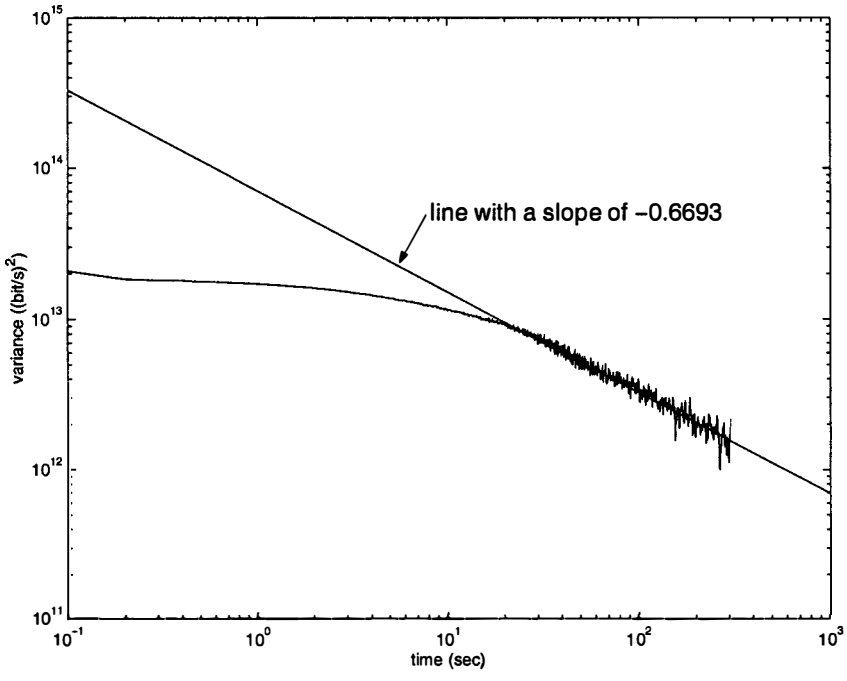


Figure 4.20: Variance time plot of the aggregate traffic rate

4.6 Summary

In this chapter we designed a measurement-based connection admission control scheme. With due consideration to practical applications of a measurement-based CAC scheme, we proposed principles that govern the design of a measurement-based CAC scheme. Robustness, flexibility and simplicity become major objectives in our CAC design. The proposed CAC scheme needs only simple traffic parameters from sources, i.e. peak cell rate and/or mean cell rate. Furthermore, the proposed measurement scheme is easy to implement and only needs measurement of the mean cell rate of the link.

Extensive simulation was carried out to investigate the performance of the proposed CAC scheme in a network with small buffers. Simulation results show that the proposed CAC scheme is able to provide robust QoS guarantees, is capable of achieving a significantly higher utilization than that achieved by effective bandwidth approach, is robust against inaccuracies in user declared traffic parameters,

and can be applied to long-range dependent variable bit rate video sources. All these features make the proposed CAC scheme an attractive option for implementation.

Chapter 5

In-Service QoS Monitoring and Estimation

In ATM networks a prime concern is to ensure that there are adequate resources to meet QoS requirements of traffic sources. In order to make effective traffic and congestion control, and network management, in-service monitoring and estimation (ISME) has to be employed as opposed to the conventional out-of-service monitoring and testing techniques [19]. The ISME can be used to:

- monitor the CLR performance under in-service conditions and verify that the CLR meets the QoS requirements;
- provide CLR information to connection admission control;
- identify the location and causes for the CLR performance degradation without affecting customers;
- conduct preventive, reactive and predictive maintenance by continuously investigating and estimating the performance trend.

Garcá-Hernández *et al.* propose an in-service monitoring method using Operation, Administration and Management (OAM) cells [119]. However one potential problem for ISME is that some QoS indicators are specified in terms of the probability of occurrence of certain rare events. For example, in ATM networks, cell

loss ratio is often specified to be as small as 10^{-9}. Monitoring using direct statistical methods is impractical for such small CLR. For example, in an OC3 link with a link utilization of 0.5 and a cell loss ratio of 10^{-9}, at least 10 billion ATM cells have to be monitored before any statistically meaningful information can be collected. This will take 15 hours. The statistical information obtained after such long monitoring period may be obsolete and the network management system's reaction may be too late.

Zhu *et al.* propose an in-service QoS monitoring and estimation method, which is used to obtain a real-time estimation of CLR [67]. Their method is based on the asymptotic relationship between CLR and buffer size. The observed cell loss ratios of several small virtual buffers are used to estimate the CLR of the actual system (refer to section 2.2.7 for details). However, their method still requires a long monitoring period and the accuracy of their method is not good. In this thesis, based on the same principle used in [67], we will propose an in-service QoS monitoring and estimation scheme which employs simple network measurement and linear regression. The proposed ISME scheme only needs a short monitoring period and is effective for a variety of traffic types. Simulation studies show that the proposed ISME scheme has a much better performance than that proposed in the literature.

5.1 Relationship Between CLR and Buffer Size in ATM Networks

In this section we shall establish the validity of using a generalized relationship between buffer size and the logarithm of cell loss ratio, denoted by $log(clr)$, for ISME. Analytical results in the literature will be reviewed and summarized to demonstrate this relationship for both Markovian traffic and long-range dependent traffic.

5.1.1 Markovian Traffic Process

For Markovian traffic, $log(clr)$ is known to decrease proportionally with increasing buffer size. This linear relationship is a natural outgrowth of the fluid flow analysis of ATM networks [31]. This relationship is also widely used in connection admission control schemes based on effective bandwidth approach for bandwidth allocation.

Based on fluid flow model, Anick *et al.* show that for an infinite queue fed by N homogeneous exponential on-off sources, and drained at a rate c, where c is normalized by the peak cell rate of the on-off source, the probability of overflow beyond buffer size x is given by [31]:

$$G(x) = Pr\{buffer\,content > x\} = 1 - \mathbf{1} \bullet \mathbf{F}(x)$$

where $\mathbf{1}$ is a $1 \times N$ vector of 1s, $\mathbf{F}(x)$ is a $N \times 1$ vector:

$$\mathbf{F}(x) = \begin{bmatrix} F_1(x) \\ \vdots \\ F_N(x) \end{bmatrix}$$

and $F_i(x)$ is the equilibrium probability that i sources are on and buffer content does not exceed x. They show that:

$$\mathbf{F}(x) = \mathbf{F}(\infty) + \sum_{i=0}^{N-\lfloor c \rfloor - 1} \alpha_i \boldsymbol{\phi_i} e^{z_i x}$$

where z_i and $\boldsymbol{\phi_i}$ are the eigenvalues and eigen vectors associated with the solution of the differential equations describing the stationary probabilities of the system, and α_i are coefficients determined by boundary conditions. Using the relationship

$$\mathbf{1} \bullet \mathbf{F}(\infty) = 1$$

$G(x)$ can be obtained:

$$G(x) = -\sum_{i=0}^{N-\lfloor c \rfloor - 1} \alpha_i \left(\mathbf{1} \bullet \boldsymbol{\phi_i}\right) e^{z_i x}. \tag{5.1}$$

In Eq. 5.1, z_0, which is the largest negative eigenvalue, is given by:

$$z_0 = -r = -\frac{(1-\rho)(1+\lambda)}{1-c/N}$$

where ρ is link utilization and λ is the ratio of the average on period to the average off period of the on-off source. Other eigenvalues z_i are negative roots of the following quadratics by setting $k = i$:

$$A(k)z^2 + B(k)z + C(k) = 0, \qquad k = 0, 1, \ldots, N$$

where

$$\begin{aligned} A(k) &= \left(\frac{N}{2-k}\right)^2 - \left(\frac{N}{2-c}\right)^2 \\ B(k) &= 2(1-\lambda)\left(\frac{N}{2-k}\right)^2 - N(1+\lambda)\left(\frac{N}{2-c}\right) \\ C(k) &= -(1+\lambda)^2\left\{\left(\frac{N}{2}\right)^2 - \left(\frac{N}{2-k}\right)^2\right\}. \end{aligned}$$

Therefore, when buffer size x becomes large, buffer overflow probability $G(x)$ can be approximated by the following equation:

$$G(x) \sim -\alpha_0 \left(\mathbf{1} \bullet \boldsymbol{\phi}_0\right) e^{z_0 x}.$$

The value of α_0 and $\mathbf{1} \bullet \boldsymbol{\phi}_0$ can be determined as:

$$\alpha_0 = -\left(\frac{\lambda}{1+\lambda}\right)^N \left\{\prod_{i=1}^{N-\lfloor c \rfloor -1} \frac{z_i}{z_i - z_0}\right\}$$

$$\mathbf{1} \bullet \boldsymbol{\phi}_0 = \left(\frac{N}{c}\right)^N.$$

It is noticed that:

$$\rho = \frac{N}{c} \times \frac{\lambda}{1+\lambda},$$

therefore buffer overflow probability is approximately given by:

$$G(x) \sim \rho^N \left\{\prod_{i=1}^{N-\lfloor c \rfloor -1} \frac{z_i}{z_i - z_0}\right\} e^{z_0 x}. \tag{5.2}$$

Buffer overflow probability is often used as an estimate of the CLR. Thus, Eq. 5.2 indicates an asymptotic linear relationship between buffer size and $log(clr)$.

This linear relationship is shown to hold within a much more general context by Elwalid *et al.* [34], [39]. Based on theoretical analysis using large deviations theory and simulation validation, Elwalid *et al.* propose the following asymptotic relationship between buffer size B and $log(clr)$ for general Markovian traffic:

$$log(clr) \sim -\alpha - \delta B \tag{5.3}$$

where α and δ are both positive constants determined by the traffic process and the link capacity. This generalized result has been used in many situations to develop algorithms for connection admission control and routing in ATM networks [33], [69], [39].

5.1.2 Long-Range Dependent Traffic Model

A lot of research shows convincingly that long-range dependence exists widely in network traffic [72], [77], [73], [79] (refer to section 2.3.1 for details). For long-range dependent traffic process, CLR decays more slowly with buffer size, and the linear relationship between buffer size and $log(clr)$ no longer exists.

Taqqu *et al.* show that the superposition of many on-off sources with strictly alternating on and off periods, and whose on periods or off periods exhibit the *Noah effect* (i.e. high variability or infinite variance) produces an aggregate traffic that exhibits the *Joseph effect* (i.e. self-similar or long-range dependent) [85]. This result reduces the self-similarity phenomenon for the aggregate traffic to properties of the individual traffic components that make up the aggregate traffic stream. Self-similarity in the aggregate traffic can therefore be explained by heavy-tailed on and/or off period distributions of individual sources [77]. They also present extensive statistical analysis of high time-resolution Ethernet LAN traffic traces, which confirms that data at the level of individual sources or source-destination pairs are consistent with the Noah effect. The proposed physical explanation based on the Noah effect suggests the essential difference between self-similar and traditional traffic modeling in the parameter settings of the well known on-off source

models: traditional traffic modeling assumes finite variance distributions for the on and off periods, while self-similar modeling is based on the assumption of the Noah effect, i.e. infinite variance distribution. In another research on heavy-tailed on-off sources by Heath *et al.*, they show independently that on-off sources with heavy-tailed on and/or off distributions lead to strong long-range dependence in the aggregate traffic [84].

Therefore, the multiplexing of on-off sources with heavy-tailed on and/or off distributions results in long-range dependence in the aggregate traffic. Heavy-tailed on-off sources are appropriate traffic models for self-similar traffic.

Likhanov *et al.* consider the superposition of a large number of on-off sources with Pareto distributed (heavy-tailed) active (on) periods [120]. For M i.i.d. on-off sources, the aggregate traffic converges to a self-similar traffic process as $M \rightarrow \infty$. Using queueing theory, they show that the overflow probability of the aggregate traffic has an asymptotic relationship with buffer size B:

$$Pr(buffer\ content > B) \sim \alpha B^{\beta} \tag{5.4}$$

where α and β are constants determined by the traffic process.

Parulekar *et al.* use the $M/G/\infty$ model for self-similar process [121]. The $M/G/\infty$ model is first mentioned by Cox as an example of an asymptotically self-similar process [122]. Using large deviations theory Parulekar *et al.* show that for self-similar traffic process the overflow probability decays hyperbolically with buffer size as that shown in Eq. 5.4.

In addition to the above models for self-similar traffic, Fractional Brownian Noise (FBN) model is also used for self-similar traffic modeling. The FBN model results in the following generalized relationship between overflow probability and buffer size [121], [81]:

$$Pr(buffer\ content > B) \sim \exp\left(-\delta B^{\gamma}\right) \tag{5.5}$$

where δ and γ are constants determined by the traffic process.

It can be argued at this stage that some self-similar traffic models are more appropriate than others. However, as a quick review of the existing literature

indicates, all traffic models have provided good fit for diverse applications. Here we choose Eq. 5.4 as the asymptotic relationship between overflow probability and buffer size. That is, we consider that $log(clr)$ has the following asymptotic relationship with buffer size:

$$log(clr) \sim log(\alpha) + \beta log(B). \tag{5.6}$$

Eq. 5.4 is chosen because in our loss performance analysis in Chapter 3, we used the on-off source model. According to Likhanov *et al.* , for heavy-tailed on-off sources the asymptotic relationship between CLR and buffer size is hyperbolic [120]. Moreover, although Gaussian distribution provides a good fit for the distribution of the aggregate traffic when the number of connections are very large, the CLR estimate resulting from Gaussian distribution is usually too optimistic, especially when the aggregate traffic distribution deviates from Gaussian. The ISME algorithm we are going to design will be used to provide a clr estimate for CAC scheme. We prefer a conservative clr estimate to an optimistic clr estimate so that the CAC scheme is robust with regard to QoS guarantees. Therefore, the Weibull-like relationship in Eq. 5.5, which results from a Gaussian distribution, is not chosen.

It is important to note that the impact of long-range dependence on network performance evaluation and bandwidth allocation is a controversial issue (refer to section 2.3.3 for details). Despite the existing controversy, $log(clr)$ versus buffer size relationship for both Markovian and self-similar traffic sources are taken into account in the development of our ISME algorithm.

5.2 A CLR Estimation Algorithm

Using the relationships between clr and buffer size for both Markovian traffic and self-similar traffic, a CLR estimation algorithm can be designed which only needs a small monitoring period. The basic idea is to use the cell loss ratios observed from several virtual buffers with much smaller buffer sizes than that of the real buffer to estimate CLR of the real system. Since virtual buffers have much

smaller buffer sizes, the cell loss ratios in the virtual buffers are much larger than that of the real buffer. In order to obtain statistical meaningful observations of the cell loss ratios of these virtual buffers, fewer cells need to be observed compared to those required for a large real buffer. Hence the observation of the cell loss ratios of virtual buffers requires much less monitoring period. Then based on the asymptotic relationship between CLR and buffer size for both Markovian traffic and self-similar traffic, the cell loss ratios observed in these virtual buffers are used to obtain an estimate of CLR in the real buffer. Therefore much less monitoring period is required in our algorithm to obtain an estimate of CLR in the real buffer than that using direct monitoring method.

Fig. 5.1 shows the system model using virtual buffers. Four virtual buffers are employed in the algorithm. Denote by B_1, B_2, B_3, B_4 the sizes of the four virtual buffers. B_1, B_2, B_3, B_4 are chosen such that:

$$B_1 < B_2 < B_3 < B_4 << B,$$

where B is the size of the real buffer. Denote by clr_t^1, clr_t^2, clr_t^3 and clr_t^4 the cell loss ratios observed in the four virtual buffers 1, 2, 3 and 4 respectively at discrete time t. These cell loss ratios of the virtual buffers are used to obtain an estimate of the CLR in the real buffer, which is denoted by $\widehat{clr_t}$.

For Markovian traffic, the cell loss ratio has an asymptotic relationship with buffer size as given in Eq. 5.3. Using this relationship, given the cell loss ratio observations in the virtual buffer 1 and virtual buffer 2, the logarithm of cell loss ratio in a buffer of size x can be estimated:

$$\log\left(\widehat{clr_t^x}_M\right) = a + b \times x \tag{5.7}$$

where $\widehat{clr_t^x}_M$ denotes the CLR estimate in a buffer with buffer size x at time t, which is estimated from the CLR-buffer size relationship for Markovian traffic,

$$\begin{aligned} a &= \frac{\log(clr_t^1) \times B_2 - \log(clr_t^2) \times B_1}{B_2 - B_1}. \\ b &= \frac{\log(clr_t^2) - \log(clr_t^1)}{B_2 - B_1}. \end{aligned}$$

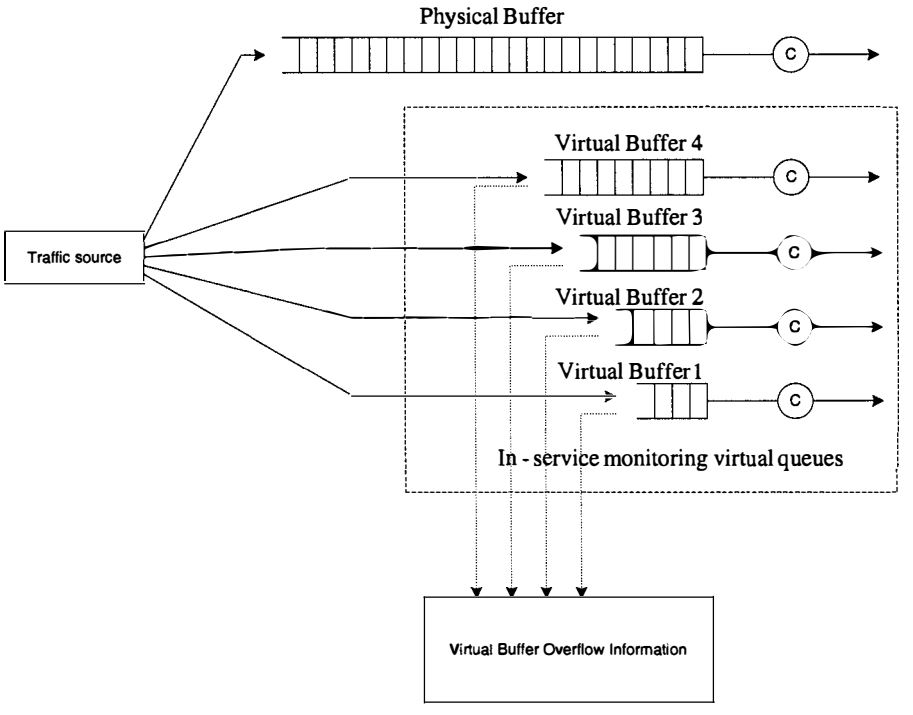


Figure 5.1: System model of the CLR estimation algorithm

The logarithm of cell loss ratio in a buffer with buffer size x can also be estimated using the CLR-buffer size relationship for self-similar traffic given in Eq. 5.6:

$$\log\left(\widehat{clr_t^x}_S\right) = c + d\log(x) \tag{5.8}$$

where $\widehat{clr_t^x}_S$ denotes the CLR in a buffer with buffer size x at time t, which is estimated from the CLR-buffer size relationship for self-similar traffic,

$$\begin{aligned} c &= \frac{\log(clr_t^1)\log(B_2) - \log(clr_t^2)\log(B_1)}{\log(B_2) - \log(B_1)}. \\ d &= \frac{\log(clr_t^2) - \log(clr_t^1)}{\log(B_2) - \log(B_1)}. \end{aligned}$$

In our algorithm, the CLR of the real buffer is estimated using Eq. 5.9:

$$\log\left(\widehat{clr_t}\right) = \alpha \times \log\left(\widehat{clr_t^B}_M\right) + \beta \times \log\left(\widehat{clr_t^B}_S\right), \tag{5.9}$$

where α and β are non-negative constants estimated using the cell loss ratio observations in the virtual buffer 3 and 4.

It is worth noting that due to the bursty nature of network traffic, sometimes cell loss ratios of zero can be observed in these virtual buffers no matter how small the virtual buffer is chosen to be or how large the monitoring period is chosen to be. A cell loss ratio of zero becomes a problem in our CLR estimation algorithm using Eq. 5.9. In our algorithm this problem is solved in the following way: if any of the virtual buffers has a cell loss ratio of zero, the cell loss ratio in the real buffer is estimated to be zero too.

5.2.1 Estimation of Parameters α and β

Based on the CLR-buffer size relationship for Markovian traffic, estimates of the logarithm of cell loss ratios in buffers with size B_3 and B_4 at time t can be obtained using Eq. 5.7. Denote them by $\log\left(\widehat{clr_t^{B_3}}_M\right)$ and $\log\left(\widehat{clr_t^{B_4}}_M\right)$ respectively. Based on the CLR-buffer size relationship for self-similar traffic, estimates of the logarithm of cell loss ratios in buffers with size B_3 and B_4 at time t can be obtained using Eq. 5.8. Denote them by $\log\left(\widehat{clr_t^{B_3}}_S\right)$ and $\log\left(\widehat{clr_t^{B_4}}_S\right)$ respectively. These estimated cell loss ratio values and the observed cell loss ratio values for the virtual buffer 3 and 4 are taken as input-output pairs to estimate the parameters α and β in Eq. 5.9. For ease of expression, denote $\log\left(\widehat{clr_t^{B_3}}_M\right)$ and $\log\left(\widehat{clr_t^{B_4}}_M\right)$ by x_t^3 and x_t^4 respectively, and denote $\log\left(\widehat{clr_t^{B_3}}_S\right)$ and $\log\left(\widehat{clr_t^{B_4}}_S\right)$ by y_t^3 and y_t^4 respectively. The logarithm of observed cell loss ratios in the virtual buffer 3 and 4, $\log(clr_t^3)$ and $\log(clr_t^4)$, are denoted by z_t^3 and z_t^4 respectively. $((x_t^3, y_t^3), z_t^3)$ and $((x_t^4, y_t^4), z_t^4)$ are used as input-output pairs in Eq. 5.10 to obtain estimates of non-negative parameters α and β used in Eq. 5.9:

$$z = \alpha x + \beta y. \tag{5.10}$$

In order to increase the precision of the estimation, past CLR information is

taken into account in estimating parameters α and β. More specifically, CLR information from time interval $(t-K,\ t]$ is used in estimating α and β. There are altogether $2K$ input-output pairs $\left(\left(x^3_{t-K+1},\ y^3_{t-K+1}\right),\ z^3_{t-K+1}\right)$, $\left(\left(x^4_{t-K+1},\ y^4_{t-K+1}\right),\ z^4_{t-K+1}\right)$, $\ldots$, $\left(\left(x^3_t,\ y^3_t\right),\ z^3_t\right)$, $\left(\left(x^4_t,\ y^4_t\right),\ z^4_t\right)$ that are used to estimate the parameters α and β according to Eq. 5.10. Least-square algorithm is employed to obtain the optimum parameters $\alpha > 0$ and $\beta > 0$ which minimize the error term:

$$f(\alpha,\ \beta) = \sum_{i=0}^{K-1} \left[\left(z^3_{t-i} - \alpha x^3_{t-i} - \beta y^3_{t-i}\right)^2 + \left(z^4_{t-i} - \alpha x^4_{t-i} - \beta y^4_{t-i}\right)^2\right].$$

Solving the equations:

$$\begin{cases} \frac{\partial f(\alpha,\beta)}{\partial \alpha} &= 0 \\ \frac{\partial f(\alpha,\beta)}{\partial \beta} &= 0 \end{cases},$$

the optimum α and β which minimize the error term $f(\alpha,\ \beta)$ can be obtained:

$$\begin{cases} \alpha &= \frac{BE-AC}{DE-A^2} \\ \beta &= \frac{CD-AB}{DE-A^2} \end{cases}, \tag{5.11}$$

where

$$\begin{aligned} A &= \sum_{i=1}^{K-1} \left(x^3_{t-i} y^3_{t-i} + x^4_{t-i} y^4_{t-i}\right) \\ B &= \sum_{i=1}^{K-1} \left(x^3_{t-i} z^3_{t-i} + x^4_{t-i} z^4_{t-i}\right) \\ C &= \sum_{i=1}^{K-1} \left(y^3_{t-i} z^3_{t-i} + y^4_{t-i} z^4_{t-i}\right) \\ D &= \sum_{i=1}^{K-1} \left(\left(x^3_{t-i}\right)^2 + \left(x^4_{t-i}\right)^2\right) \\ E &= \sum_{i=1}^{K-1} \left(\left(y^3_{t-i}\right)^2 + \left(y^4_{t-i}\right)^2\right). \end{aligned}$$

It is often the case that Eq. 5.11 results in a negative α or β. Care must be taken when this happens. If parameter α given in Eq. 5.11 is negative, then Eq. 5.12 is

used to obtain the optimum non-negative parameters α and β:

$$\begin{cases} \alpha &= 0 \\ \beta &= \frac{C}{E} \end{cases} . \tag{5.12}$$

Or, if parameter β given in Eq. 5.11 is negative, then Eq. 5.13 is used to obtain the optimum non-negative parameters α and β:

$$\begin{cases} \alpha &= \frac{B}{D} \\ \beta &= 0 \end{cases} . \tag{5.13}$$

The parameter K is determined as:

$$K = \min\{J,\ M\},$$

where J is the maximum integer such that the cell loss ratios observed in the virtual buffers in the interval $(t - J,\ t]$ are all non-zero values, and M is an integer determined by the CLR estimation algorithm. Therefore there are a maximum of $2M$ input-output pairs that are used by the least square algorithm in estimating the parameters α and β using Eq. 5.10, which enables us to obtain accurate estimates of α and β.

5.2.2 Low-pass FIR Filter

In our CLR estimation algorithm, the CLR estimated using Eq. 5.9 is not used directly as a CLR estimate of the real buffer. On the contrary, the CLR estimated from Eq. 5.9 is used as an input to a low-pass filter. It is the output from the low-pass filter that is used as the CLR estimate for the real buffer. The low-pass filter is applied because our empirical observation shows that since the buffer sizes of virtual buffers are much less than that of the real buffer, the cell loss ratios of virtual buffers are easily affected by some short traffic bursts. Thus some transient high frequency components exist in the CLR observations of virtual buffers. However usually these short traffic bursts can be effectively absorbed by the real buffer and will not cause cell loss in the real buffer. Therefore, it becomes necessary that a low-pass filter be applied to eliminate the effects of these transient high frequency traffic bursts.

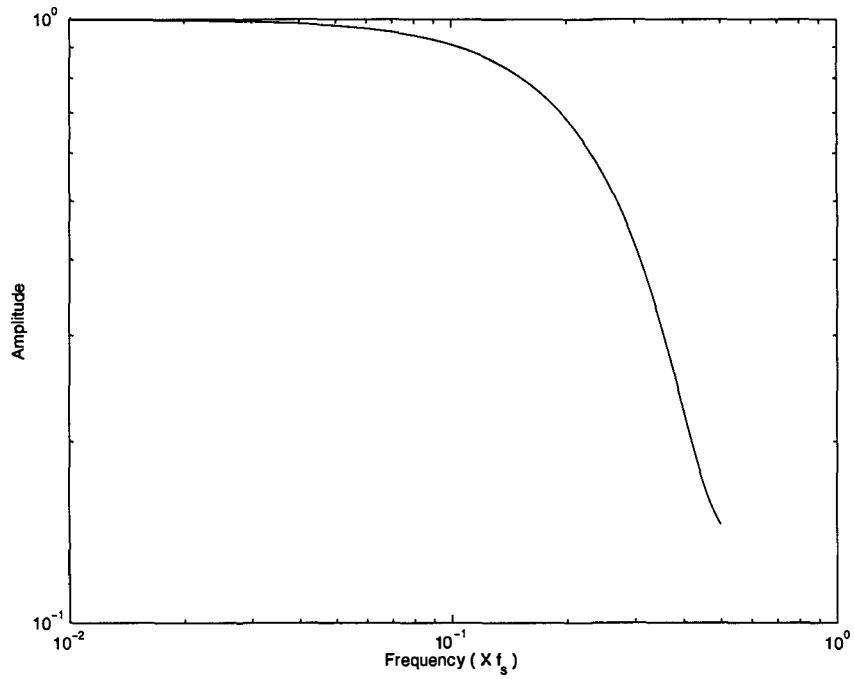


Figure 5.2: Amplitude-frequency response of the FIR filter

The low-pass filter used in our algorithm is a FIR (Finite Impulse Response) filter of the form:

$$y(t) = \sum_{i=0}^{j} b_i \times x(t-i).$$

The FIR filter is empirically designed using Matlab. Fig. 5.2 shows the amplitude-frequency response of the FIR filter. In Fig. 5.2, f_s denotes the inverse of the CLR monitoring period in our algorithm.

Fig. 5.3 shows the architecture of the CLR estimation algorithm.

5.2.3 Choice of Parameters

There are several critical parameters in the CLR estimation algorithm which need to be determined.

Parameter M determines the maximum number of past CLR information taken

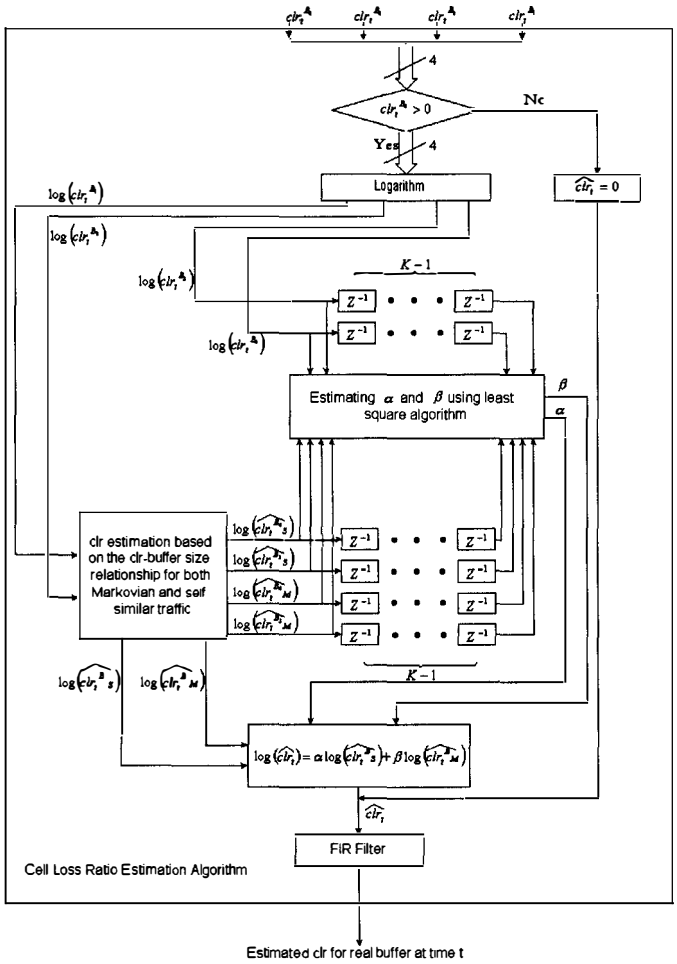


Figure 5.3: Architecture of the CLR estimation algorithm

into account in the least square algorithm for estimating parameters α and β. Taking into account some past CLR information helps to improve the precision of the estimation. However as more information is taken into account, more memory is required to store these past information and more computing power is required in the least square algorithm. Moreover, parameters α and β are to some extent indicators of the nature of network traffic (Markovian or self-similar). They change when network traffic changes. When too much past information is taken into account, it may restrict the ability of the algorithm to track variations in network traffic, which in turn decreases the precision of the estimation. In our algorithm,

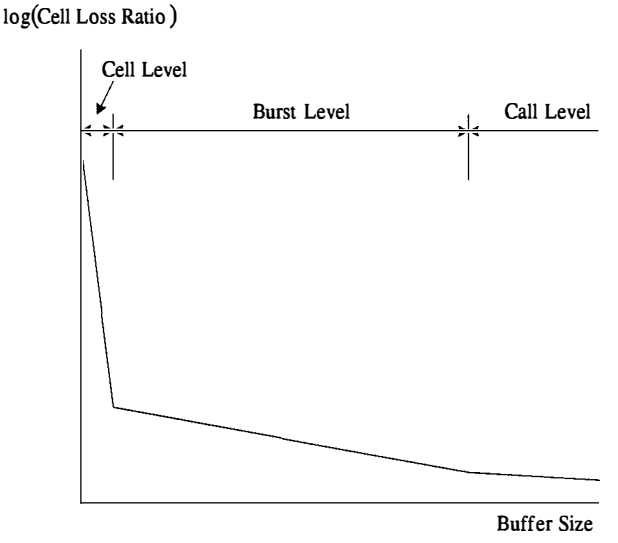


Figure 5.4: Cell level, burst level and call level cell loss ratio

M is empirically chosen to be 10. That is, there is a maximum number of 20 input-output pairs used in the least square algorithm for estimating α and β.

The sizes of the virtual buffers are also critical parameters which determine the performance of the CLR estimation algorithm. The smaller the sizes of the virtual buffers are, the larger the cell loss ratios in these virtual buffers. Hence less monitoring period is required to make a CLR estimation when virtual buffer sizes are small. It is therefore desirable that the sizes of the virtual buffers are chosen to be as small as possible. On the other hand, the CLR estimation algorithm is based on the asymptotic relationships between CLR and buffer size. These asymptotic relationships only apply for large buffers, therefore the sizes of virtual buffers can not be chosen to be too small. In-depth understanding of the asymptotic CLR-buffer size relationship is required to make a good choice of virtual buffer size.

Hui suggests that congestion should be evaluated at different levels, namely, the cell level, the burst level and the call level [14]. Accordingly the CLR-buffer size relationship can be approximately divided into three regions, namely cell level region, burst level region and call level region. This has been verified by a lot of experimental studies [117], [99], [60], [50]. Fig. 5.4 illustrates the three regions.

When buffer size falls into the cell level region, cell loss occurs because of the

simultaneous arrivals of cells from independent sources. However the aggregate traffic rate of these independent sources may be below the link capacity. When network buffer size falls into the burst level region, cell loss occurs when the aggregate traffic rate is momentarily greater than the link capacity. Buffer content grows to the limit that cell loss occurs, as long as the aggregate rate excess exists. When network buffer size falls into the call level region, cell loss occurs because excessive number of connections are admitted into the network. Therefore aggregate traffic rate exceeds link capacity for a large time-scale (comparable to the connection duration time). This different nature of cell loss determines that CLR-buffer size relationship is different in different regions.

The boundary of cell level region is a non-decreasing function of the number of connections on the link. Our empirical observations show that the boundary of cell level region usually varies between a buffer size of several ATM cells and a size of up to 20 cells, depending on the number of connections on the link. This conforms to the observation reported in [60]. Burst level region spans from several ATM cells to several thousands of cells, depending on the system size (e.g. number of connections, link rate, etc.). The size of the real buffer usually lies in this region. Moreover, when analyzing the asymptotic relationship between CLR and buffer size, a lot of researchers adopt the fluid flow model [39], [121], [81], and their loss performance analysis naturally rests on the burst level. At last, no call level dynamics are considered in the analysis of the asymptotic relationship between CLR and buffer size. Therefore, the asymptotic relationship between CLR and buffer size actually applies for a buffer size in the burst level region. Accordingly the sizes of virtual buffers should be chosen in the burst level region.

In our CLR estimation algorithm, for a small system where the number of connections is small, the minimum virtual buffer size B_1 is chosen to be 10 cells. For a large system where there are a lot of connections on the link, the minimum virtual buffer size B_1 is chosen to be 20 cells. Sizes of virtual buffers are chosen to be 5 cells apart.

Another important parameter in our CLR estimation algorithm is the CLR

monitoring period, which is the minimum time required to make valid CLR observations in the virtual buffers. There are a lot of factors that affect the CLR monitoring period, including the maximum size of the virtual buffers, the size of the real buffer, link capacity, link utilization, the resolution of the CLR estimation algorithm (the minimum cell loss ratio that can be estimated by the algorithm), etc. This makes it difficult to formulate the process of determining the CLR monitoring period. The CLR monitoring period is determined empirically. Here we try to illustrate the process through the following example. Consider a case where the CLR estimation algorithm is designed to estimate a cell loss ratio as small as 10^{-9} in a system with a buffer size of $1,000$ cells. The link capacity is $150Mbps$. The maximum virtual buffer size B_4 is chosen to be 35 cells in our CLR estimation algorithm. Our empirical observations show that when the CLR in the real buffer is around 10^{-9}, the cell loss ratio in the virtual buffer B_4 usually varies in the range from 10^{-3} to 10^{-5}. Assume that when the cell loss ratio in the real buffer is 10^{-9}, the lowest link utilization is 0.5 (when the link utilization is too low, no cell loss will occur in the real buffer). Then, if a CLR monitoring period of $50s$ is chosen, it implies that a minimum number of 88 cell losses can be observed in the monitoring period, which is enough to generate a valid observation of CLR in the virtual buffers. Therefore for this specific example, the CLR monitoring period can be chosen to be approximately $50s$. The CLR monitoring period is significantly reduced compared with the CLR monitoring period of 13 hours required using direct monitoring method.

5.3 Simulation Study

In this section, we investigate the performance of the CLR estimation algorithm using simulation. Both on-off traffic source model and real vbr video trace are used in the simulation. The objectives of the simulation are to:

- validate the effectiveness of the CLR estimation algorithm for a variety of traffic sources, and

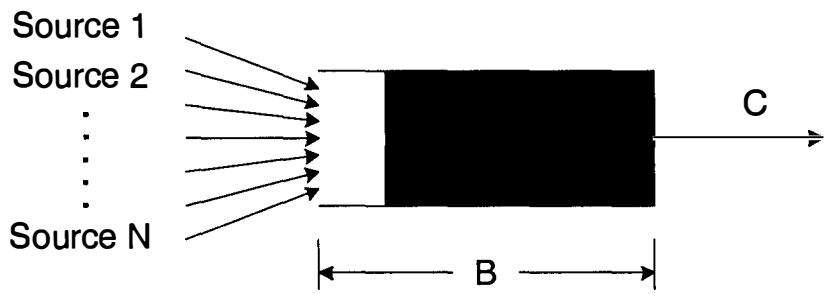


Figure 5.5: Simulation model for the CLR estimation algorithm

- evaluate the accuracy of the CLR estimation algorithm compared to other algorithms in the literature.

The simulation is carried out using OPNET. Fig. 5.5 illustrates the simulation model. There are N traffic sources in the simulation. The traffic generated by the traffic source is encapsulated into ATM cells using AAL5 protocol and then transported into a single server queue drained at a speed equal to the link capacity C. The queue has a buffer size of B. The link capacity C is engineered such that an average bandwidth utilization of 0.8 is achieved in the simulation. This is the utilization that is achieved by our CAC scheme shown in Chapter 4.

Two simulation scenarios are considered in the simulation. In the first simulation scenario, referred to as the Markovian scenario, Markovian traffic sources are used. In the second scenario, referred to as the Self-similar scenario, real vbr video traffic sources, which are known to be self-similar, are used. These two typical scenarios are used to establish the effectiveness of our CLR estimation algorithm for a variety of traffic sources.

5.3.1 Markovian Scenario

30 exponential on-off sources, which are typical Markovian traffic sources, are used in this scenario. Each on-off source has independent exponentially distributed on and off periods with means τ and γ respectively. The peak cell rate of the on-off source is pcr and the mean cell rate of the on-off source is mcr.

Table 5.1: Traffic parameters of the exponential on-off source

	pcr (kb/s)	mcr (kb/s)	τ (sec)	γ (sec)
Exponential on-off source	64	22	0.384	0.734

Table 5.1 shows the traffic parameters of the exponential on-off source. These parameters are typical parameters of telephone voice traffic [71].

Seven simulations are run in the Markovian scenario with buffer size B varying from 100 cells to 700 cells. Each simulation is run for $10,000$ seconds. When buffer size B is increased further above 700 cells, it becomes difficult to make valid cell loss ratio observations in the real buffer within the limited simulation time. In our simulation, the CLR of the real buffer and the estimated CLR are observed, and compared to investigate the performance of the CLR estimation algorithm. As a typical example, Fig. 5.6 shows the CLR of the real buffer and our estimation using a simulation where buffer size B is 300 cells. For comparison purpose, we also present the estimated CLR by the CLR estimation algorithm of Li [70].

Fig. 5.6 shows that significant improvements are achieved by our algorithm. The improvements are in two respects:

- First, our algorithm requires much less monitoring period. In our algorithm only one valid CLR observation of the virtual buffers is sufficient in order to generate a CLR estimate. In the algorithm by Li, a lot of CLR observations of the virtual buffers have to be made in order to generate a CLR estimate of the real buffer. In the simulation reported in [70], 100 CLR observations in the virtual buffers have to be made in order to generate a CLR estimate. This implies that their monitoring period is approximately 100 times larger than that required by our algorithm.

- Second, our algorithm is more accurate than the algorithm by Li. This is clearly shown in the figure.

Estimation errors do exist. In addition to the estimation error due to the algorithm itself, part of the error shown in the figure comes from CLR observation.

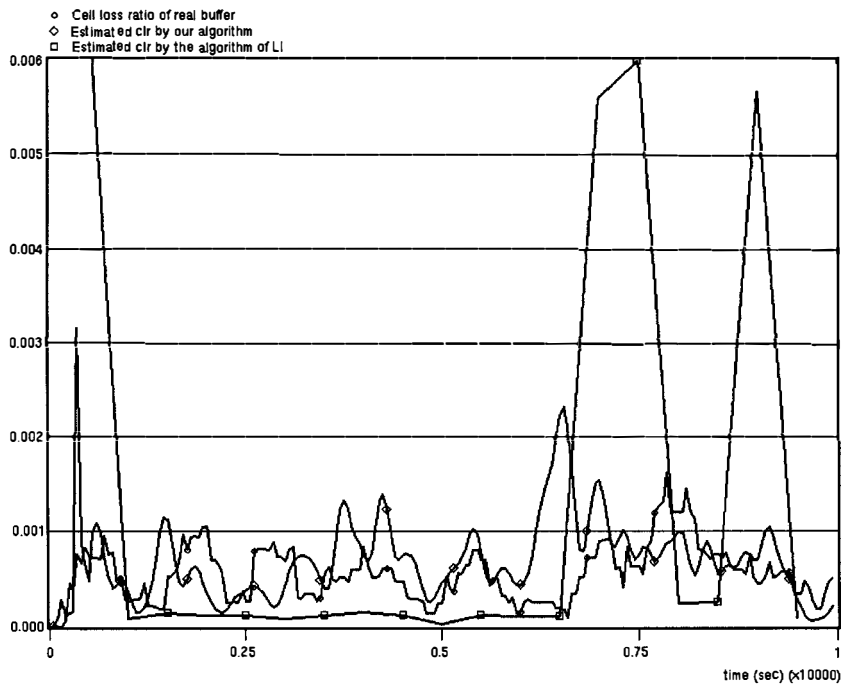


Figure 5.6: CLR estimation in the simulation with a buffer size of 300 cells

The estimated CLR of our algorithm in a monitoring period T gives an estimate of the CLR of the real buffer in the same interval. However, it is difficult to make a valid observation of CLR of the real buffer in the small monitoring period. Therefore the CLR of the real buffer shown in the figure is the ratio of lost cells to the total cells offered to the link for transmission in a much larger time interval. This mismatch becomes an error source in Fig. 5.6 as well.

Fig. 5.7 shows the estimated CLR as well as the CLR of the real buffer for different buffer sizes in the Markovian scenario. Since the Markovian scenario is a stationary scenario in the sense that there is a fixed number of connections on the link, for ease of comparison, averages of the CLR are shown in the figure. For example, the CLR of the real buffer shown in the figure for a buffer size of 400 cells is the average value of CLR observations of the real buffer in the simulation where buffer size is set to be 400 cells. Accordingly the estimated CLR shown in the figure is also the average value of estimated cell loss ratios.

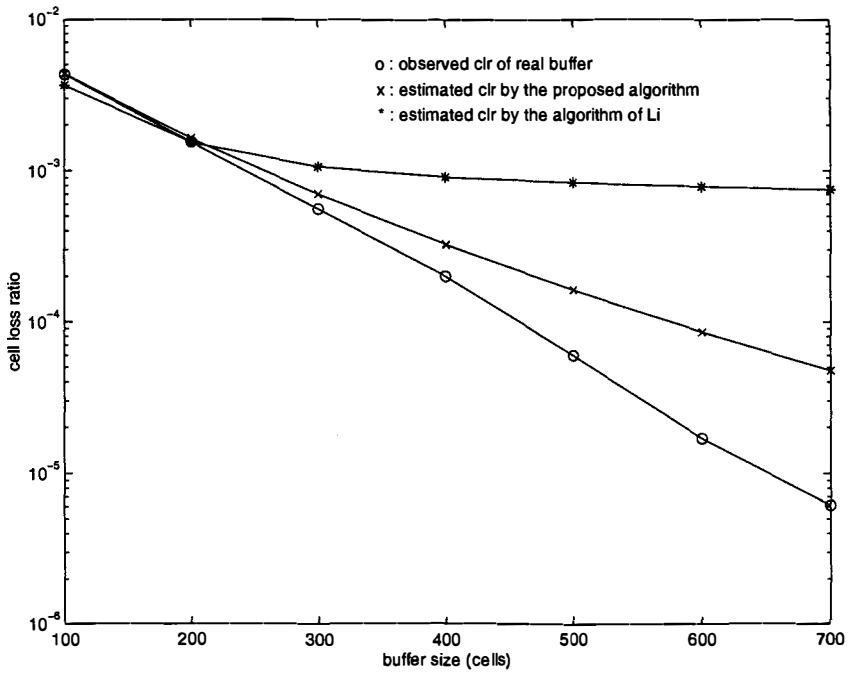


Figure 5.7: Cell loss ratio estimation in the Markovian scenario

Fig. 5.7 shows that the estimation error of our CLR estimation algorithm increases with buffer size. When buffer size is below 300 cells, the estimation algorithm is able to give an accurate estimate of CLR in the real buffer. When the buffer size is increased to 700 cells, the estimation algorithm can only give a CLR estimate that is accurate within one order of magnitude of the actual value. Fig. 5.7 also shows that the accuracy of our algorithm is much better than that proposed by Li. In our algorithm, instead of using the theoretical asymptotic relationship for Markovian and self-similar traffic directly, some amendments are made by introducing the parameters α and β so that the asymptotic relationship is captured more accurately. As an example, Fig. 5.8 shows the value of the two parameters in the simulation where buffer size is 300 cells. It can be seen that the parameters α and β are neither close to 0 nor close to 1, but fluctuate around 0.5. This implies that these two parameters do contribute to making our CLR estimation algorithm more accurate than that of Li.

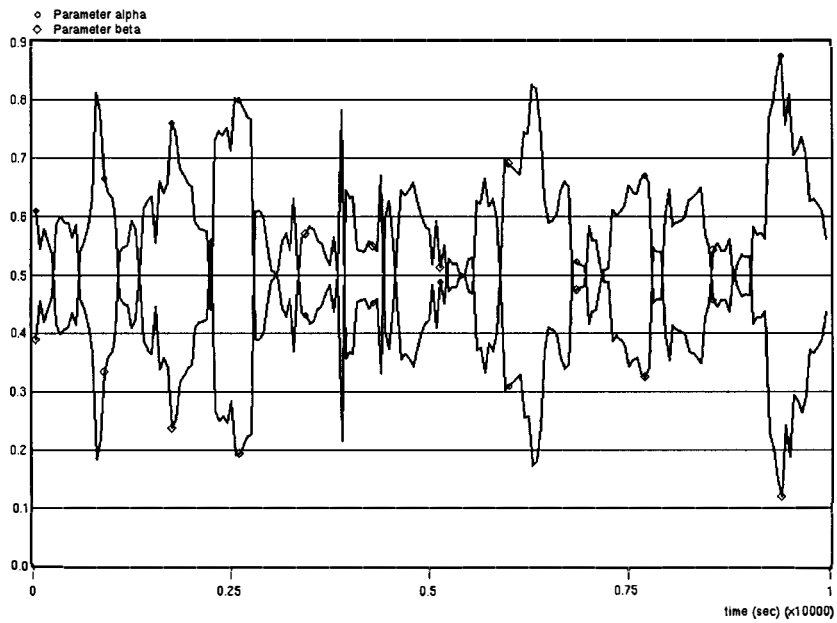


Figure 5.8: Parameters α and β in the simulation with a buffer size of 300 cells

5.3.2 Self-Similar Scenario

In the self-similar scenario, MPEG encoded movie "Starwars" is used as vbr video traffic source [123], [124][1]. This video trace is well-known to be self-similar. Five vbr video traffic sources are used in this scenario. At the beginning of the simulation, each traffic source starts to read from a random position in the vbr video file and generates traffic accordingly. When the end of the video file is reached, the traffic source continues generating traffic from the beginning of the file. Ten simulations are run in the self-similar scenario with buffer size varying from 100 cells to 1000 cells. Each simulation is run for 10,000 seconds. As a typical example, Fig. 5.9 shows the results of the simulation with a buffer size of 500 cells.

Fig. 5.9 shows that our estimation algorithm is able to estimate the cell loss ratio of self-similar traffic accurately. In sharp contrast, the algorithm of Li can only generate two CLR estimates during the 10,000 seconds simulation time, i.e.

[1]This trace is available via anonymous ftp from ftp.telcordia.com, in directory pub/vbr.video.trace.

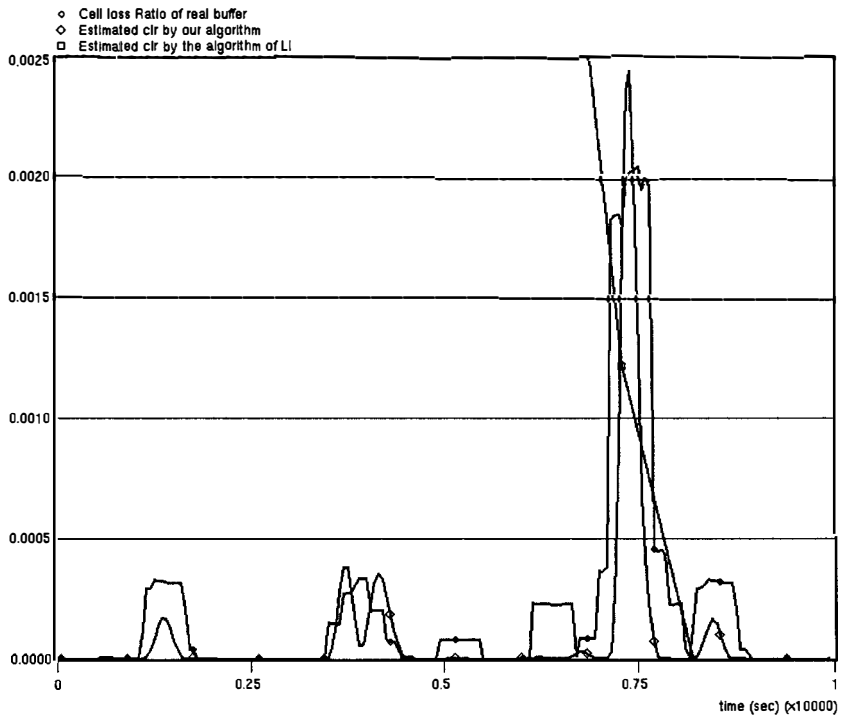


Figure 5.9: CLR estimation in a simulation with a buffer size of 500 cells

a CLR estimate of 1.217×10^{-3} at $7300s$ and a CLR estimate of 1.942×10^{-17} at $8200s$. The results of Fig. 5.9 implies that Li's algorithm fails to capture the variations of cell loss ratio with time, and is not suitable for real applications. The reason is that in Li's algorithm, in addition to much longer monitoring period required, he fails to consider the fact that in real applications whatever the virtual buffer size is chosen to be, CLR observations of zero in the virtual buffer cannot be avoided. In our algorithm the problem is solved by generating a CLR estimate of zero, then passing it to a low pass filter, when a CLR of zero is observed in virtual buffers.

Fig. 5.10 shows the estimated CLR as well as the CLR of the real buffer for different buffer sizes in the Self-similar scenario. Since the Self-similar scenario is a stationary scenario in the sense that there is a fixed number of connections on the link, for ease of comparison, only averages of the CLR are shown in the figure.

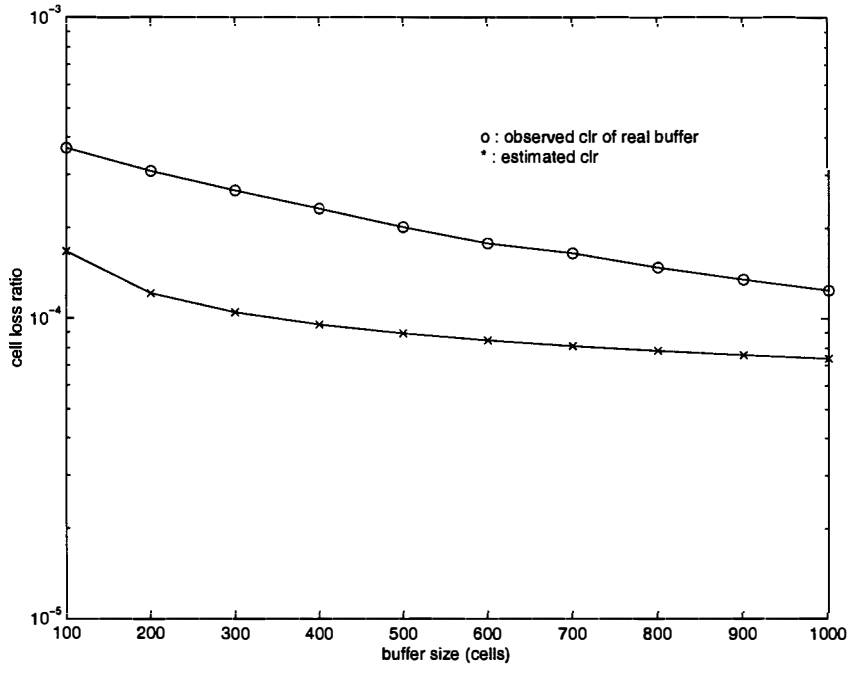


Figure 5.10: Cell loss ratio estimation in the Self-similar scenario

Fig. 5.10 shows that the estimation algorithm generally gives a CLR estimate that is accurate within one order of magnitude around the true value. This estimation error must be taken into account when the proposed CLR estimation algorithm is applied to real applications.

5.4 Summary

In this chapter, we designed an in-service QoS monitoring and estimation scheme which can be used for on-line estimation of cell loss ratio. The proposed ISME scheme is based on the asymptotic relationship between the cell loss ratio and the buffer size for both Markovian and self-similar traffic. Virtual buffer techniques used in the proposed ISME scheme significantly reduce the CLR monitoring time, which makes the proposed ISME suitable for real-time applications.

Simulation using both theoretical Markovian traffic model and real vbr video source was carried out to investigate the performance of the proposed ISME scheme.

Simulation results indicate that the proposed ISME scheme requires much less monitoring period and achieves better accuracy than that proposed in the literature. The proposed ISME scheme is therefore suitable for real applications.

Generally speaking, the proposed CLR estimation algorithm is able to generate CLR estimation that is accurate within one order of magnitude around the true value. This estimation error must be taken into account when the proposed CLR estimation algorithm is applied to real applications.

Chapter 6

Connection Admission Control - Closing the Loop

To date, a lot of CAC schemes have been proposed. Chapter 2 reviewed some of the major work in this area. Those CAC schemes, although varying in details, can be generally described by the two open-loop architectures shown in Fig. 6.1 and Fig. 6.2.

Most CAC schemes adopt the architecture of Fig. 6.1. They obtain traffic parameters from either traffic measurements or traffic descriptors. These parameters are then fed into the chosen traffic and network model to obtain model parameters. Based on performance analysis of traffic sources in the network model, either an estimate of QoS parameter when the new connection request is admitted, or an estimate of the residual bandwidth which is unused by existing network connections, or an admissible region, is obtained. If the QoS parameter when the new connection is admitted is less than QoS objective, or the residual bandwidth is greater than the bandwidth requirement of the new connection request, or the number of connections falls within the admissible region, the new connection is admitted. Otherwise, the new connection is rejected.

There are several other measurement-based admission control schemes which are based on in-service QoS monitoring [69], [70], [71]. They adopt the architecture shown in Fig. 6.2. These CAC schemes obtain an estimate of the CLR

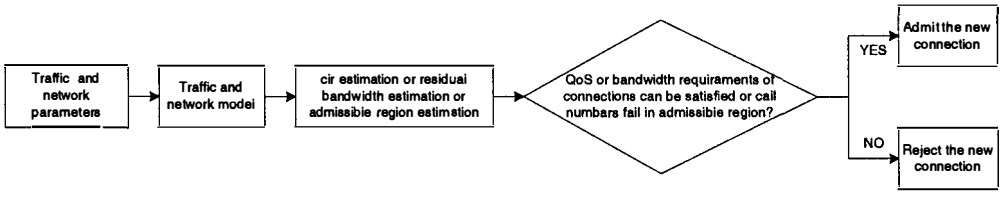


Figure 6.1: Architecture of the open-loop CAC scheme Part I

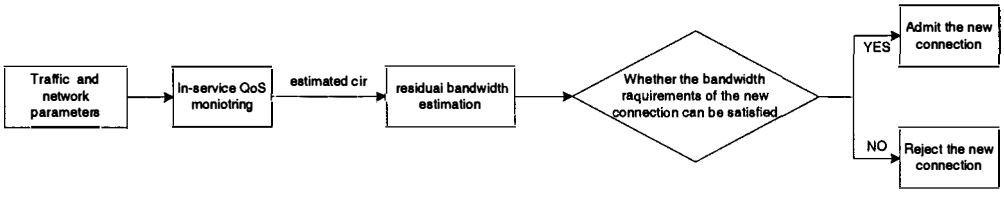


Figure 6.2: Architecture of the open-loop CAC scheme Part II

of existing traffic sources in the network through an in-service QoS monitoring and estimation scheme. This CLR estimation is then used to estimate the residual bandwidth. If the residual bandwidth is greater than the bandwidth requirement of the new connection request the new connection is admitted. Otherwise it is rejected.

There are inherent drawbacks in these open-loop architectures. The performance of a CAC scheme based on an open-loop architecture relies on accurate traffic and network models, accurate model parameters, and accurate loss performance analysis. In the open-loop architecture there is no feedback on the performance of the CAC scheme available, which can be used to adapt the CAC scheme to achieve the optimum performance of the CAC scheme. Unfortunately the complexity and heterogeneity of network traffic make it very difficult to model traffic and obtain model parameters accurately, as well as perform accurate loss performance analysis.

For CAC schemes adopting the architecture of Fig. 6.2, they represent a novel theoretical research direction. But they are based on a very rough relationship between CLR and bandwidth. Errors in estimation of both residual bandwidth and CLR may ruin the performance of these CAC schemes. Although these CAC schemes adopt a different architecture, problems puzzling these CAC schemes are

the same as those in CAC schemes adopting the architecture of Fig. 6.1.

In this chapter, we shall design a CAC scheme based on a novel closed-loop architecture which is able to overcome the inherent drawbacks of the open-loop architecture.

6.1 Architecture of the Closed-Loop CAC Scheme

Based on our work on measurement-based CAC and in-service CLR estimation algorithm, we shall propose a closed-loop architecture for CAC scheme in this section.

On the basis of our loss performance analysis using on-off traffic source model and bufferless fluid flow model, we proposed a measurement-based CAC scheme. Simulation studies showed that the proposed MBAC scheme is capable to achieve a high network utilization in a network with small buffer size. However the proposed MBAC adopts the open-loop architecture of Fig. 6.1, hence inherent drawbacks exist in the MBAC scheme using the open-loop architecture. To be more specific, the drawbacks are on two aspects:

- First, the proposed MBAC is based on loss performance analysis of the bufferless fluid flow model. For a network with small buffer size, it is able to efficiently utilize network resources - both bandwidth and buffer. But for a network with large buffer size, due to model limitations, it will generate conservative QoS estimate, and is unable to efficiently utilize network buffer. However, both bandwidth and buffer are precious network resources. Under certain conditions they are exchangeable. Therefore it is desirable that both bandwidth and buffer are efficiently utilized.

- Second, the proposed MBAC is based on loss performance analysis using on-off source model. For traffic sources that can be accurately described by the on-off source model, like voice traffic, network resources can be efficiently utilized. But for other traffic sources which cannot be accurately described by the on-off source model, the proposed MBAC will generate

conservative QoS estimate which results in in-efficient network resources utilization.

It is worth noting that these drawbacks not only exist in our MBAC, they are actually inherent drawbacks of the open-loop architecture. It could be argued that some other traffic source models are more accurate than on-off source model for specific traffic source or specific network scenario (e.g. Gaussian traffic distribution is more accurate for the aggregate traffic under heavy traffic condition). However, heterogeneity and complexity of network traffic makes it very difficult, if at all possible, to find a traffic model that can accurately model all kinds of traffic sources under all possible network scenarios. Moreover, for CAC based on complex traffic source models, the difficulty involved in obtaining model parameters and performance analysis may outweigh the advantages of these models. Therefore in real applications CAC schemes based on complex traffic source and network models may not necessarily perform better than those based on simple traffic source and network models. It is difficult to solve all these problems within the open-loop architecture itself, therefore new architectures have to be considered.

We shall now have a short review of the proposed MBAC in Chapter 4 to investigate the implementation of the closed-loop architecture.

Assume there are n independent heterogeneous traffic sources, denoted by $X_1, \ldots, X_n$, on the link. The declared peak cell rate of traffic source $X_{i\,(i=1,\ldots,n)}$ is pcr_i. All traffic rates are normalized with regard to a traffic rate unit u, where $u \geq \max\{pcr_1, \ldots, pcr_n\}$. It can be shown that the cell loss ratio of $\sum_{i=1}^{n} X_i$ is less than or equal to that of $\sum_{i=1}^{m} Y_i$, where $Y_{i\,(i=1,\ldots,m)}$ is an independent on-off source with peak cell rate 1 and activity parameter p, and $m = \lceil \sum_{i=1}^{n} pcr_i \rceil$. An estimate of p is obtained from the traffic measurements:

$$\widehat{p} = \frac{r}{m} + \alpha\sqrt{\frac{\frac{r}{m}(1-\frac{r}{m})}{m}}$$

where r is the measured mean cell rate of the link and α is the safety margin. The CLR of $\sum_{i=1}^{m} Y_i$ in the bufferless fluid flow model is calculated and taken as a

CLR upper bound for $\sum_{i=1}^{n} X_i$. A method for choosing an appropriate value for α in a bufferless system was introduced in Chapter 4.

In the proposed MBAC, the physical interpretation of the safety margin α is that a portion of bandwidth is set aside to account for estimation errors. By properly controlling the safety margin, we are able to control the achieved CLR close to the CLR objective, therefore maximize the network resources utilization while satisfying QoS requirements of connections. Fig. 6.3 illustrates the effect of controlling the safety margin. The same parameters used in the saturation scenario of section 4.3.2 are used here. By controlling the safety margin, we actually change the network resources (e.g. buffer, bandwidth) allocated to existing connections. If the CAC scheme is over-allocating the resources, the safety margin can be chosen to be smaller so that less resources are allocated to existing connections and vice versa. Therefore by choosing appropriate safety margin parameter we are able to compensate for the over-allocated bandwidth due to errors of the on-off traffic source model. Moreover, buffer and bandwidth are exchangeable under certain conditions. For a network with large buffers, less bandwidth can be allocated to network connections because large buffer is able to absorb some traffic congestion. Therefore, for a network with large buffers safety margin can be chosen to be smaller, that is, less bandwidth is allocated to connections, to efficiently utilize buffer. So, by choosing appropriate safety margin, we are able to compensate for the over-allocated bandwidth due to model errors of on-off source model and bufferless fluid flow model.

Since network traffic is changing with time, the safety margin must also change with the changing traffic in order to achieve the optimum performance of the CAC scheme.

Based on our work on measurement-based CAC and in-service CLR estimation algorithm, we propose the closed-loop architecture of Fig. 6.4 for CAC scheme. In-service QoS monitoring is introduced into the closed-loop architecture to provide feedback on the performance of the CAC scheme. The CLR information from in-service QoS monitoring and the CLR objective are fed to a control

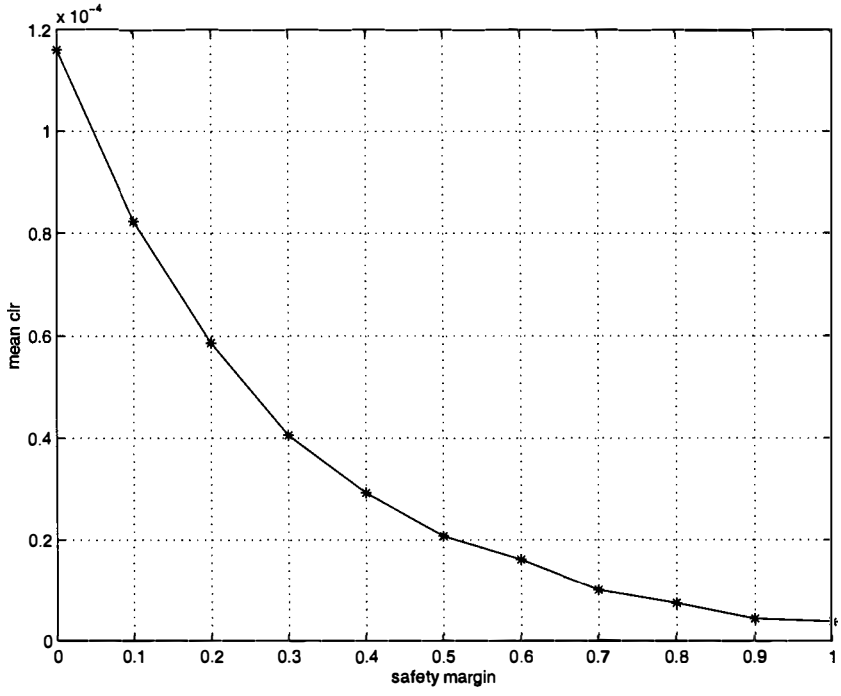


Figure 6.3: Controlling effect of the safety margin

system. Based on the CLR information the control system tunes the control parameter of the CAC, i.e. safety margin parameter. This adjusts the bandwidth allocation in the CAC scheme to account for model errors in connection admission and bandwidth allocation. In contrast to those CAC using open-loop architecture, our CAC using closed-loop architecture is intelligent - it is aware of its performance and is able to adjust its performance to maximize network resources utilization.

It is worth noting that in some other measurement-based CAC schemes, there are similar parameters which can be used to adjust connection admission and bandwidth allocation. Lee summarizes some features of these measurement-based CAC schemes [58]. Therefore, although the closed-loop architecture of Fig. 6.4 was proposed based on our MBAC, it is of general significance for CAC scheme design. That is, the proposed closed-loop architecture can also be implemented in other measurement-based CAC schemes to improve their performance.

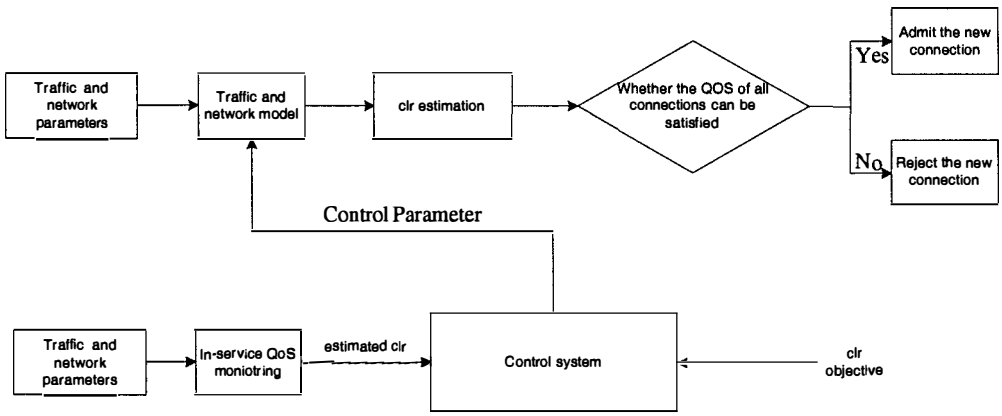


Figure 6.4: Closed-loop architecture for CAC

In the closed-loop CAC scheme, fuzzy knowledge based controller is used for the control system.

6.2 Introduction to Fuzzy Systems

In this section, we briefly introduce fuzzy logic system and show how they can be used to represent an unknown mathematical model.

Fuzzy systems, contrary to some common belief, are precisely defined with a strong mathematical basis. What is "fuzzy" in a fuzzy system is the information it deals with. The fuzzy system theory describes incomplete or vague concepts which might be difficult to formulate mathematically.

Fuzzy systems are knowledge-based systems. Since the fuzzy system described in this section is mainly used for control purpose, it is often called fuzzy knowledge based controller (FKBC). A multi-input single-output FKBC is basically made of a fuzzification module, a defuzzification module, an inference engine, and a knowledge base as shown in Fig. 6.5 [125].

The heart of a fuzzy system is a knowledge base which consists of a database and a fuzzy rule base. The basic function of the database is to provide the necessary information for the proper functioning of the fuzzification module, the rule base, and the defuzzification module. This information includes

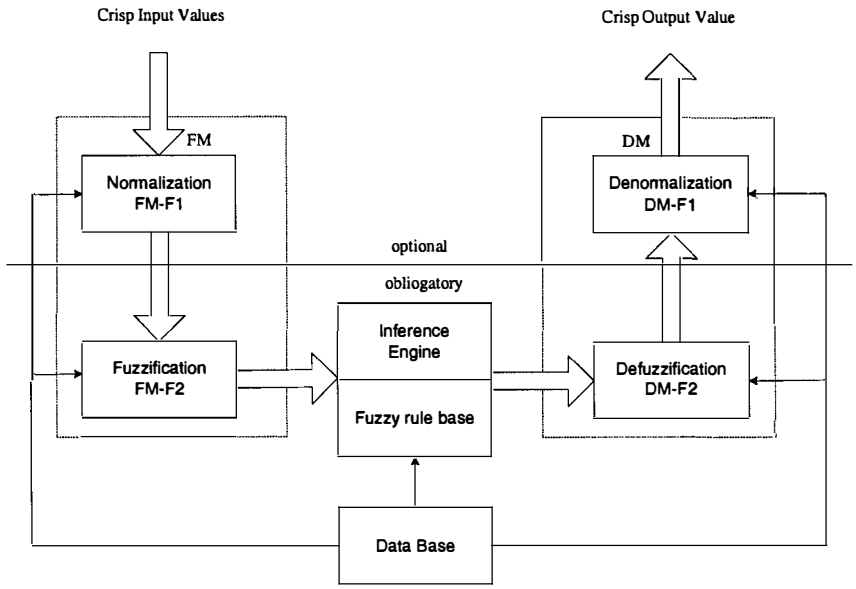


Figure 6.5: Architecture of a fuzzy knowledge based controller

- fuzzy sets (membership functions) representing the meaning of the linguistic values of the input variables and output variables, and
- physical domains and their normalized counterparts, together with the normalization/denormalization (scaling) factors.

A fuzzy rule base is a collection of fuzzy IF-THEN rules. The fuzzy inference engine uses these fuzzy IF-THEN rules to determine a mapping from fuzzy sets in the input universe of discourse $U \subset R^n$ to fuzzy sets in the output universe of discourse $V \subset R$ based on fuzzy logic principles. The fuzzy IF-THEN rules are of the following form:

$$R^{(l)}: \; IF \; x_1 \; is \; F_1^l \; and \; \ldots \; and \; x_n \; is \; F_n^l, \; THEN \; y \; is \; G^l \tag{6.1}$$

where F_i^l and G^l are fuzzy sets, $\mathbf{x} = (x_1, \ldots, x_n)^T \in U$ and $y \in V$ are input and output fuzzified vectors respectively, and $l = 1, 2, \ldots, M$. M is the number of fuzzy rules. The IF part of the rule is often called the rule antecedent and the THEN part of the rule is called the rule consequent. Practice has shown

that these fuzzy IF-THEN rules provides a convenient framework to incorporate human experts' knowledge. Each fuzzy IF-THEN rule of (6.1) defines fuzzy set $F_1^l \times \cdots \times F_n^l \rightarrow G^l$ in the product space $U \times V$.

The role of the fuzzification module is to map the crisp input values to fuzzy sets in U. A defuzzification module maps fuzzy sets in V to the crisp output value.

FM-F1 in the normalization module performs a scale transformation (i.e. an input normalization) which maps the physical values of the input variables into a normalized universe of discourse (normalized domain). FM-F2 performs the so-called fuzzification which converts crisp input variables into a fuzzy set A' in U, in order to make them compatible with the fuzzy sets representation of the process state variables in the rule antecedent. There are many types of fuzzifiers among which, the most commonly considered in practical applications is the singleton fuzzifier which maps the crisp input to a singleton fuzzy set. Intuitively, this means the membership function of the input can only be zero or one. Thus A' becomes a fuzzy singleton with support $\mathbf{x}$, that is, $\mu_{A'}(\mathbf{x}') = 1$ for $\mathbf{x}' = \mathbf{x}$ and $\mu_{A'}(\mathbf{x}') = 0$ for all other $\mathbf{x}' \in U$ with $\mathbf{x}' \neq \mathbf{x}$. $\mu_{A'}$ is the membership function of fuzzy set A'. A membership function μ_F of a fuzzy set F is a function $\mu_F : U \rightarrow [0, 1]$, where U is the input universe of discourse of F. In this thesis, the singleton fuzzifier is chosen to implement our fuzzy system.

The inference engine computes the overall value of the control output variable based on the individual contributions of each rule in the rule base. Each such individual contribution represents the value of the control output variable as computed by a single rule. Let a fuzzy set A' in U be the input to the fuzzy inference engine, then each fuzzy IF-THEN rule of (6.1) determines a fuzzy set B^l in V using the sup-star composition. That is:

$$\mu_{B^l}(\mathbf{y}) = sup_{\mathbf{x} \in U} \left[\mu_{F_1^l \times \cdots \times F_n^l \rightarrow G^l}(\mathbf{x}, y) \star \mu_{A'}(\mathbf{x}) \right]. \tag{6.2}$$

There are many rules that can be chosen for the fuzzy implication $\mu_{F_1^l \times \cdots \times F_n^l \rightarrow G^l}(\mathbf{x}, y)$, to name but a few, mini-operation rule, product-operation rule, arithmetic rule, maxmin rule, etc. In this thesis, the product-operation rule is chosen, therefore:

$$\mu_{F_1^l \times \cdots \times F_n^l \rightarrow G^l}(\mathbf{x}, y) = \mu_{F_1^l \times \cdots \times F_n^l}(\mathbf{x}) \mu_{G^l}(y). \tag{6.3}$$

Moreover, product-operation is also chosen for the $\star$ operation. That is:

$$\mu_{B^l}(y) = sup_{\mathbf{x}\in U}\left[\mu_{F_1^l\times\cdots\times F_n^l}(\mathbf{x})\mu_{G^l}(y)\mu_{A'}(\mathbf{x})\right]. \tag{6.4}$$

In Eq. 6.4, $\mu_{F_1^l\times\cdots\times F_n^l}(\mathbf{x})$ is defined either according to the mini-operation rule:

$$\mu_{F_1^l\times\cdots\times F_n^l}(\mathbf{x}) = \min\left\{\mu_{F_1^l}(x_1), \cdots, \mu_{F_1^l}(x_n)\right\} \tag{6.5}$$

or according to the product-operation rule:

$$\mu_{F_1^l\times\cdots\times F_n^l}(\mathbf{x}) = \mu_{F_1^l}(x_1)\times\cdots\times\mu_{F_1^l}(x_n). \tag{6.6}$$

Again product operation rule is chosen for the definition of $\mu_{F_1^l\times\cdots\times F_n^l}(\mathbf{x})$.

The defuzzification module maps the obtained fuzzy sets in V to a crisp output value. The most commonly used defuzzifier is the center-average defuzzifier which calculates a weighted average of the centers of the output fuzzy sets to produce the crisp output. The center average defuzzifier is defined as:

$$y = \frac{\sum_{l=1}^{M}\overline{y}^l\left(\mu_{B^l}\left(\overline{y}^l\right)\right)}{\sum_{l=1}^{M}\left(\mu_{B^l}\left(\overline{y}^l\right)\right)}, \tag{6.7}$$

where $\overline{y}^l$ is the center of the fuzzy set G^l, that is, the point in V at which $\mu_{G^l}(y)$ achieve its maximum value, and $\mu_{B^l}(y)$ is given by Eq. 6.2. This defuzzifier is chosen to implement our fuzzy system.

Lemma 3 *The fuzzy logic systems with center average defuzzifier in Eq. 6.7, product-inference rule in Eq. 6.4 and 6.6, and singleton fuzzifier are of the following form:*

$$f(\mathbf{x}) = \frac{\sum_{l=1}^{M}\overline{y}^l\left(\prod_{i=1}^{n}\mu_{F^l}(x_i)\right)}{\sum_{l=1}^{M}\left(\prod_{i=1}^{n}\mu_{F^l}(x_i)\right)} \tag{6.8}$$

where $\overline{y}^l$ is the point at which μ_{G^l} achieves its maximum value, and it is assumed that $\mu_{G^l}\left(\overline{y}^l\right) = 1$ [126].

If denormalization unit is adopted, then the output of the above fuzzy logic system becomes $\eta f(\mathbf{x})$, where η is the denormalization factor.

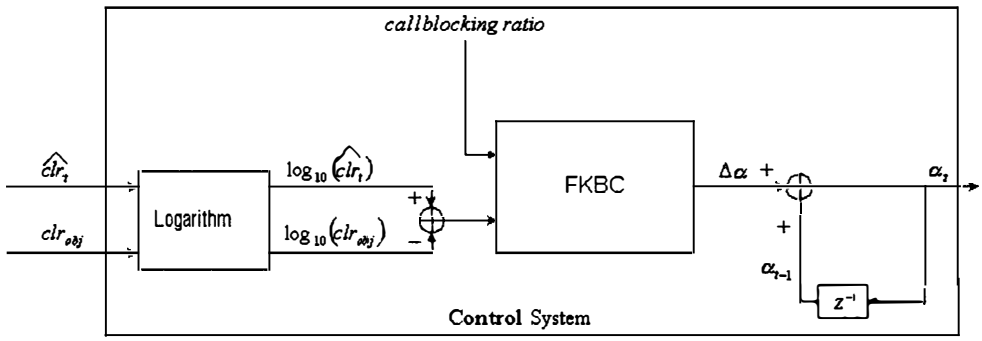


Figure 6.6: Architecture of the control system

6.3 Control System Design

Fig. 6.6 shows the architecture of the control system used in our closed-loop CAC scheme.

The CLR estimate made by the in-service CLR estimation algorithm at discrete time t, denoted by $\widehat{clr_t}$, and the CLR objective, denoted by clr_{obj}, are used as inputs to the control system. The output of the control system is the safety margin α which can be used to control the connection admission and bandwidth allocation in the CAC scheme.

It is noticed that the CLR of the system is lower than the CLR objective does not necessary imply that the CAC scheme is conservative, it can also be caused by low call arrival rate. Based on this observation, the call blocking ratio is also taken into account in the control system design. Here *call blocking ratio* is defined as the ratio of the number of calls rejected by the CAC scheme to the number of connection requests during the same CLR monitoring period when CLR estimation $\widehat{clr_t}$ is made. Only when the CLR of the system is lower than the CLR objective while at the same time a non-zero call blocking ratio is observed, it means that the CAC scheme is conservative. Hence the safety margin parameter of the CAC scheme needs to be changed such that less bandwidth is allocated to existing connections.

Call blocking ratio, and the difference between logarithm of the estimated CLR $\log_{10}\left(\widehat{clr_t}\right)$ and CLR objective $\log_{10}(clr_{obj})$

$$\varepsilon = \log_{10}\left(\widehat{clr_t}\right) - \log_{10}(clr_{obj}),$$

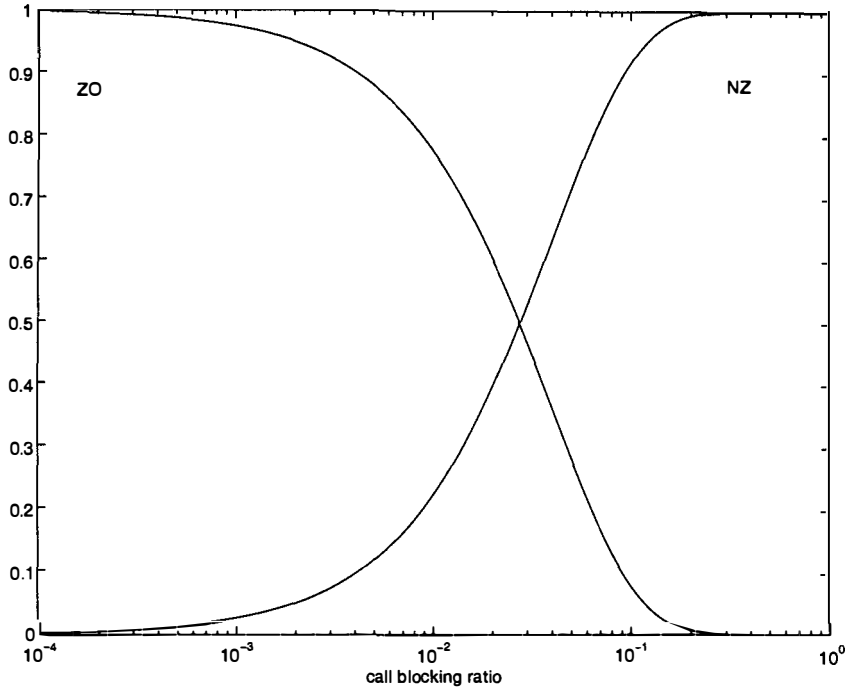


Figure 6.7: The fuzzy sets ZO and NZ on the domain [0, 1] for call blocking ratio

are fed into the FKBC as input values. The output of the FKBC is $\triangle\alpha$, which is the change made on the safety margin α.

Input value normalization unit FM-F1 is not adopted in our FKBC design. The ranges of input values, call blocking ratio and ε, are taken into account in the design of the fuzzy sets.

Fig. 6.7 shows the fuzzy sets ZO (zero) and NZ (non-zero) on the domain [0, 1] for call blocking ratio.

The membership function of the fuzzy set NZ is defined as:

$$\mu_{NZ}^{cbr}(x) = \frac{1 - e^{-ax}}{1 - e^{-a}}$$

where a is a positive constant. In our FKBC a is chosen to be 25. The membership function of the fuzzy set ZO is defined as $\mu_{ZO}^{cbr}(x) = 1 - \mu_{NZ}^{cbr}(x)$.

Fig. 6.8 shows the fuzzy sets NL (negative large), NS (negative small), ZO (zero), PS (positive small) and PL (positive large) for ε. Gaussian like membership

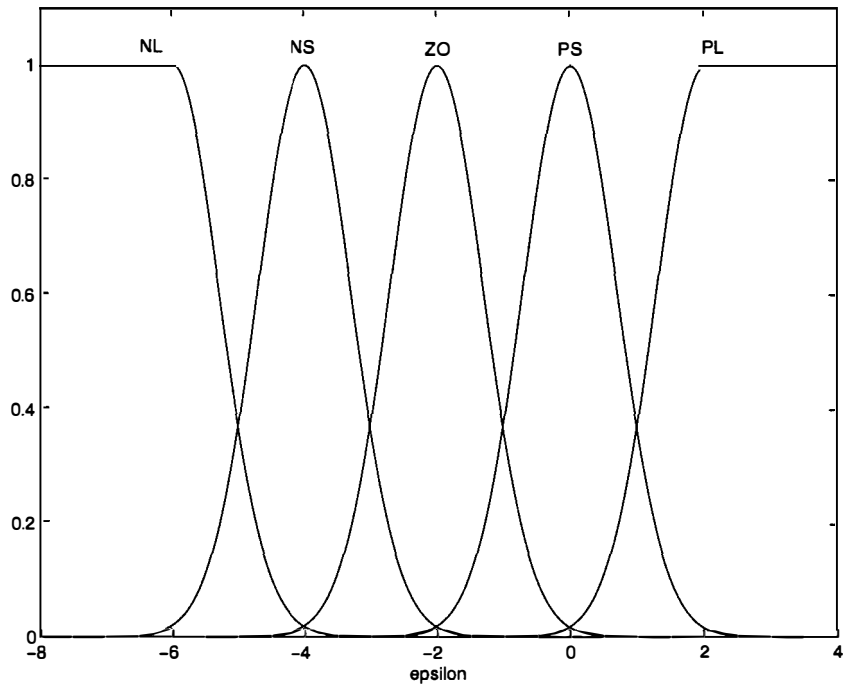


Figure 6.8: Fuzzy sets NL, NS, ZO, PS and PL for ε

functions of the form

$$\mu_i^{\varepsilon}(x) = e^{-\frac{(x-a_i)^2}{\sigma_i^2}}$$

are chosen for these fuzzy sets. For the special case of $\widehat{clr_t} = 0$, it can be shown that $\varepsilon = -\infty$. The corresponding membership functions are: $\mu_{NL}^{\varepsilon}(-\infty) = 1$, and the membership functions of other fuzzy sets $\mu_{NS}^{\varepsilon}(-\infty)$, $\mu_{ZO}^{\varepsilon}(-\infty)$, $\mu_{ZO}^{\varepsilon}(-\infty)$, $\mu_{PS}^{\varepsilon}(-\infty)$, $\mu_{PL}^{\varepsilon}(-\infty)$ are zero.

It is noted that in Fig. 6.8 the center of the fuzzy set ZO is not chosen to be zero, but -2. The reasons are that:

- the CAC scheme not only requires that the control system controls the CLR of the system to be close to the CLR objective clr_{obj} in order to maximize the network resources utilization;
- moreover, in order to guarantee the QoS requirements of the connections,

the CAC scheme also requires that the CLR of the system is controlled to be less than the CLR objective.

Therefore the center of the ZO fuzzy set has to be chosen to be less than zero in order to guarantee that the QoS requirements of the connections are satisfied. Moreover, as shown in Chapter 5, our in-service CLR monitoring algorithm can only generate a CLR estimate that is accurate within one order of magnitude around the true value. This estimation error must also be taken into account when designing the control system.

It is shown in [127] that if the membership functions of fuzzy sets are Gaussian, then the fuzzy system with a singleton fuzzifier, product inference engine and center-average defuzzifier of the form in Eq. 6.8 has universal approximation capabilities. That is, for any given real continuous function $g(\mathbf{x})$ and arbitrary $\zeta > 0$, there exists a fuzzy system in the form of Eq. 6.8 such that

$$\sup_{\mathbf{x} \in R^n} |f(\mathbf{x}) - g(\mathbf{x})| < \zeta.$$

This justifies our choice of the singleton fuzzifier, product inference engine and center average defuzzifier with Gaussian membership functions for ε to implement our fuzzy system.

Denote by y the fuzzy output from the inference engine. The fuzzy IF-THEN rules are defined as follows:

- IF call blocking ratio is ZO *and* ε is NL THEN y is ZO
- IF call blocking ratio is ZO *and* ε is NS THEN y is ZO
- IF call blocking ratio is ZO *and* ε is ZO THEN y is ZO
- IF call blocking ratio is ZO *and* ε is PS THEN y is PS
- IF call blocking ratio is ZO *and* ε is PL THEN y is PL
- IF call blocking ratio is NZ *and* ε is NL THEN y is NL
- IF call blocking ratio is NZ *and* ε is NS THEN y is NS

- IF call blocking ratio is NZ *and* ε is ZO THEN y is ZO
- IF call blocking ratio is NZ *and* ε is PS THEN y is PS
- IF call blocking ratio is NZ *and* ε is PL THEN y is PL

The centers of fuzzy sets NL, NS, ZO, PS and PL of the output variable y are chosen to be $\overline{y}_{NL} = -5, \overline{y}_{NS} = -1, \overline{y}_{ZO} = 0, \overline{y}_{PS} = 10, \overline{y}_{PL} = 50$ respectively. The centers of PL and PS are chosen to be much larger than those of NL and NS in order to ensure a quick response of the control system when QoS violation happens, and to ensure the robustness of the QoS guarantees of the CAC scheme.

The outputs from the fuzzy inference engine are defuzzified according to Eq. 6.7. The defuzzified output control variable is denormalized so that the obtained $\triangle\alpha$ is mapped into the practical range of interest. The denormalization factor η is chosen to be 0.001. The denormalization factor is chosen to be small because any change in the value of the safety margin parameter takes time to be reflected in the real system in the form of an increase or decrease in CLR. Thus, it must be ensured that changes in the value of safety margin parameter is small in any short time period. Moreover, a small denormalization factor also helps to keep the system under control when occasional large CLR estimation error occurs.

The changes in the safety margin are caught by the period update of the clrf function. The procedure of updating the clrf function is introduced in Chapter 4. Therefore the closed-loop architecture that is used to control the safety margin parameter and the measurement-based CAC scheme are almost two independent processes. The closed-loop architecture controls the tightness of the CAC scheme by controlling the safety margin parameter, but the normal running of the CAC scheme is not affected by the closed-loop architecture. The connection admission is not interrupted by the closed-loop control architecture.

6.4 Simulation Study

In this section, we shall test the performance of the closed-loop CAC scheme using simulation. The CAC scheme used in the closed-loop architecture is the same

as that used in section 4.3. However the closed-loop architecture is introduced to control the safety margin parameter of the CAC scheme to achieve optimum performance of the CAC scheme.

Buffer size in the simulation of this section is chosen to be $1,000$ ATM cells, in contrast to the simulation in section 4.3 where a small buffer size of 20 ATM cells was chosen. This large buffer size is used to test the ability of the closed-loop CAC scheme to overcome the restriction of model error. The CAC scheme reported in Chapter 4 is based on the bufferless fluid flow model. This CAC scheme is able to efficiently utilize network resources in a network with a small buffer size, but in a network with a large buffer size it fails to efficiently utilize the large buffer size due to the restriction of the bufferless fluid flow model. The CAC scheme reported in Chapter 4 achieves the same bandwidth utilization in a network with a large buffer size as that in a network with a small buffer size. However buffer size and bandwidth are exchangeable network resources under certain conditions. Therefore a higher bandwidth utilization should be achieved in a network with a large buffer size. We except that the closed-loop CAC scheme is able to overcome the restrictions of model error and achieve higher bandwidth utilization in a network with a large buffer size. So, a large buffer size of $1,000$ ATM cells is chosen in this section.

All other parameters are chosen to be the same as those in section 4.3. These parameters are recited below for convenience.

The cell loss ratio objective is set to be 10^{-4}. The switching speed of the ATM switch is set to be infinity, hence every incoming cell is placed immediately in the output buffer. Link utilization and cell loss ratio are observed in the simulation. Link utilization is calculated as the ratio of instantaneous link traffic rate to the link capacity. Cell loss ratio at time t is calculated as the ratio of the number of cell loss occurred in the interval $(t - T_c,\ t]$ to the total number of cells offered to the link for transmission in the same interval, where T_c equals $500s$. In each scenario, there are several types of traffic sources multiplexed on the link. The connection arrival process of each type of traffic source is a Poisson process with a mean of λ calls per second. The connection holding time for all traffic types is

exponentially distributed with a mean of 100 seconds.

6.4.1 Simulation Using Exponential On-Off Sources

First, simulation using exponential on-off source model is carried out. Exponential on-off source is a typical Markovian traffic source. The duration of the on and off periods are independent and exponentially distributed with means β and γ respectively. Three types of traffic sources are multiplexed on the link in the simulation. Traffic rate unit u is set to be $100kb/s$, which is the maximum pcr of the three traffic types. Clrf update period is chosen to be 0.2 second.

As that in section 4.3, two scenarios are considered. In the first scenario, referred to as the saturation scenario, the call arrival rate is chosen to be as high as that in the saturation scenario of section 4.3.2. The high call arrival rate means that the system is continually receiving new connection requests. Thus the CAC scheme is expected to achieve the maximum utilization in the saturation scenario. This scenario is used to establish the performance of the closed-loop CAC scheme with regard to QoS guarantees, because if calls are offered at a very high rate, the rate at which calls are admitted in error becomes very large too [16]. In the second scenario, referred to as the moderate scenario, the call arrival rate is chosen to be the same as that in the moderate scenario of section 4.3.3. Moderate scenario is used to test the performance of the closed-loop CAC scheme under moderate call arrival rate.

Each simulation starts with conservative guess of safety margin parameter rather than a randomly selected value. Under the control of the FKBC, the safety margin will converge to a stable value. Starting with conservative guess of safety margin parameter rather than a random value helps the simulation to arrive at a stable state quicker, hence significantly reduce the simulation time. However, the validity of the simulation results is not affected.

Table 6.1: Parameters of the three traffic types in the saturation scenario

	$\lambda\,(s^{-1})$	$pcr\,(kb/s)$	$burstiness$	$\beta\,(s)$	$\gamma\,(s)$
type 1	10	100	10	0.424	3.816
type 2	50	50	5	0.424	1.696
type 3	100	10	2	0.424	0.424

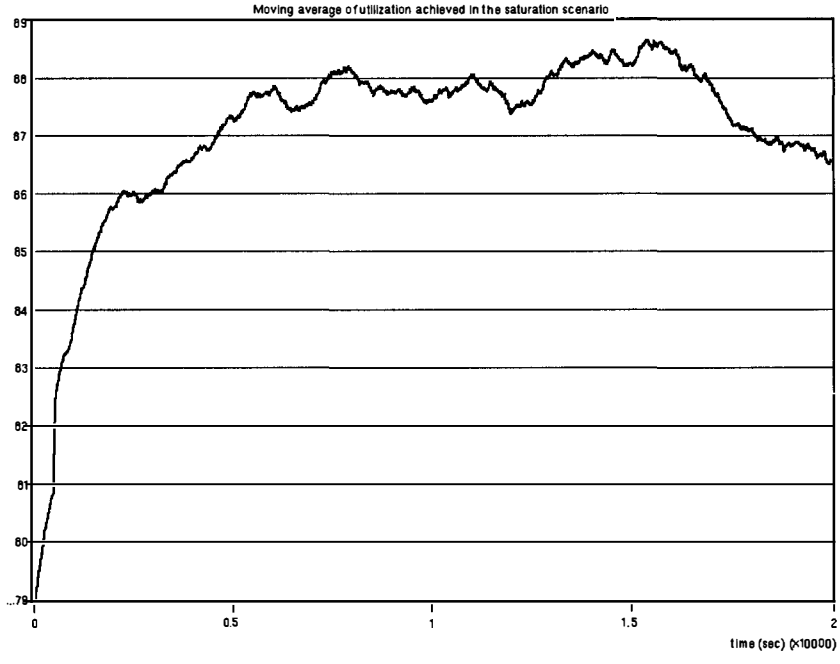


Figure 6.9: Moving average of utilization achieved in the saturation scenario

Saturation Scenario

The parameters of the three traffic types of the saturation scenario are listed in Table 6.1.

The simulation was run for 20,000 seconds. Fig. 6.9 shows the moving average of the observed utilization. Moving average window size is 1,000 seconds. Fig. 6.10 shows the observed cell loss ratio, Fig. 6.11 shows the number of each connection type on the link, and Fig. 6.12 shows the variations of safety margin parameter in the simulation.

Fig. 6.12 shows that under the control of the FKBC, the safety margin parameter α starts from a conservative guess of $\alpha = 0$ and arrives at a stable value of

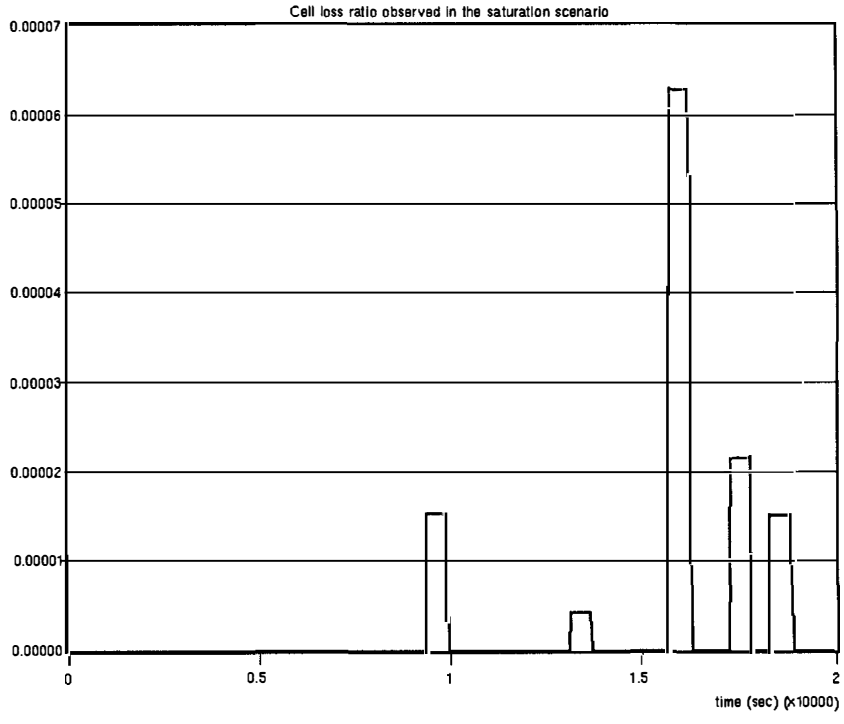


Figure 6.10: CLR observed in the saturation scenario

around -0.40. With the decrease of safety margin, the utilization achieved by the CAC scheme rises from an initial value of around 0.79 to 0.88. During this process, an increase in cell loss ratio is also observed. However the cell loss ratio is successfully controlled below the CLR objective.

This simulation builds the validity of our closed-loop CAC scheme. Compared with the CAC scheme without closed-loop control reported in Chapter 4, network utilization increases from 0.76 to 0.88. Therefore the closed-loop architecture is able to significantly increase network utilization. This increase in utilization is achieved by utilizing the large buffer in the network efficiently. Therefore, this simulation demonstrates the ability of the closed-loop CAC scheme to overcome the restriction of model error, and achieves better network resources utilization.

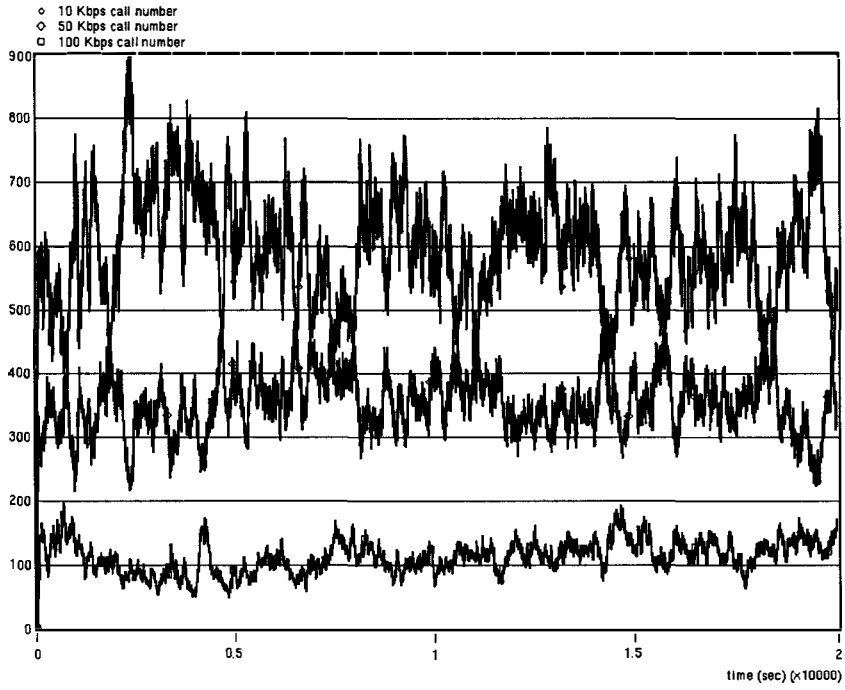


Figure 6.11: Call number in the saturation scenario

Table 6.2: Parameters of the three traffic types in the moderate scenario

	$\lambda\,(s^{-1})$	$pcr\,(kb/s)$	$burstiness$	$\beta\,(s)$	$\gamma\,(s)$
type 1	0.202	100	10	0.424	3.816
type 2	0.756	50	5	0.424	1.696
type 3	1.512	10	2	0.424	0.424

Moderate Scenario

In this scenario, we study the case when calls of each type arrive at a moderate rate. The parameters of the three traffic types of the moderate scenario are listed in Table 6.2.

The simulation was run for 20, 000 seconds. Fig. 6.13 shows the moving average of the observed utilization, Fig. 6.14 shows the number of each traffic type on the link, and Fig. 6.15 shows the call blocking ratio, where call blocking ratio is defined as ratio of the number of calls rejected by the CAC scheme to the number of connection requests during the CLR monitoring period T of the in-service CLR

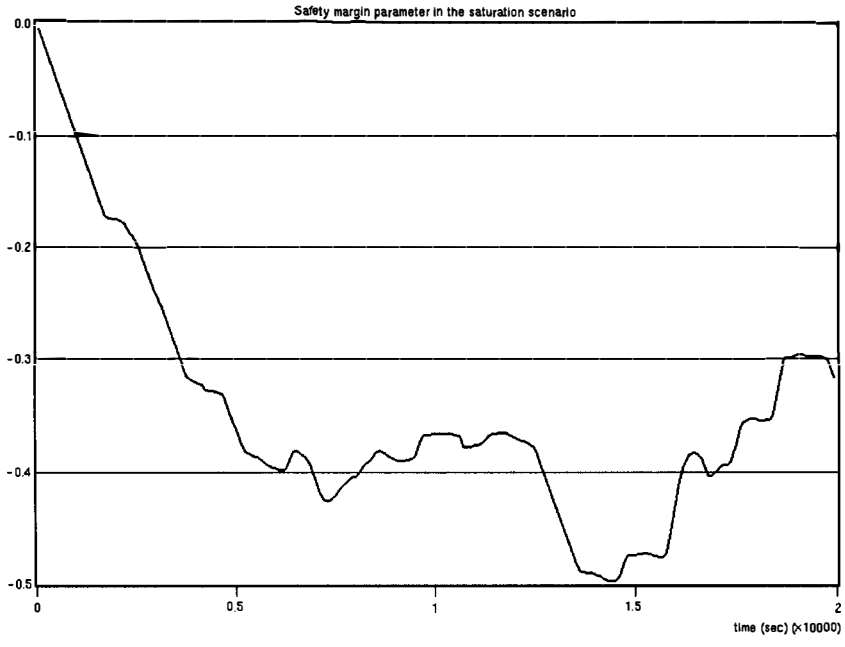


Figure 6.12: Safety margin parameter in the saturation scenario

estimation algorithm. T is chosen to be $50s$. Fig. 6.16 shows the variations of the safety margin parameter in the simulation.

An utilization of about 0.69 is achieved in the moderate scenario and no cell loss is observed. Compared with the utilization achieved in section 4.3.3, there is only a modest increase in utilization. However, Fig. 6.15 shows that the closed-loop CAC scheme is able to decrease the call blocking ratio by decreasing the safety margin parameter according to the feedback information from the in-service QoS monitoring algorithm. When the safety margin decreases to a stable value of around -0.83, all network connections are admitted. Therefore the low utilization observed in this scenario is because there are not enough connection requests in this scenario, not because the CAC scheme is conservative. Fig. 6.16 shows the variations in safety margin parameter. Safety margin parameter decreases very slowly in the moderate scenario. When the call blocking ratio is very low, the control system is very cautious in decreasing the safety margin parameter. The reason is that in this case the cell loss ratio is below the CLR objective does not necessarily mean the CAC scheme is conservative, it may also be caused by a low

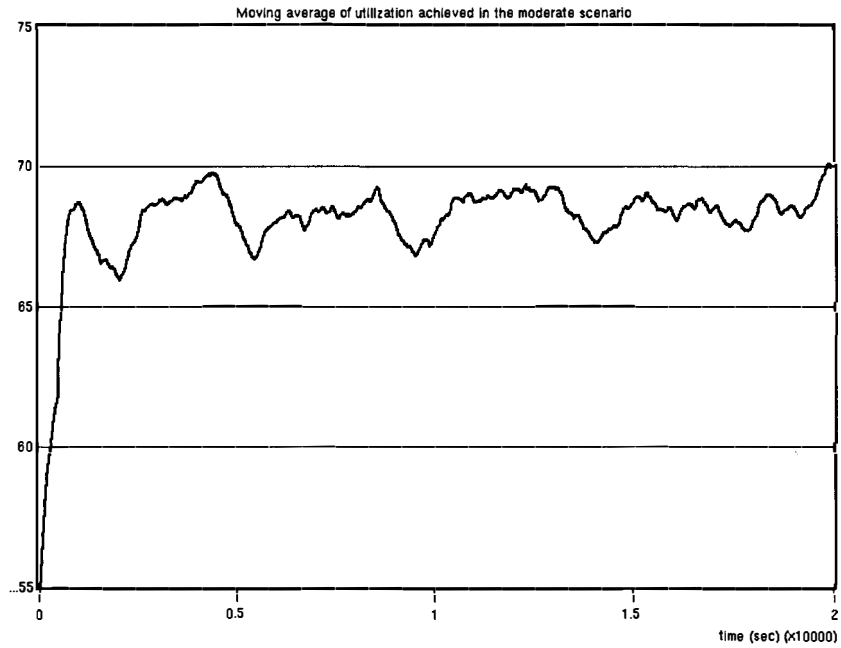


Figure 6.13: Moving average of utilization achieved in the moderate scenario

connection arrival rate. The safety margin parameter stops decreasing when the observed call blocking ratio is zero.

It is worth noting that, in order to establish the performance improvement of the closed-loop CAC scheme over the CAC scheme without closed-loop control, the parameters in this scenario are chosen to be the same as those in section 4.3.3. If either the bandwidth requirements of connections or the call arrival rate of each traffic type is increased, the link utilization in this scenario will increase until a performance similar to that shown in the saturation scenario in the last section is achieved.

6.4.2 Simulation Using Real Traffic Sources

In previous sections we studied the performance of the closed-loop CAC scheme using the exponential on-off source model. In this section, we will further investigate the performance of our CAC scheme using variable bit rate video sources.

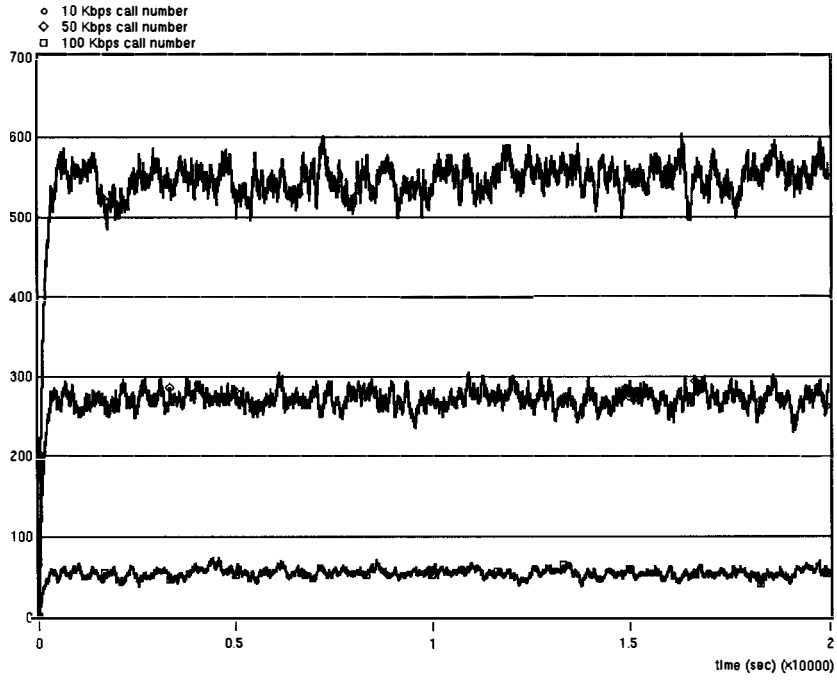


Figure 6.14: Call number in the moderate scenario

Eight Motion-JPEG (M-JPEG) encoded movies are used in the simulation. Accordingly, there are eight traffic types. Connections of each type have an exponentially distributed duration with a mean of $100s$, and the connection arrival process is a Poisson process with a mean of $1call/s$. When a connection of a type is admitted, it starts reading the corresponding M-JPEG encoded movie file from the beginning and generates traffic according to the movie file. The statistics of the M-JPEG encoded movies are shown in Table 6.3. The frame rate of the M-JPEG encoded movies is $30frames/s$. Details about the M-JPEG encoded movies can be found in [118]. Traffic rate unit u is chosen to be $7.288Mbps$, which is the maximum peak cell rate of the movie sources. OC3 link is used in the simulation. The measurement window size is chosen to be 0.10 second. Clrf update period is chosen to be 0.2 second.

The simulation was run for $10,000s$. Fig. 6.17 shows the moving average of the observed utilization. Fig. 6.18 shows the cell loss ratio observed in this scenario. Fig. 6.19 and 6.20 show the number of calls of each type multiplexed

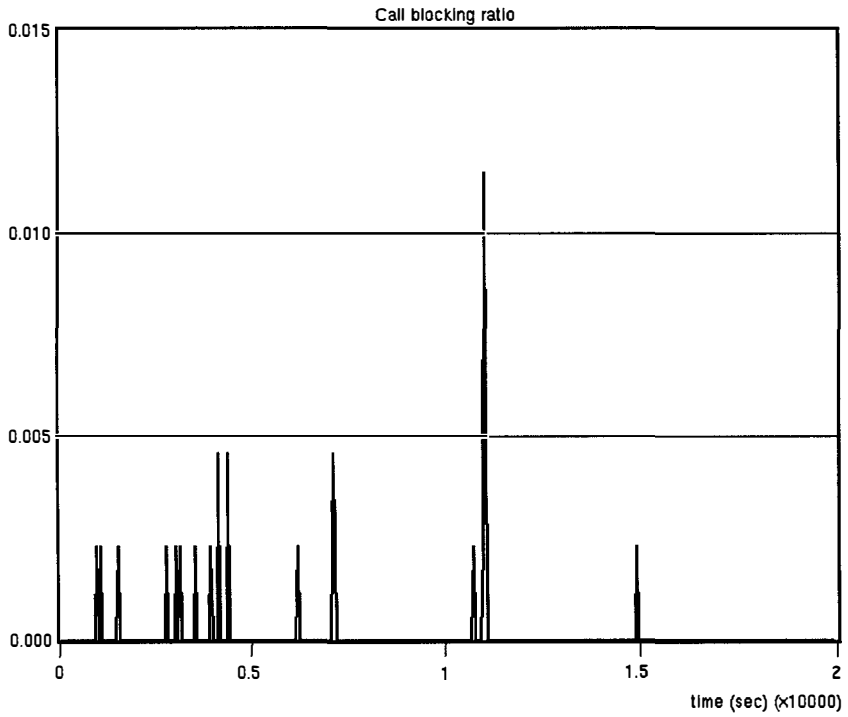


Figure 6.15: Call blocking ratio in the moderate scenario

on the link. Fig. 6.21 shows the variations of the safety margin parameter in the simulation.

When our CAC scheme based on bufferless fluid flow model and the on-off traffic source model is applied to real vbr video traffic sources, in addition to the error caused by the bufferless fluid flow model, the on-off traffic source model error also affects the performance of the CAC scheme. Therefore, the open-loop CAC scheme reported in Chapter 4 achieves much lower utilization when it is applied to real vbr video sources than that achieved when on-off source model is used.

Fig. 6.17 shows that when the closed-loop architecture is employed, a utilization of about 0.87 can be achieved when the closed-loop CAC scheme is applied to real vbr video sources. This utilization is significantly higher than that achieved by the corresponding open-loop CAC scheme under the same conditions, and is close to that achieved by the closed-loop CAC scheme in the saturation scenario

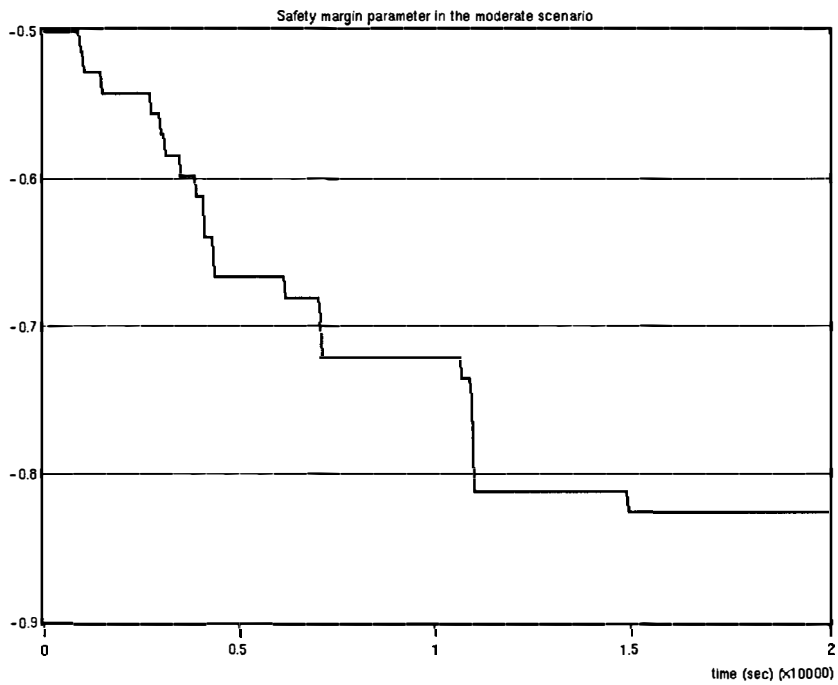


Figure 6.16: Safety margin in the moderate scenario

where on-off sources are used. This simulation again demonstrates that the closed-loop CAC scheme is able to overcome the restriction of all these model errors to achieve higher network resources utilization.

Using simulation and analysis, we showed that the measurement-based CAC scheme reported in Chapter 4 is robust against inaccuracies in declared traffic parameters. This property is due to the measurement-based nature of the CAC scheme reported in Chapter 4. The closed-loop CAC scheme proposed in this chapter is based on the measurement-based CAC scheme shown in Chapter 4. The only difference is that a closed-loop architecture is introduced to control the safety margin parameter. Therefore, the property of the measurement-based CAC scheme that it is robust against inaccuracies in declared traffic parameters is inherited by the closed-loop CAC scheme. The closed-loop CAC scheme is also robust against inaccuracies in declared traffic parameters. This is verified by our simulation.

Table 6.3: Traffic rate of the M-JPEG encoded movies ($bytes/frame$)

Type	Name	peak rate	mean rate
1	Sleepless in Seattle	16617	9477.6
2	Crocodile Dundee	19439	10772.9
3	Home Alone, II	22009	11382.8
4	Jurassic Park	23883	11363.0
5	Rookie of the Year	27877	12434.9
6	Speed	29385	12374.4
7	Hot Shots, Part Duex	29933	12766.1
8	Beauty and the Beast	30367	12661.5

6.5 Summary

Based on our work on measurement-based CAC scheme and in-service QoS monitoring algorithm, we have proposed a novel closed-loop CAC scheme in this chapter. The closed-loop architecture employs the CLR estimate from the in-service QoS monitoring and estimation algorithm as an input, and gives as its output the safety margin parameter of the CAC scheme. By controlling the safety margin parameter, the closed-loop architecture controls the call admission and bandwidth allocation of the CAC scheme. Due to the difficulty in obtaining an analytical relationship between the CLR and the safety margin parameter, fuzzy knowledge based controller is employed. FKBC provides a convenient framework to incorporate human experts' knowledge.

Simulation was carried out to test the performance of the closed-loop architecture. Simulation results showed that the proposed closed-loop CAC scheme is able to overcome the restrictions of model errors which are inherent drawbacks of an open-loop CAC scheme. Specifically, our CAC scheme is based on the bufferless fluid flow model and on-off traffic source model. By employing the closed-loop architecture, the closed-loop CAC scheme is able to overcome the model errors of the bufferless fluid flow model and on-off traffic source model, and efficiently utilize network resources even in the case when the buffer size is large and/or traffic source deviates from on-off source. Simulation results have

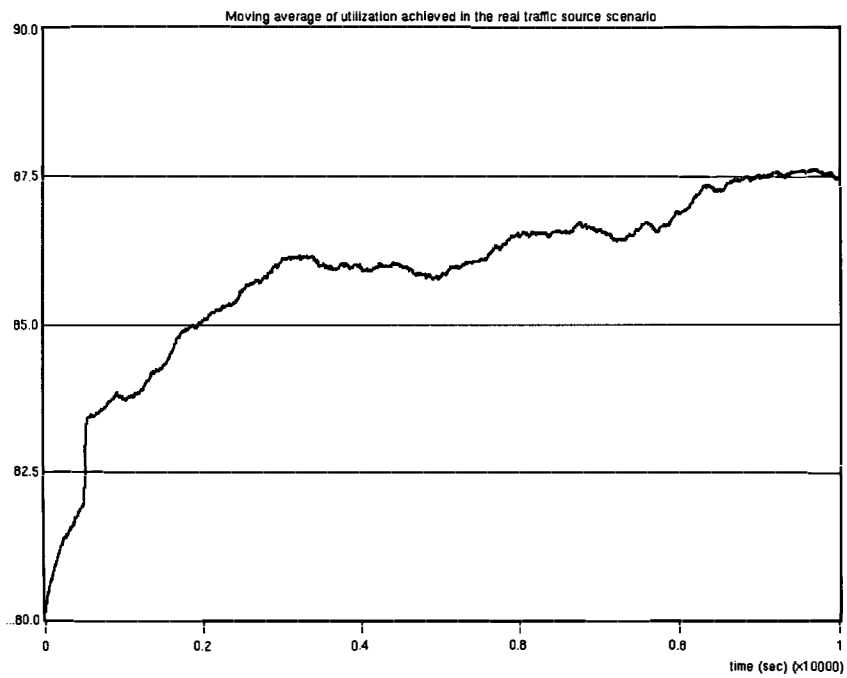


Figure 6.17: Moving average of utilization achieved in the real traffic source scenario

shown that the closed-loop CAC scheme is able to achieve a significantly higher network resources utilization than that achieved by the open-loop CAC scheme while still be able to provide robust QoS guarantee to network connections.

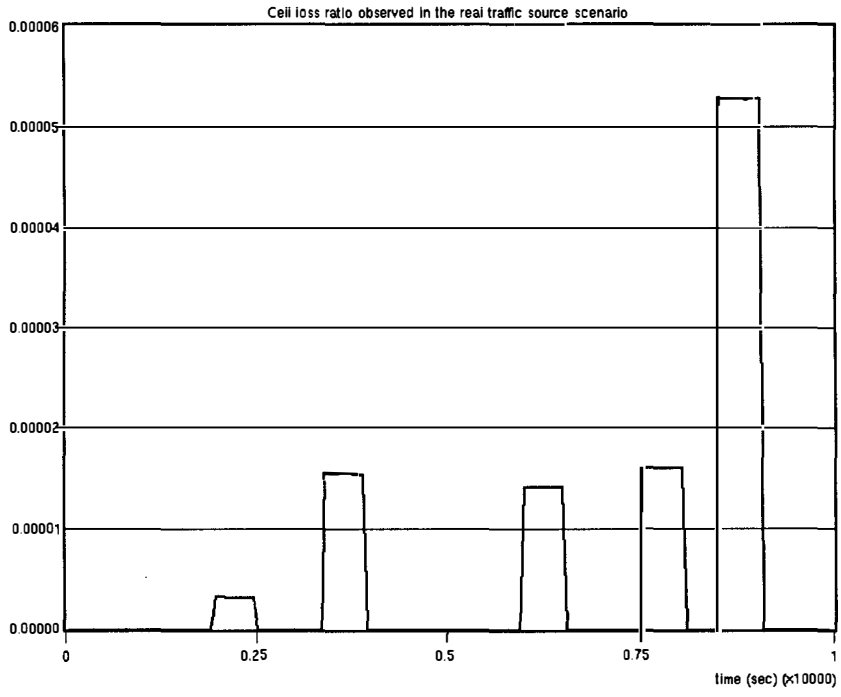
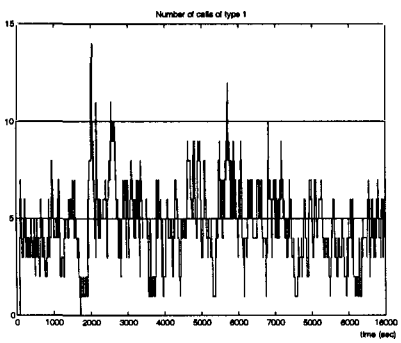
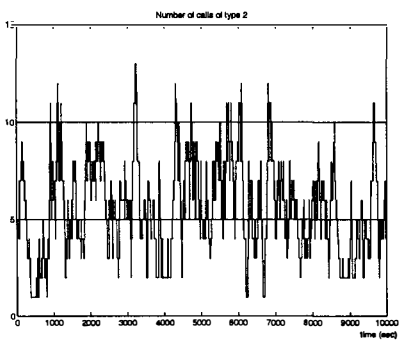


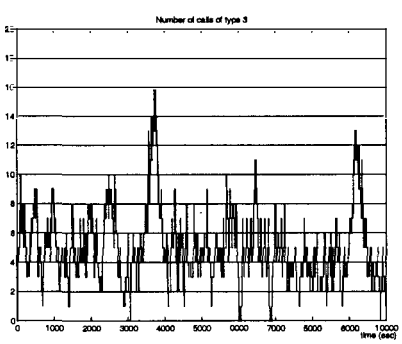
Figure 6.18: CLR observed in the real traffic source scenario



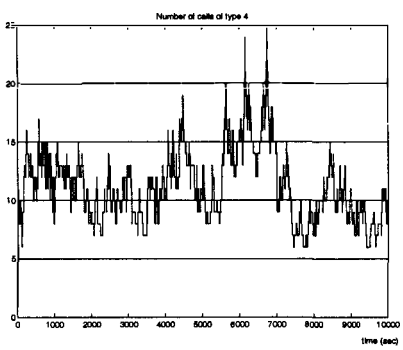
(a) Call number of traffic type 1



(b) Call number of traffic type 2

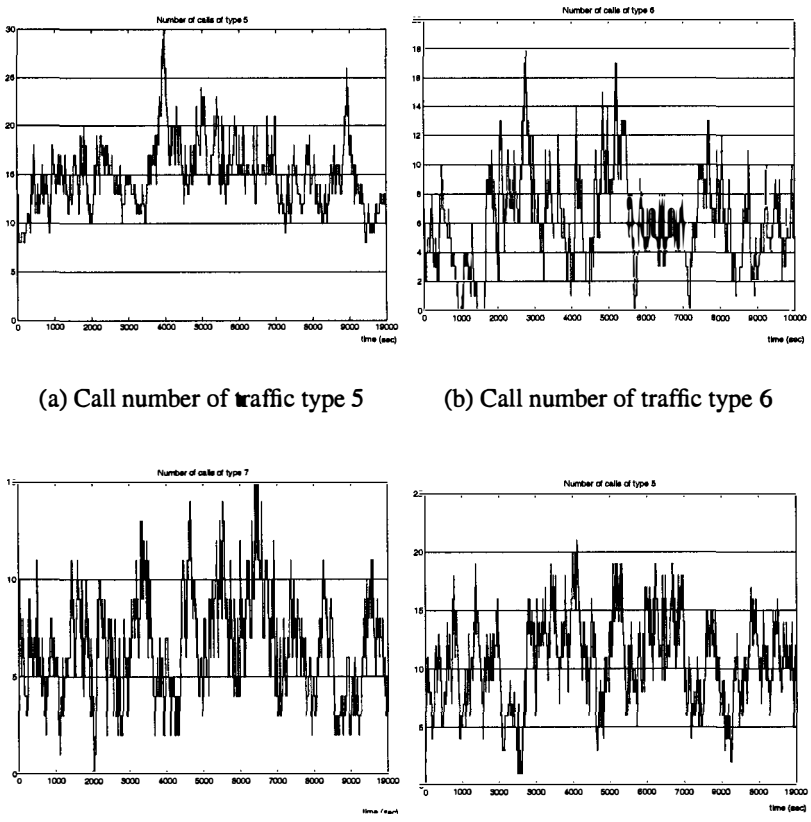


(c) Call number of traffic type 3



(d) Call number of traffic type 4

Figure 6.19: Number of each type of calls multiplexed on the link in the real traffic source scenario Part I



(a) Call number of traffic type 5 (b) Call number of traffic type 6

(c) Call number of traffic type 7 (d) Call number of traffic type 8

Figure 6.20: Number of each types of calls multiplexed on the link in the real traffic source scenario Part II

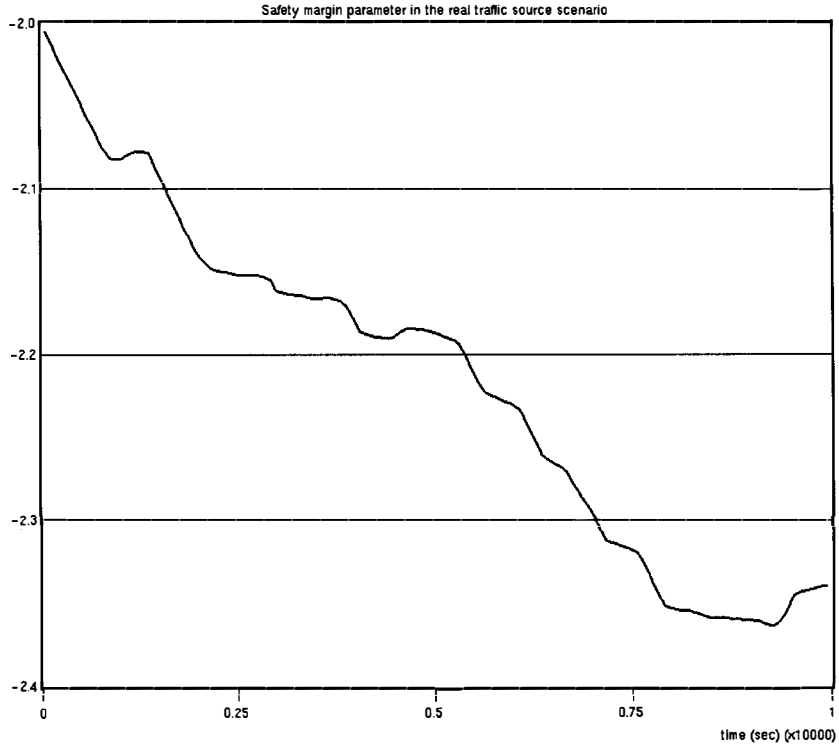


Figure 6.21: Safety margin in the real traffic source scenario

Chapter 7

Conclusion

In this thesis, we investigated the statistical characteristics and loss performance of heterogeneous network traffic. A measurement-based CAC scheme was proposed which is able to achieve higher utilization than those using effective bandwidth approach while at the same time provide robust QoS guarantees. The proposed CAC scheme is also robust against inaccuracies in user declared traffic parameters. Moreover, we analyzed the characteristics of the CAC schemes in the literature and pointed out that almost all of these CAC schemes employ open-loop control. CAC scheme based on the open-loop architecture lacks the ability to adjust its performance to account for model errors, which are inevitable in ATM networks due to the complexity and heterogeneity of network traffic. Therefore we proposed a novel closed-loop architecture for implementing the CAC scheme. An in-service QoS monitoring algorithm was developed which is employed to provide feedback to the closed-loop CAC scheme on its performance. A fuzzy knowledge based controller was designed which uses the estimated CLR from the in-service QoS monitoring algorithm as input. The FKBC controls the performance of the CAC scheme to compensate for model errors. Simulation indicated that the closed-loop architecture is able to overcome the inherent drawbacks of the open-loop architecture and achieve significantly higher network resources utilization. More importantly, our research revealed a new direction in the development of CAC schemes. We believe that this is a promising approach to solve the problem of

CAC in ATM networks.

This chapter summarizes the contributions of this thesis and recommends future research work.

7.1 Contributions of The Thesis

The main contributions of this thesis in brief are:

- First, loss performance analysis of heterogeneous traffic sources was performed. A new methodology was proposed to investigate the loss performance of heterogeneous traffic sources. An efficient and effective means of investigating loss performance of heterogeneous traffic sources in the bufferless fluid flow model was developed. Specifically, a cell loss rate function was defined and used to characterize the loss performance of traffic sources in the bffm. The clrf significantly simplifies theoretical analysis as well as computation.

  Stochastic ordering theory was used to analyze the clrf, The introduction of the stochastic ordering theory not only simplifies the theoretical analysis but also makes it possible to extend applications of our theoretical analysis to a broader area. That is, our loss performance analysis not only applies to cell loss in the bffm, but it can also be applied to any model whose cell loss function is a convex function of the traffic rate.

  A set of theorems were proposed and proved which have great significance in both loss performance analysis, and connection admission control scheme design. Moreover we proposed a cell loss upper bound for heterogeneous on-off sources. The proposed upper bound is tighter than those in the literature. It was shown that the proposed upper bound for heterogeneous on-off sources can be extended to traffic sources with any traffic rate distribution.

- Second, the application of our loss performance analysis in CAC scheme design was presented. With due considerations to practical applications of

a measurement-based admission control (MBAC) scheme, some principles were proposed which govern the design of a MBAC. A MBAC was designed which only needs simple traffic parameters from traffic sources, i.e. peak cell rate and/or mean cell rate. The measurement scheme for the proposed MBAC is easy to implement, and only needs measurement of the mean cell rate of the link. Extensive simulation was carried out which is indicative of good performance of the MBAC. The proposed MBAC was shown to be robust with regard to QoS guarantees, robust against inaccuracies in user-declared traffic parameters, and capable of achieving a high link utilization. All these features make the proposed MBAC an attractive option for real implementation.

Real vbr Motion-JPEG encoded video traffic sources were also used in the simulation. Statistical analysis was performed which showed that these vbr video sources present significant self-similar behavior. Simulation results indicated that despite the self-similarity in these traffic sources, the proposed measurement-based CAC scheme is able to provide robust QoS guarantees to these traffic sources. On the basis of this observation and research results in the literature, it was pointed out that the *critical time scale* widely exists in the measurement-based CAC scheme. Traffic fluctuations slower than the critical time scale have no impact on the performance of a measurement-based CAC. The existence of the critical time scale compromises the effect of self-similarity, which enables the proposed measurement-based CAC scheme to provide QoS guarantees to long-range dependent traffic.

- Third, based on the asymptotic relationship between cell loss ratio and buffer size, an in-service QoS monitoring and estimation (ISME) scheme was proposed which is able to provide on-line CLR estimates. Virtual buffer techniques were used in the ISME scheme. The proposed ISME scheme using virtual buffers significantly reduces the monitoring period required for making a valid QoS estimation. Simulation results showed that the proposed ISME scheme achieves better performance than those in the literature.

- Finally, a novel closed-loop architecture for connection admission control was proposed based on the proposed MBAC and ISME. It was pointed out that there are fundamental drawbacks in CAC schemes using open-loop architecture. These drawbacks affect the performance of a CAC scheme, and cannot be solved within the framework of open-loop architecture itself. Therefore a novel closed-loop architecture was proposed for implementing the CAC scheme. The ISME was employed to provide feedback to the closed-loop CAC scheme. A fuzzy knowledge based controller was designed which utilizes the estimated CLR from the in-service QoS monitoring algorithm as input. The FKBC controls the performance of the CAC scheme by controlling the safety margin parameter in the CAC scheme. Simulation indicated that the closed-loop architecture is able to overcome the inherent drawbacks of the open-loop architecture and achieve significantly higher network resources utilization than that using the open-loop architecture, while still able to provide robust QoS guarantees to network connections. More importantly, our research revealed a new direction towards the development of CAC scheme which is a promising approach to solve the problem of CAC in ATM networks.

7.2 Future Work

In this section we provide some suggestions for extensions to the work presented in this thesis.

7.2.1 Future Work on Traffic Measurements

The CAC scheme proposed in this thesis is a measurement-based CAC scheme. By employing traffic measurements, the proposed CAC scheme relieves the pressure on traffic sources to tightly characterize their traffic parameters.

The proposed MBAC only needs measurement of mean cell rate (mcr) of the aggregate traffic. Measurement of mcr is easier than measurement of other statistical quantities. However accurate measurement of mcr is not a simple task.

- First, the choice of measurement window size remains an open problem. In this thesis the measurement window size was chosen empirically. It is desirable that the choice of measurement window size is formulated. This entails statistical analysis of the characteristics of real network traffic, which include traffic burst length, connection holding time, call level dynamics, etc.

- Second, the method of obtaining an accurate mcr estimate from traffic measurements needs improvement. In the proposed CAC scheme, connection admission in the measurement window is delayed in order to eliminate the impact of call level dynamics on the mcr measurement. Although an alternative method was proposed which does not need to delay connection admission, the alternative method requires that the sum of declared mean cell rates of connections admitted during the measurement window is added to the estimated mean cell rate, which may result in over estimation of mcr. In real applications, an mcr estimation algorithm is desirable which does not affect connection admission while at the same time is able to provide an accurate estimate of mcr. In order to realize it, the mcr estimation algorithm needs to take into account call level dynamics explicitly.

7.2.2 Future Work on FKBC

A fuzzy knowledge based controller was used in the proposed closed-loop CAC scheme to control the CAC scheme based on the feedback information provided by the in-service QoS monitoring and estimation algorithm. Simulation results showed that the FKBC is able to control the CAC scheme to achieve much higher utilization than that achieved by the open-loop CAC scheme, while at the same time guarantee the QoS requirements of all connections. However there is still further improvement that can be done on the FKBC.

At this stage, the design of the FKBC is highly empirical. This makes it difficult to implement the proposed closed-loop CAC scheme practically. Moreover, it is difficult to obtain the optimum FKBC through empirical design. Therefore,

it is desirable that an adaptive fuzzy knowledge based controller [126], [127] be employed. A properly designed adaptive fuzzy knowledge knowledge based controller is able to automate the training process and come up with the optimum FKBC. This entails formulating the control objective and optimizing the training process. This problem remains open for future work.

7.2.3 Future Work on Engineering The Buffer Size

In real networks, in addition to cell loss ratio, cell transfer delay and cell delay variation are also important QoS parameters. In this thesis, following the main stream of CAC research, we used cell loss ratio as the QoS index and assumed that other parameters can be controlled within the desired bound by engineering the buffer size.

In network management, choosing an optimum network buffer size to satisfy the QoS requirements is not a simple task. A lot of factors have to be considered which include QoS requirements of connections, network topology, costs, etc. Further study is required to solve this problem.

Bibliography

[1] The ATM Forum, *ATM User-Network Interface Specification Version 3.1*, Sept. 1994.

[2] J. Burgin and D. Dorman, "Broadband ISDN Resource Management: The Role of Virtual Path," *IEEE Communications Magazine*, vol. 29, no. 9, pp. 44–48, Sept. 1991.

[3] V. J. Friesen, J. J. Harms, and J. W. Wong, "Resource Management with Virtual Paths in ATM Networks," *IEEE Network*, vol. 10, no. 5, pp. 10–19, Sept. 1996.

[4] W. Stallings, *ISDN and Broadband ISDN with Frame Relay and ATM*, Prentice Hall, Upper Saddle River, New Jersey 07458, 4th edition, 1999.

[5] R. Jain, "Congestion Control in Computer Networks: Issue and Trends," *IEEE Network Magazine*, vol. 4, no. 3, pp. 24–30, May 1990.

[6] H. Gilbert, O. Anoul-Magd, and V. Phung, "Developing a Cohesive Traffic Management Strategy for ATM Networks," *IEEE Communications Magazine*, vol. 29, no. 10, pp. 36–45, Oct. 1991.

[7] J. W. Roberts, "Traffic Control in the B-ISDN," *Computer Networks and ISDN Systems*, vol. 25, no. 10, pp. 1055–1064, May 1993.

[8] Z. -L. Zhang, J. Kurose, J. D. Salehi, and D. Towsley, "Smoothing, Statistical Multiplexing, and Call Admission Control for Stored Video," *IEEE*

Journal on Selected Areas in Communications, vol. 15, no. 6, pp. 1148–1166, Aug. 1997.

[9] P. Newman, "Backward Explicit Congestion Notification for ATM Local Area Networks," in *IEEE Globecom'93*, Geneva, 1993, vol. 2, pp. 719–723.

[10] K. Kawahara, Y. Oie, M. Murata, and H. Miyahara, "Performance Analysis of Reactive Congestion Control for ATM Networks," *IEEE Journal on Selected Areas in Communications*, vol. 13, no. 4, pp. 651–661, May 1995.

[11] J. W. Roberts, "LAN-LAN Interconnection and B-ISDN," Cost 242 doc. td(92)67, 1992.

[12] A. Gersht and K. J. Lee, "A Congestion Control Framework for ATM Networks," *IEEE Journal on Selected Areas in Communications*, vol. 9, no. 7, Sept. 1991.

[13] G. Ramamurthy and R. S. Dighe, "A Multidimensional Framework for Congestion Control in B-ISDN," *IEEE Journal on Selected Areas in Communications*, vol. SAC-9, no. 9, pp. 1440–1451, 1991.

[14] J. Y. Hui, "Resource Allocation for Broadband Networks," *IEEE Journal on Selected Areas in Communications*, vol. 6, no. 9, pp. 1598–1608, 1988.

[15] M. Grossglauser and D. N. C. Tse, "A Time-Scale Decomposition Approach to Measurement-Based Admission Control," in *IEEE INFOCOM'99*, 1999, vol. 3, pp. 1539–1547.

[16] 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, Aug. 1995.

[17] Y. Kim and S. Q. Li, "Timescale of Interest in Traffic Measurement for Link Bandwidth Allocation Design," in *IEEE INFOCOM'96*, Apr. 1996, vol. 2, pp. 738–748.

[18] J. Roberts, U. Mocci, and J. Virtamo, *Broadband Network Teletraffic*, vol. 1155, Springer, 1 edition, 1996.

[19] N. Sato, K. Asatani, H. Murakami, R. E. Mallon, S. R. Hughes, and T. L. Groff, "In-Service Monitoring Methods - Better Way to Assure Quality of Digital Transmission," *IEEE Journal on Selected Areas in Communications*, vol. 12, no. 2, pp. 355–360, Feb. 1994.

[20] A. E. Eckberg, "B-ISDN/ATM Traffic and Congestion Control," *IEEE Network Magazine*, vol. 6, no. 5, pp. 28–37, Sept. 1992.

[21] T. -H. Lee, K. -C. Lai, and S. -T. Duann, "Design of A Real-Time Call Admission Controller for ATM Networks," *IEEE/ACM Transaction on Networking*, vol. 4, no. 5, pp. 758–765, Oct. 1996.

[22] S. Lee and J. Song, "A Measurement-Based Admission Control Algorithm Using Variable-Sized Window in ATM Networks," *Computer Communications*, vol. 21, no. 2, pp. 171–178, 1998.

[23] J. T. Lewis, R. Russell, F. Toomey, B. McGurk, S. Crosby, and I. Leslie, "Practical Connection Admission Control for ATM Networks Based on On-Line Measurements," *Computer Communications*, vol. 21, no. 17, pp. 1585–1596, 1998.

[24] M. Reisslein, K. W. Ross, and S. Rajagopal, "Guaranteeing Statistical QoS to Regulated Traffic: The Single Node Case," in *IEEE INFOCOM'99*, Apr. 1999, vol. 3, pp. 1061–1072.

[25] S. Lee and J. Song, "An Adaptive Call Admission Control in ATM Networks Using Optimized Measurements Windows," *Computer Communications*, vol. 23, pp. 863–870, 2000.

[26] S. Jamin Sugih, P. B. Danzig, S. J. Shenker, and L. Zhang, "A Measurement-based Admission Control Algorithm for Integrated Service Packet Networks," *IEEE/ACM Transactions on Networking*, vol. 5, no. 1, pp. 56–70, 1997.

[27] D. Mitra, M. I. Reiman, and J. Wang, "Robust Dynamic Admission Control for Unified Cell and Call QoS in Statistical Multiplexers," *IEEE Journal on Selected Areas in Communications*, vol. 16, no. 5, pp. 692–707, June 1998.

[28] H. Che and S. Q. Li, "Fast Algorithms for Measurement-Based Traffic Modeling," *IEEE Journal on Selected Areas in Communications*, vol. 16, no. 5, pp. 612–625, 1998.

[29] K. Shiomoto and S. -I. Chaki, "Adaptive Connection Admission Control Using Real-Time Traffic Measurements in ATM Networks," *IEICE Transactions on Communications*, vol. E78-B, no. 4, pp. 458–464, 1995.

[30] R. Bolla, F. Davoli, and M. Marchese, "Bandwidth Allocation and Admission Control in ATM Networks with Service Separation," *IEEE Communications Magazine*, vol. 35, no. 5, pp. 130–137, 1997.

[31] D. Anick, D. Mitra, and M. M. Sondhi, "Stochastic Theory of Data-Handling Systems with Multiple Sources," *Bell System Technical Journal*, vol. 61, no. 8, pp. 1871–1893, 1982.

[32] R. J. Gibbens and P. J. Hunt, "Effective Bandwidths for Multi-Type UAS Channel," *Queueing System*, vol. 9, no. 1-2, pp. 17–28, 1991.

[33] R. Guerin, H. Ahmadi, and M. Naghshineh, "Equivalent Capacity and its Application to Bandwidth Allocation in High-Speed Networks," *IEEE Journal on Selected Area in Communication*, vol. 9, no. 7, pp. 968–981, 1991.

[34] 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, June 1993.

[35] F. P. Kelly, "Effective Bandwidths at Multi-Class Queues," *Queueing Systems*, vol. 9, no. 1-2, pp. 5–16, Oct. 1991.

[36] A. Shwartz and A. Weiss, *Large Deviations for Performance Analysis - Queues, Communications and Computing*, Chapman & Hall, London, 1995.

[37] 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.

[38] C. -S. Chang, "Stability, Queue Length, and Delay of Deterministic and Stochastic Queueing Networks," *IEEE Transaction on Automatic Control*, vol. 39, no. 5, pp. 913–931, May 1994.

[39] A. Elwalid, D. Heyman, T. V. Lakshman, D. Mitra, and A. Weiss, "Fundamental Bounds and Approximations for ATM Multiplexers with Applications to Video Teleconferencing," *IEEE Journal on Selected Areas in Communications*, vol. 13, no. 6, pp. 1004–1016, Aug. 1995.

[40] D. D. Botvich and N. G. Duffield, "Large Deviations, the Shape of the Loss Curve, and Economies of Scale in Large Multiplexers," *Queueing Systems*, vol. 20, no. 3-4, pp. 293–320, 1995.

[41] A. Simonian and J. Guibert, "Large Deviations Approximation for Fluid Queues Fed by a Large Number of on/off Sources," *IEEE Journal on Selected Areas in Communications*, vol. 13, no. 6, pp. 1017–1027, Aug. 1995.

[42] 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.

[43] A. W. Berger and W. Whitt, "Extending the Effective Bandwidth Concept to Networks with Priority Classes," *IEEE Communications Magazine*, vol. 36, no. 8, pp. 78–83, Aug. 1998.

[44] J. Bucklew, *Large Deviation Techniques in Decision, Simulation and Estimation*, New York : Wiley & Sons, 1990.

[45] E. Cinlar, *Superposition of Point Processes*, pp. 549–606, New York: Wiley, 1972.

[46] H. Heffes and D. M. Lucantoni, "A Markov Modulated Characterization of Packetized Voice and Data Traffic and Related Statistical Multiplexer Performance," *IEEE Journal on Selected Areas in Communications*, , no. SAC-4, pp. 856–868, 1986.

[47] K. Sriram and W. Whitt, "Characterizing Superposition Arrival Processes in Packet Multiplexers for Voice and Data," *IEEE Journal on Selected Areas in Communications*, vol. SAC-4, pp. 833–846, 1986.

[48] K. W. Fendick, V. R. Sakasena, and W. Whitt, "Investigating Dependence in Packet Queues with the Index of Dispersion for Work," *IEEE Transactions on Communications*, vol. 39, no. 08, pp. 1231–1244, Aug. 1991.

[49] W. Whitt, "Tail Probabilities with Statistical Multiplexing and Effecive Bandwidths for Multi-Class Queues," *Telecommunication Systems*, vol. 2, pp. 71–107, 1993.

[50] G. L. Choudhury, D. M. Lucantoni, and W. Whitt, "Squeezing the Most out of ATM," *IEEE Transactions On Communications*, vol. 44, no. 2, pp. 203–217, Feb. 1996.

[51] P. Billingsley, *Probability and Measure*, New York: Wiley, 2nd edition, 1986.

[52] T. Murase, H. Suzuki, S. Sato, and T. Takeuchi, "A Call Admission Control Scheme for ATM Networks Using a Simple Quality Estimate," *IEEE Journal on Selected Areas in Communications*, vol. 9, no. 9, pp. 1461–1470, Dec. 1991.

[53] A. Elwalid, D. Mitra, and R. H. Wentworth, "A New Approach for Allocating Buffers and Bandwidth to Heterogeneous, Regulated Traffic in an ATM Node," *IEEE Journal on Selected Areas in Communications*, vol. 13, no. 6, pp. 1115–1127, Aug. 1995.

[54] H. Saito and K. Shiomoto, "Dynamic Call Admission Control in ATM Networks," *IEEE Journal on Selected Area in Communications*, vol. 9, no. 7, pp. 982–989, Sept. 1991.

[55] B. T. Doshi, "Deterministic Rule Based Traffic Descriptors for Broadband ISDN: Worst Case Behavior and Connecction Acceptance Control," in *Proc. ITC-14*, J. Labetoulle and Eds. J. Roberts, Eds. 1994, pp. 591–600, New Yprk: Elsevier.

[56] T. Worster, "Modeling Deterministic Queues: the Leaky Bucket as an Arrival Process," in *Proc. ITC-14*, J. Labetoulle and Eds. J. W. Roberts, Eds. 1994, pp. 581–590, New York: Elsevier.

[57] N. Yamanaka, Y. Sato, and K. I. Sato, "Performance Limitations of Leaky Bucket Algorithm for Usage Parameter Control of Bandwidth Allocation Methods," *IEICE Transactions on Communications*, vol. E74-N, no. 2, pp. 82–86, 1992.

[58] B. Lee, J. Sugih, and S. Scott, "Comments on the Performance of Measurement-Based Admission Control," in *IEEE INFOCOM'00*, 2000, vol. 3, pp. 1233–1242.

[59] 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, June 1999.

[60] 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.

[61] S. M. Srinidhi, W. H. Thesling, and V. K. Konangi, "An Adaptive Scheme for Admission Control in ATM Networks," *Computer Networks and ISDN Systems*, vol. 29, no. 5, pp. 569–582, 1997.

[62] H. Saito, "Call Admission Control in an ATM network Using Upper Bound Cell Loss Probability," *IEEE Transactions on Communications*, vol. 40, no. 9, pp. 1512–1521, Sept. 1992.

[63] S. Shioda and H. Saito, "Connection Admission Control Guaranteeing Negotiated Cell-Loss Ratio of Cell Streams Passing Through Usage Parameter Control," *IEICE Transactions on Communications*, vol. E80-B, no. 3, pp. 399–411, Mar. 1997.

[64] T. E. Tedijanto and L. Gun, "Effectiveness of Dynamic Bandwidth Management Mechanisms in ATM Networks," in *IEEE INFOCOM'93*, 1993, vol. 1, pp. 358–367.

[65] A. Baiocchi, N. B. Melazzi, M. Listanti, A. Roveri, and R. Winkler, "Loss Performance Analysis of an ATM Multiplexer Loaded with High-Speed On-Off Sources," *IEEE Journal on Selected Areas in Communications*, vol. 9, no. 3, pp. 388–393, Apr. 1991.

[66] 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.

[67] H. Zhu and V. S. Frost, "In-Service Monitoring for Cell Loss Quality of Service Violations in ATM Networks," *IEEE/ACM Transactions on Networking*, vol. 4, no. 2, pp. 240–248, Apr. 1996.

[68] 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, Feb. 1995.

[69] C. Courcoubetis, G. Kesidis, A. Ridder, J. Walrand, and R. Weber, "Admission Control and Routing in ATM Networks Using Inferences from Measured Buffer Occupancy," *IEEE Transactions on Communications*, vol. 43, no. 2/3/4, pp. 1778–1784, Feb. 1995.

[70] J. -S. Li, "Measurement and In-Service Monitoring for QoS Violations and Spare Capacity Estimations in ATM Network," *Computer Communications*, vol. 23, no. 2, pp. 162–170, 2000.

[71] B. Bensaou, S. T. C. Lam, H. -W. 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.

[72] W. E. Leland, M. S. Taqqu, W. Willinger, and D. V. Wilson, "On the Self-Similar Nature of Ethernet Traffic (Extended Version)," *IEEE/ACM Transactions on Networking*, vol. 2, no. 1, pp. 1–15, Feb. 1994.

[73] 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, no. 2/3/4, pp. 1566–1579, Feb. 1995.

[74] 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, June 1995.

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

[76] D. P. Heyman and T. V. Lakshman, "What Are the Implications of Long-Range Dependence for VBR-Video Traffic Engineering?," *IEEE/ACM Transactions on Networking*, vol. 4, no. 3, pp. 301–317, June 1996.

[77] W. Willinger, M. S. Taqqu, R. Sherman, and D. V. Wilson, "Self-Similarity Through High-Variability: Statistical Analysis of Ethernet Lan Traffic at the Source Level," *IEEE/ACM Transactions on Networking*, vol. 5, no. 1, pp. 71–86, Feb. 1997.

[78] A. Erramilli, O. Narayan, and W. Willinger, "Experimental Queueing Analysis with Long-Range Dependent Packet Traffic," *IEEE/ACM Transcations on Networking*, vol. 4, no. 2, pp. 209–223, Apr. 1996.

[79] M. E. Crovella and A. Bestavros, "Self-Similarity in World Wide Web Traffic: Evidence and Possible Causes," *IEEE/ACM Transactions on Networking*, vol. 5, no. 6, pp. 835–846, Dec. 1997.

[80] P. Pruthi and A. Erramilli, "Heavy-Tailed On/Off Source Behavior and Self-Similar Traffic," in *IEEE ICC'95*, 1995, vol. 1, pp. 445--450.

[81] I. Norros, "A Storage Model with Self-Similar Input," *Queueing Systems*, vol. 16, pp. 387–396, 1994.

[82] P. L. Conti, H. Saito, and L. D. Giovanni, "A Robust Connection Admission Control Applicable to Long Range Dependence Traffic," *IEICE Trans. Commun.*, vol. E81-B, no. 5, pp. 849–857, May 1998.

[83] O. J. Boxma, "Fluid Queues and Regular Variation," *Performance Evaluation*, vol. 27&28, pp. 699–712, 1997.

[84] D. Heath, S. Resnick, and G. Samorodnitsky, "Heavy Tails and Long Range Dependence in On/Off Processes and Associated Fluid Models," *Mathematics of Operations Research*, vol. 23, no. 1, pp. 145–165, Feb. 1998.

[85] M. S. Taqqu, W. Willinger, and R. Sherman, "Proof of a Fundamental Result in Self-Similar Traffic Modeling," *Computer Communication Review*, vol. 27, no. 02, pp. 5–23, 1997.

[86] P. R. Jalenkovic, A. A. Laznr, and N. Semret, "The Effect of Multiple Time Scales and Subexponentiality in MPEG Video Streams on Queueing Behavior," *IEEE Journal on Selected Areas in Communications*, vol. 15, no. 6, pp. 1–16, Aug. 1997.

[87] R. Roughan, D. Veitch, and P. Abry, "Real-Time Estimation of the Parameters of Long-Range Dependence," *IEEE/ACM Transactions on Networking*, vol. 8, no. 4, pp. 467--478, Aug. 2000.

[88] V. Klemes, "The Hurst Phenonmenon: A Puzzle?," *Water Resources Research*, vol. 10, pp. 675–688, 1974.

[89] B. B. Mandelbrot, "Robust R/S Analysis of Long Run Serial Correlation," in *Proc. 42nd Session ISI, Book 2*, 1979, pp. 69–99.

[90] M. S. Taqqu, V. Teverovsky, and W Willinger, "Estimators for Long-Range Dependence: An Empirical Study," *Fractals*, vol. 3, no. 4, pp. 785–798, 1995.

[91] J. Beran, *Statistics for Long-Memory Processes*, Monographs on Statistics and Applied Probability. London: Chapman and Hall, 1994.

[92] P. J. Brockwell and R. A. David, *Time Series: Theory and Methods*, Springer Series in Statistics. New York:Springer-Verlag, 2nd edition, 1991.

[93] G. Babic, B. Vandalore, and R. Jain, "Analysis and Modeling of Traffic in Modern Data Communication Networks," Techical report, Department of Computer and Information, Ohio State University, Feb., 5 1998.

[94] K. R. Krishnan, "A New Class of Performance Results for Fractional Brownian Traffic Model," Bellcore tech. rep., Bellcore, 1995.

[95] M. W. Garrett, "Analysis, Modeling and Generation of Self-Similar VBR Video Traffic," in *ACM SIGCOMM'94*, Aug. 1994.

[96] M. Grossglauser and J. -C. Bolot, "On the Relevance of Long-Range Dependence in Network Traffic," *IEEE/ACM Transactions on Networking*, vol. 7, no. 5, pp. 629--640, Oct. 1999.

[97] B. K. Ryu and A. Elwalid, "The Importance of Long-Range Dependence of VBR Video Traffic in ATM Traffic Engineering: Myths and Realities," *Computer Communication Review*, vol. 24, no. 4, Oct. 1996.

[98] H. Yamada and S. Sumita, "A Traffic Measurement Method and its Application for Cell Loss Probability Estimation in ATM Networks," *IEEE Journal on Selected Areas in Communications*, vol. 9, no. 3, pp. 315–324, Apr. 1991.

[99] C. -L. Hwang and S. Q. Li, "On Input State Space Reduction and Buffer Noneffective Region," in *IEEE INFOCOM'94*, 1994, vol. 3, pp. 1018–1028.

[100] H. -D. Sheng and S. Q. Li, "Second Order Effect of Binary Sources on Characteristics of Queue and Loss Rate," *IEEE Transactions on Communications*, vol. 42, no. 2/3/4, pp. 1162–1173, Feb. 1994.

[101] W. -C. Lau and S. Q. Li, "Statistical Multiplexing and Buffer Sharing in Multimedia High-Speed Networks: a Frequency-Domain Perspective," *IEEE/ACM Transactions on Networking*, vol. 5, no. 3, pp. 382–396, 1997.

[102] L. A. Kulkarni and S. Q. Li, "Measurement-Based Traffic Modeling: Capturing Important Statistics," *Journal of Stochastic Model*, vol. 14, no. 5, 1998.

[103] 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.

[104] J. Choe and N. B. Shroff, "A New Method to Determine the Queue Length Distribution at an ATM Multiplexer," in *IEEE INFOCOM'97*, 1997, pp. 549–556.

[105] R. Szekli, *Stochastic Ordering and Dependence in Applied Probability*, Springer-Verlag New York, Inc., 1995.

[106] S. M. Ross, *Stochastic Processes*, John Wiley & Sons, Inc., 1983.

[107] Z. -L. Zhang, J. Jurose, J. Salehi, and D. Towsley, "Smoothing, Statistical Multiplexing and Call Admission Control for Stored Video," Umass cmpsci technical report um-cs-96-29, Department of Computer Science, University of Massachusetts, Amherst, 1996, Available via FTP from gaia.cs.umass.edu in pub/Zhan96:Smoothing.ps.gz.

[108] C. Rasmussen, J. H. Sorensen, K. S. Kvols, and S. B. Jacobsen, "Source-Independent Call Acceptance Procedures in ATM Networks," *IEEE Journal on Selected Areas in Communications*, vol. 9, no. 3, pp. 351–358, Apr. 1991.

[109] B. Maglaris, D. Anastassiou, P. Sen, G. Kaelsson, and J. D. Robbins, "Performance Models of Statistical Multiplexing in Packet Video Communications," *IEEE Transactions on Communications*, vol. 36, no. 7, pp. 834–844, 1988.

[110] K. Sohraby, "On the Theory of General ON-OFF Sources with Applications in High-Speed Networks," in *IEEE INFOCOM'93*, 1993, vol. 2, pp. 401--410.

[111] P. T. Brady, "A Statistical Analysis of On-Off Patterns in 16 Conversations," *Bell System Technical Journal*, pp. 73–91, 1968.

[112] R. L. Easton, P. T. Hutchinson, R. W. Kolor, R. C. Moncello, and R. W. Muise, "TASI-E Communication System," *IEEE Transactions in Communications*, vol. 30, pp. 803–807, 1982.

[113] S. Q. Li and C. L. Hwang, "Queue Response to Input Correlation Functions: Continuous Spectral Analysis," *IEEE/ACM Transactions on Networking*, vol. 1, pp. 678–692, Dec. 1993.

[114] N. G. Bean, *Statistical Multiplexing in Broadband Communication*, Ph.D. thesis, Univ. Cambridge, Cambridge, 1993.

[115] R. Griffiths and P. Key, "Adaptive Call Admission Control in ATM Networks," in *Proc. 14th Int. Teletraffic Cong.-ITC 14*, June 1994.

[116] M. Fontaine and D. G. Smith, "Bandwidth Allocation and Connection Admission Control in ATM Networks," *Electronics and Communication Engineering Journal*, vol. 8, no. 4, pp. 156–164, 1996.

[117] J. W. Roberts, "Variable-Bit-Rate Traffic Control in B-ISDN," *IEEE Communications Magazine*, vol. 29, no. 9, pp. 50–56, Sept. 1991.

[118] W. -C. Feng, *Video-on-Demand Services: Efficient Transportation and Decompression of Variable Bit Rate Video*, Ph.D. thesis, Univ. of Michigan, Apr. 1996, Available at http://www.cis.ohio-state.edu/ wuchi/.

[119] J. Garcia-Hernandez and M. Ghanbari, "In-Service Monitoring of Quality of Service in ATM Networks Using OAM Cells," *IEE Proceedings-Communications*, vol. 146, no. 2, pp. 102–106, Apr. 1999.

[120] N. Likhanow, B. Tsybakov, and N. D. Georganas, "Analysis of an ATM buffer with Self-Similar ("Fractal") Input Traffic," in *IEEE INFOCOM'95*, 1995, vol. 3, pp. 985–992.

[121] M. Parulekar and A. M. Makowski, "Tail Probabilities for a Multiplexer with Self-Similar Traffic," in *IEEE INFOCOM'96*, Mar. 1996, vol. 3, pp. 1452–1459.

[122] D. R. Cox, *Long-Range Dependence: A Review*, pp. 55–74, The Iowa State University, 1984.

[123] M. W. Garrett, *Contributions Towards Real-Time Services on Packet Switched Networks*, Ph.D. thesis, Columbia University, May 1993.

[124] A. Wong, C. -T. Chen, D. J. L. Gall, F. -C. Heng, and K. M. Uz, "MCPIC: A Video Coding Algorithm for Transmission and Storage Applications," *IEEE Communications Magazine*, vol. 28, no. 11, pp. 24–32, Nov. 1990.

[125] D. Driankov, H. Hellendoorn, and M. Reinfrank, *An Introduction to Fuzzy Control*, Springer-Verlag, 2 edition, 1996.

[126] L. -X. Wang, *Adaptive Fuzzy Systems and Control - Design and Stability Analysis*, PTR Prentice Hall, Englewood Cliffs, New Jersey 07632, 1994.

[127] L. X. Wang, *A Course in Fuzzy Systems and Control*, Prentice Hall, Englewood Cliffs, NJ, 1997.