The University of Kansas

Information and Telecommunication Technology Center

Technical Report

Dynamic Source-Coding Rate Control Scheme for ATM Adaptation Layer Type 2

Sampat Sreepathi-Komanduri
David W. Petr

ITTC-FY99-TR-15664-02

January 1999

Project Sponsor:
Sprint Corporation

Abstract

Asynchronous Transfer Mode has been used to implement high speed networks providing multi-gigabit services for multimedia applications. The introduction of new delay sensitive applications, like Voice and Telephony Over ATM (VTOA), necessitated the modification of ATM protocol aspects in order to provide bandwidth-efficient transmission of low-rate, short and variable length packets and high utilization of the limited bandwidth. ATM Adaptation Layer 2 (AAL2) is the result of this modification.

The complexity of traffic management increases with the addition of new services. In this project, one of the traffic management issues called connection admission control (CAC) is dealt in the context of AAL2. ATM level CACs are modified to suit in an AAL2 framework. The results of analysis were compared with simulations done in a simulation package (BONeS).

A Dynamic source coding Rate Control (DRC) scheme is proposed for AAL2 which helps in increasing the link utilization. In this scheme the source coding rate is reduced momentarily in response to congestion.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Recent years have witnessed tremendous growth in the field of telecommunications. One of the key enabling technologies is the deployment of optical fiber which made wider bandwidths (higher bit rates) available to the user. Due to the availability of wider bandwidths, man was encouraged to plan about new user services like video-on-demand, video telephony, video library, high speed data, etc. The introduction of new communication services, especially in packet-switched networks which are used primarily to transmit data of various types, is giving rise to increase in the volume of traffic flowing through the network. The pace of increase in the volume of data traffic is much faster than the pace of increase in voice or telephony traffic carried by legendary circuit-switched networks. Looking at the trends, the telecommunications industry felt the need to have a common connectivity network instead of running different services on different networks. The need of integration of services resulted in the introduction of *broadband integrated services digital network* (B-ISDN) by *International Telecommunications Union* (ITU), formerly known as *Consultative Committee for Telephone and Telegraph* (CCITT). The key design objective of B-ISDN is "The provision of wide range of services to a broad variety of users utilizing a limited set of connection types and multi-purpose user-network interfaces" [11, 5].

One of the enabling technologies for deployment of broadband infrastructure is *asynchronous transfer mode* (ATM) chosen by ITU as the target transfer mode. The final goal is to have ATM networks serve as universal transport and/or switching backbones for a wide variety of applications including voice, data and video as well as other multimedia interactive communications.

ATM is a mixture of packet-switching and circuit-switching technologies. It is packet switching in the sense that all the traffic is transported via fixed size (53 octet) cells and these cells are relayed and routed through the network based on information contained in the cell header. It is partly circuit-switching because it is connection-oriented service.

ATM promises to deliver a cost-effective solution for heterogeneous types of services by promoting dynamic network resource sharing, supporting bandwidth on demand and exploiting the statistical multiplexing property of aggregated traffic. Before such promise can be fulfilled, many problems have to be solved. One of the key challenges is traffic management of ATM networks. In this thesis, we are looking at one of the aspects of traffic and resource management in ATM networks for voice transport.

1.1 Motivation and Contribution

In traditional circuit-switched networks every connection is assigned a dedicated circuit of fixed bandwidth via *synchronous time division multiplexing* (STDM). This results in minimal interaction among traffic streams corresponding to different connections. However, this is no longer true for ATM networks which aggregate all the traffic coming from different connections in order to obtain bandwidth savings (statistical multiplexing gain). As a result, traffic from one connection affects the *quality of service* (QoS) of all other connections sharing common resources. In order to ensure the QoS of all users some additional resource management schemes are needed. One of such schemes is called *connection admission control* or *call admission control* (CAC). CAC deals with the question of whether or not to accept a new connection request with certain traffic characteristics and QoS requirements. CAC, on receipt of a new connection request, calculates the amount of network resources required to be allocated to this new connection to support its QoS. If enough resources are available with the network, the new connection is accepted; otherwise it is rejected. In other words, CAC calculates the *maximum* number of users whose QoS can be satisfied with given set of network resources. Over the last few years, many CAC schemes were proposed [2, 3, 12, 13, 5, 14, 15, 1] and [5] gives a nice overview of some of these CAC schemes.

New communication services require new or modified CAC schemes in order to do a better estimation of the network resources required to support a new connection request.

The current trends indicates that the voice and telephony traffic which represents the largest volume in telecommunications networks, both in terms of traffic and revenue generated, is moving from being circuit-switched to being packet-switched. The network service providers (NSPs) have to take care that this transition is done smoothly, without any negative impact on feature of the richness and reliability of this service. A new ATM adaptation layer (AAL-2) was standardized in early 1997 to achieve this objective and to support delay sensitive applications like *voice and telephony over ATM* (VTOA). AAL-2 provides bandwidth efficient transmission of low-rate, short, and variable length packets in delay sensitive applications like voice [10]. Just like in any other application using packet-switched technology, VTOA also requires traffic management to ensure QoS of all the users.

In this thesis, existing CAC schemes are modified in order to make them suitable to be used in an AAL-2 framework. The results of the analysis are comparable to the simulation results. In the work presented here, we have selected the nature of the ATM *virtual circuit*(VC) to be *constant bit rate* as opposed to the alternative of using a *variable bit rate* VC. In chapter 3 we have presented the justification for using CBR VC in the framework of AAL-2. From the results of analysis and simulations, it is found that a significant amount of bandwidth is wasted by static CAC schemes.

To overcome this problem, a dynamic source coding rate control scheme is proposed which sits in the AAL-2 transmitter and depending on the packet queue length, changes the coding rate of the source. This can be done with zero feedback delay whenever the AAL-2 coder sits in the AAL-2 transmitter . It is found by simulations that this scheme can achieve significant amount of gain with negligible-but-controllable degradation of user QoS.

1.2 Organization of thesis

Chapter 2 gives the required background and reviews some of the recent work done in the field of congestion control in ATM networks.

- Section 2.1 presents a brief background about ATM.
- In section 2.2, we discuss B-ISDN and its layers. This section provides a brief introduction of all the *ATM adaptation layers* (AALs).
- Section 2.3 presents the concepts and design of AAL2.
- In Section 2.4 we discuss different service classes and section 2.5 presents different QoS categories.
- Section 2.6 provides an overview of traffic management and congestion control options in ATM networks.

Chapter 3 discusses the fundamental question of whether the virtual circuits (VCs) for AAL2 should be CBR or VBR.

In chapter 4, a few CAC schemes which were proposed for ATM are mapped to be used in an AAL-2 context. These schemes are then evaluated using analysis and simulations.

To suit the need of AAL2 and overcome the limitations of any of the static CAC schemes mentioned in chapter 4, a dynamic source-coding rate control scheme is defined and the results of this scheme are presented in chapter 5. Conclusions and scope of extensions are presented in chapter 6.

Chapter 2

Background

2.1 Asynchronous Transfer Mode (ATM)

ATM has been adopted as the transfer mode to provide integration of various traffic types to be supported by B-ISDN. ATM is a high-speed packet-like switching and multiplexing technique. All the information to be transferred across the network is segmented into fixed sized cells (of 53 bytes length) which are then transported across the network depending on the information contained in the header of each cell in a manner similar to packet switching. ATM has three important features which are not present in the traditional synchronous transfer mode (STM). The first one is that the source information rate is not constrained physically by the system except for the maximum link speed. This allows integration of different systems with different and variable information rates to share same link. The second important feature is that, if one of the sources is not generating cells, the corresponding bandwidth can used by other sources which are sharing the same link and are generating cells. This effect gives rise to the so-called statistical multiplexing gain which can provide increased bandwidth utilization when compared with systems where a fixed bandwidth is allocated to each connection. The third feature is the uniform cell format of ATM cell which simplifies broadband switch architectures since the bandwidth required by the connection does not influence the switching algorithm [21]. The concept of ATM can be well explained using figure 2.1. Here the ATM switching nodes are connected by means of ATM links. The users are connected to the ATM switches either directly or via multiplexers in order to increase access link utilization. Cells are routed via the network using the tag included in the cell header. The tag has two sections associated with *virtual path* (VP) and *virtual circuit* concepts: a *virtual path identifier* (VPI) and a *virtual circuit identifier* (VCI). A particular VPI (or VCI) defines a *virtual path link* (VPL) (or a *virtual channel link* (VCL)) which corresponds to a transport of all ATM cells with common VPI (or VCI) between a point where the VPI (or VCI) is assigned and the point where that value is translated or terminated. A *virtual path connection* (VPC) (or *virtual channel connection* (VCC)) is defined by a concatenation of VP links (or VC links) and can have different VPIs (or VCIs) on different ATM

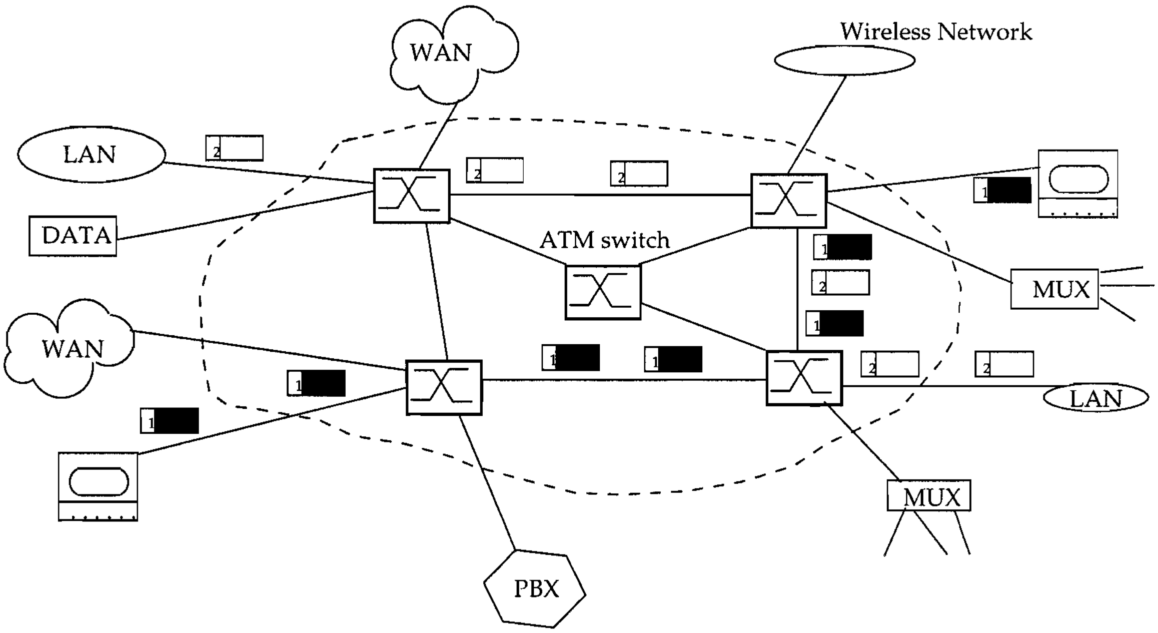


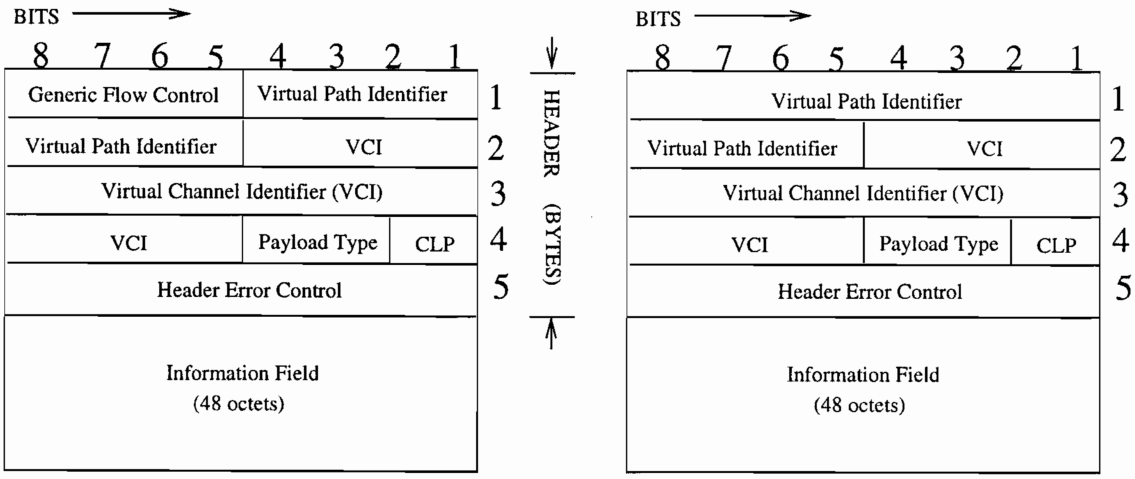
Figure 2.1: ATM Network

links constituting the VP (or VC) [16]. The ATM cell header is shown in figure 2.2. The B-ISDN standards define two different header formats. One format for the User-Network Interface (UNI) and another for the Network-Network Interface (NNI). The VPI and VCI are used to identify the specific call or connection to which the cell belongs. As mentioned earlier, VPI and VCI are also used to switch the cell at every intermediate ATM switch. More on ATM switch architecture can be found in [16]. The cell header also has reserved bits for special functions such as priority and congestion indication.

2.2 B-ISDN Protocol Stack

The B-ISDN standards are based on a layered protocol stack, just as all other recent communication standards. The protocol stack can be broadly divided into three layers. Each of these layers is further subdivided. Layer 1 deals with physical transmission, layer 2 deals with the network and transport functions and layer 3 contains the user, management and control functions. Figure 2.3 shows the protocol stack and table 2.1 indicates the services offered by B-ISDN.

Physical Transmission Layer (SONET) - The physical layer resides just below the ATM layer. The physical layer is responsible for moving the cells presented by the ATM layer from one network node to another. *Synchronous Optical Network* (SONET) is chosen as physical layer proto-



User-Network Interface

Network-Network Interface

Figure 2.2: ATM Header Structure

Class	Bit Rate	Timing Relation	Connection Mode
A	Constant Bit Rate (CBR)	Required	Connection-oriented
B	Variable Bit Rate (VBR)	Required	Connection-oriented
C	VBR	Not required	Connection-oriented
D	VBR	Not required	Connection-less
X	VBR	Not required	Connection-oriented

Table 2.1: B-ISDN service classes

col by the B-ISDN standards body. SONET is currently used by most of the telephone companies for their high bandwidth optical links. SONET comprises the *Synchronous Transport Signal* (STS - electrical domain) and *Optical Carrier* (OC - optical domain) series of standards. The OC-3 ($\approx 150Mbps$) and the OC-12 ($\approx 600Mbps$) are expected to be the rates adopted at the UNI for B-ISDN.

ATM Adaptation Layer (AAL) - The ATM layer resides above the physical transmission layer. The ATM layer requires all information presented in the form of cells. The *ATM Adaptation Layer* (AAL) is between the ATM layer and upper layers. This layer provides functions such as segmentation and reassembly of user information to ATM cells.

A number of adaptation layer types have been defined to correspond to the B-ISDN service indicated in table 2.1. One AAL layer can serve one or more classes of users; for example, one AAL type can serve both connection-oriented and connection-less data traffic. Apart from the func-

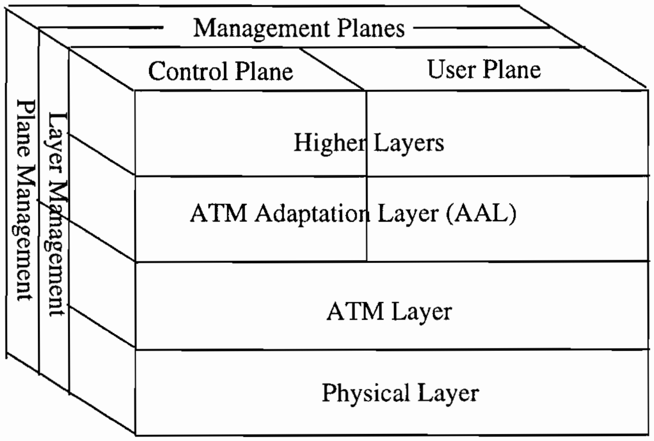


Figure 2.3: B-ISDN model

tion mentioned above (segmentation and reassembly), AAL is supposed to carry out additional, service-dependent functions also. To distinguish between these two functions, the adaptation layer is divided into two sub layers, a convergence sublayer (CS) resting below the higher (user) layer and a segmentation and reassembly (SAR) sublayer (figure 2.4).

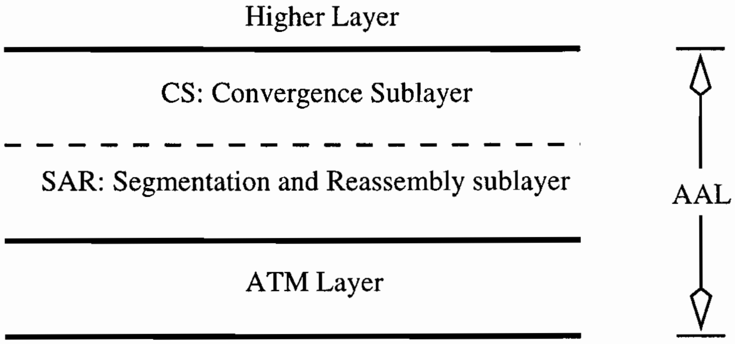


Figure 2.4: ATM adaptation layer (AAL)

Four different types of AAL have been proposed or defined. These are labeled AAL types 1,2, 3/4, and 5.We will explain the functions AALs 1, 3/4 and 5 in brief here. More details about the AAL layers and their functions can be found in [27].

- AAL 1 is used to support *Constant Bit Rate* (CBR) connections across the network. This layer provides transfer of timing information, data structure information and indication of lost or errored information between source and destination.

- AAL 2 is used to transfer variable bit rate data which is time dependent. It sends timing information along with the data so that the timing dependency may be recovered at the destination. AAL-2 provides error recovery and indicates errored information which could not be recovered. As the source generates a variable bit rate some of the cells transfered may not be full and therefore additional features are required at the segmentation and recovery layer. More on AAL-2 will be found in next section.

- AAL 3/4 is recommended to be used for transfer of data which is sensitive to loss, but not to delay. This AAL may be used for connection-oriented as well as for connection-less data communication. This AAL has a substantial overhead in terms of sequence numbers and multiplexing indicators and is rarely used.

- AAL 5 is an outgrowth of data communication industry and is optimized for data transport. In AAL 5 and AAL 3/4, the *convergence sublayer* (CS) is split into two parts, a higher *service specific convergence sublayer* (SSCS) and a lower *common part convergence sublayer* (CPCS). AAL 5 is a much simpler protocol when compared with AAL3/4. AAL 5 has less control overhead and correspondingly less functionality.

Transport Layer - The transport layer is the link between User layer and the lower layers. This layer is responsible for providing a reliable end-to-end transfer of user data. The functionalities of transport layer includes connection management, data acknowledgment, error handling and congestion control.

User, Management and Control Planes - The B-ISDN protocol stack is further divided into three planes (see figure 2.3), each of which incorporates all of the layers previously discussed. This is the uppermost layer of the B-ISDN protocol stack and is divided into three planes. The user plane, control plane and management plane. The user plane is responsible for providing the network services to the user applications running in the end systems. For variable bit rate users, both connection oriented and connection-less services are supported, whereas for constant bit rate users, connection oriented service is supported. The management plane is responsible for *Operation, Administration and Management* (OA& M) of the network. It monitors the network for faults and transmission performance. The control plane takes care of signaling, call set up and call clear functions. The signaling messages are sent on the same network as specially marked ATM cells.

2.3 AAL2: In Detail

With the recent advances in telecommunications industry, most of the applications are moving towards packet-based technology. Different applications have different requirements. ATM is able to answer the needs of most of the applications efficiently but one of the areas where it is not very successful (untill recently) is *support for VBR-rt*. ATM with the help of AAL1 is able to support applications with CBR traffic and with the help of AAL5 it has been supporting VBR-nrt applications.

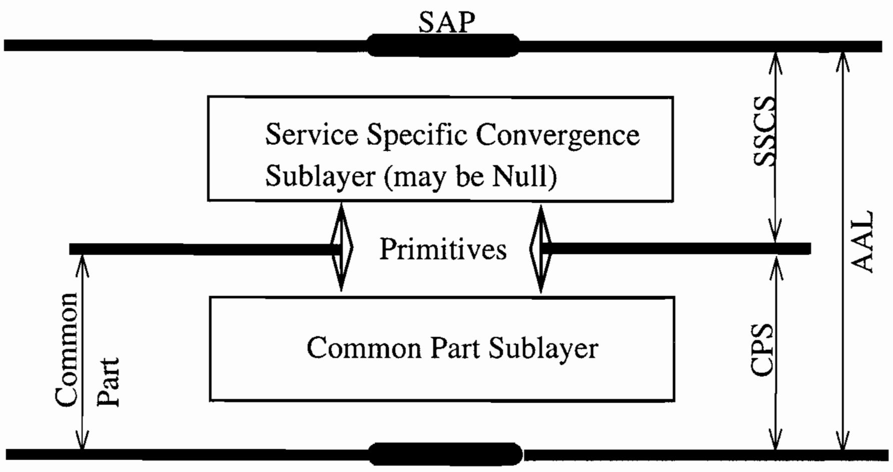
In order to support applications where the "bandwidth required" is not constant throughout the connection time (unlike CBR) but the QoS constraints (like delay) are very stringent (like CBR), ATM adaptation layer 2 (AAL2) is proposed. AAL2 is used to multiplex more than one low bit rate user information stream on a single ATM virtual connection [10]. This AAL provides bandwidth efficient transmission of low-rate, short, and variable length packets in delay sensitive applications. In situations where multiple low bit rate, delay-sensitive data streams, for example voice, need to be connected on end systems, a lot of precious bandwidth is wasted in using conventional VCs for each of the connections. Moreover, most network carriers charge on the basis of number of open VCs, hence it is efficient both in terms of bandwidth and cost to multiplex as many of these as possible on a single connection. A preliminary standard has been published by the International Telecommunication Union (ITU-T). Although some portions require further refinement, the standard is well described in [10]. The AAL type 2 is subdivided into the *Common Part Sublayer* (CPS) and *Service Specific Convergence Sublayer* (SSCS) as shown in Fig 2.5. Different SSCS protocols may be defined to support specific AAL2 user services or groups of services. The SSCS may also be null, providing merely for the required mapping between the CPS and higher layers. AAL2 provides the capabilities to transfer *AAL service data units* (AAL-SDUs) from one *AAL-service access point* (AAL-SAP) to another through the ATM network. Multiple AAL2 connections may utilize a single underlying ATM connection. The multiplexing and de-multiplexing of connections occurs at the CPS.

2.3.1 CPS to ATM data interface

The CPS hands a 1 bit ATM -User-to-User (AUU) indication, a loss priority (called the Submitted Loss Priority (SLP)) and a 48 byte ATM payload to the ATM layer below it. SLP is used by the ATM layer to set its own CLP bit. CPS also receives from the ATM layer a 48 byte SDU and a loss priority bit (called the Received Loss Priority). The RLP may differ from SLP in case the network changed CLP along the way.

2.3.2 CPS to SSCS data interface

The CPS-Interface data packets are handed over to the SSCS (1 to 64 bytes) by CPS. The format and actual length of the data are determined at setup time. The CPS also hands a 5 bit User to User Indication to the SSCS. This is data used optionally by the SSCS entity to decide the destination of the *protocol data unit* (PDU). The CPS also receives the same two units from the SSCS entity.



CPS	Common PartSublayer
SAP	Service Access Point
SSCS	Service Specific Convergence Sublayer

Figure 2.5: AAL2 Structure: Sections and Data Interfaces

2.3.3 The Common Part Sublayer

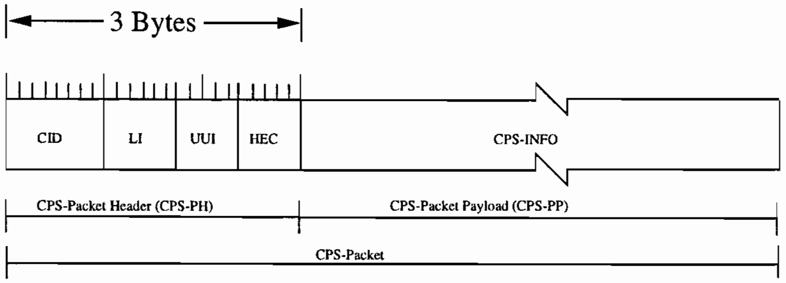
AAL2 CPS offers the following peer to peer operation:

- Data transfer of CPS-SDUs of up to 44 (default) or 64 bytes.
- Multiplexing and de-multiplexing of multiple AAL2 channels.
- Maintains the CPS-SDU sequence integrity on each AAL2 channel.
- Un-assured operation, i.e. lost CPS-SDUs are not retransmitted.
- Bidirectional virtual channel connection, using the same VC number in either direction. The VC can be permanent or switched.

The CPS interacts with both the management layer and the control layer. The control layer establishes the VC as required. Switching at the channel level has not yet been defined.

2.3.4 Format and Encoding of CPS Packet

A CPS Packet consists of a 3 byte Packet Header (CPS-PH), followed by up to 64 bytes of Packet Payload (CPS-PP). CPS Packets are the data exchange mechanism between CPS and SSCS. Fig 2.6 shows the field lengths and format.



CID:	Channel Identifier	(8 bits)
LI:	Length Indicator	(6 bits)
UUI:	User-to-User Indication	(5 bits)
HEC:	Header Error Control	(5 bits)
CPS-INFO:	Information	(1 ..... 44/64 oct)

Figure 2.6: AAL2 CPS Packet Format

- Channel Identifier (CID) value identifies the AAL2 channel user. The AAL2 channel is a bidirectional-medium, and both directions use the same value of CID.

CID value	use
0	not used
1	reserved for layer management peer-to-peer operations
2 ... 7	reserved
8 ... 255	identification of SSCS entity

- Length Indicator (LI) is a binary encoded value that corresponds to the length of the payload of the CPS-Packet. The default maximum length is 44 bytes but it can be set to a maximum of 64 bytes. The maximum length needs to be negotiated at setup time. Maximum lengths between 44 and 64 are not allowed.

- User-to-User Indication (UUI):

  - ◇ It is used to convey specific information to SSCS entities transparently through the CPS.
  - ◇ It is used to distinguish between the SSCS entities and Layer Management users of the CPS

  The 5 bit UUI field is handed without change by CPS to the SSCS entity. Its usage by the SSCS entity is optional.

- Header Error Control (HEC) is the reminder (modulo 2) of the division, by generator polynomial $X^5 + X^2 + 1$, of the product of X^5 and the contents of the first 19 bits of the CPS-PH. The receiver uses the HEC field to detect errors in the CPS-PH.

2.3.5 Format and Encoding of the CPS-PDU

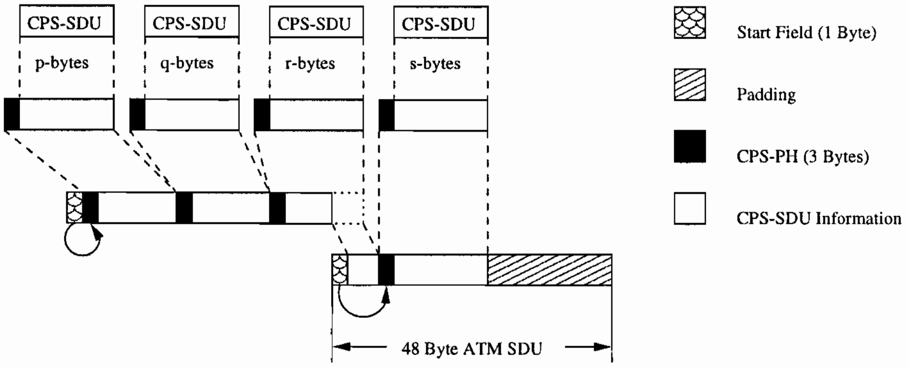
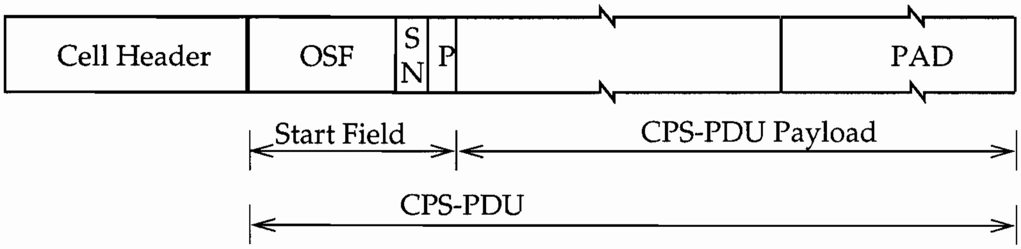


Figure 2.7: Translating CPS-SDUs to ATM SDUs

The CPS-PDU consists of a one byte start field (STF), and 47 byte payload. The 48 byte CPS-PDU is the ATM cell SDU (Fig 2.7). A CPS-PDU may carry zero, one or more full or partial CPS-Packets. The packets may overlap over more than one PDUs. Any unused space in the PDU is padded with zeros. The CPS-Packet may be partitioned anywhere along it's length (Fig 2.8). The start field values are



OSF:	Offset Field	(6 bits)
SN:	Sequence Number	(1 bit)
P:	Parity	(1 bit)
PAD:	Padding	(0 to 47 octets)

Figure 2.8: AAL2 CPS PDU Format

◇ Offset Field (OSF): This field carries the binary value of the offset, measured in number of bytes, of the first start of a CPS-Packet or, in the absence of a start of a CPS-Packet, to the beginning of the PAD field. A value of 47 indicates that there are no CPS-Packet starts in this PDU.

- ◇ Sequence Number (SN): This 1 bit field is a modulo 2 sequence number of the stream of CPS-PDUs.
- ◇ Parity (P): To detect errors in the STF, a 1 bit odd parity is set as the last bit of the STF.

2.4 Service Classes in high-speed networks

2.4.1 Classification based on bit rate characteristics

The traffic presented to the network can be classified in two categories,

- Constant bit rate (CBR) services - This class includes those services in which bits are generated periodically and hence there is a constant flow of information. Uncompressed voice is an example of CBR traffic.
- Variable bit rate (VBR) services - This class includes the services which present information to the network at a variable rate. The mean bit rate is usually much smaller than peak cell rate. There are two types of VBR services possible:
  - ◇ Bursty VBR Services - A class of services in which flow of digital information is interrupted (on-off) variable rate. An example of this class of service is bursty data communications.
  - ◇ Stream VBR Services - These are real-time applications which involve uninterrupted flow of variable rate information. An example of this class of service is encoded video.

2.4.2 Classification based on connection modes

The connection mode can be one of the two classes of services, connection-oriented or connection-less.

- Connection-oriented service - In such services, connection has to be setup before any transfer of information. Examples for this class of services are voice calls and video connections.
- Connection-less service - These class of services does not require the connection set up phase. The user packets are presented to network along with some routing information in the header. Current datagram networks are typical examples of this class of services.

B-ISDN's classification of types of services based on connection characteristics is shown in table 2.1.

2.4.3 Classification based on applications

Services provided by B-ISDN, depending on their applications, fall into one of the two categories

- Interactive services - Services which require constant interaction between both the ends. These services are further classified into three types - conversational services, retrieval services and messaging services. In the work presented in this thesis, we are primarily concerned with conversational and interactive service (i.e. voice).
- Distributive services - These are services of broadcast type with many users receiving same information. These services are subdivided into two types: distributive services without user individual presentation control and distributive services with user individual presentation control.

2.5 Quality of service requirement in B-ISDN

In ATM-based B-ISDN networks, QoS requirements are mentioned at cell-level and connection-level. The cell-level QoS requirements includes transmission delay, cell loss probabilities, jitters and the like, whereas the connection-level QoS requirements are similar to the existing circuit-switched networks, for example call blocking probability, call setup time, etc. The focus of this project is mainly cell-level QoS requirements but it is important to realize the possible interactions and trade-offs between cell-level and connection-level QoS requirements. For example, a strict connection admission control strategy might result in lowering cell loss probability at the expense of increased call rejection probability. More details on such trade-offs and their role in the design of optimal call-admission control schemes can be found in [23]. There are different types of cell-level QoS guarantees:

- Deterministic guarantee: Deterministic guarantee requires the value of QoS parameters to be always within some specified range. In order to provide such a guarantee, the network has to consider the worst case traffic scenario which results in poor utilization of network resources even with the most efficient resource allocation schemes. An example of deterministic guarantee is upper bound on the end-to-end transmission delay.
- Statistical guarantee: Statistical guarantees require the statistical values of the QoS parameters to lie in a predefined-range. An important factor in this approach is *time-scale of interest.* Most of the research done in the area of traffic management is based on an infinite-horizon approach. Recent studies [19, 20] have shown that schemes based on long-term QoS parameters, such as average cell loss probability, may result in serious and unacceptable QoS violations because of infrequent overload periods. It is therefore necessary to supplement the infinite-horizon based QoS parameters with with additional ones [18, 22]. Examples of statistical guarantees are [18]:

- ◇ Prob{end-to-end cell-loss ratio} $\leq \epsilon$ for some $\epsilon > 0$;
- ◇ E{nodal-delay} $\leq \tau_1$ for some $\tau_1 > 0$ and
- ◇ the n-th percentile of end-to-end delay is less than τ_2 for some $\tau_2 > 0$.
  In our work, we will consider this option and it will be defined further in section 4.2.

- Best-effort but no guarantee. No QoS guarantees are provided but the user is allowed to submit traffic to the network at his own risk. If any bandwidth is available with the network, it is allocated to the user temporarily. Traffic class *unspecified bit-rate* falls under this category.

2.6 Congestion Control in ATM networks

Future broadband networks are expected to support a lot of bandwidth-hungry applications like high-resolution images, high-resolution video, multimedia traffic along with the existing interactive data, voice and computer traffic. It is difficult to characterize or model this mixed future traffic and it becomes more complicated if differing qualities of services (QoS) are required for each. Due to availability of *wider bandwidths*, the delay-bandwidth has also increased. The problem can be better understood using the following example.
Consider an OC-12 link (operating at 622 *Mbps*) between New York and San Francisco ($\approx 3000 miles$) with a one-way propagation delay of 24 ms, transmitting 53 bytes ATM cells. By the time the first cell reaches the destination, around 35,207 cells have already left the source and are in transit. There will not be any effect of feedback congestion control scheme on these cells in transit. For the reasons mentioned above, the application of conventional congestion control schemes is not straight forward. Below are some other challenges of congestion control in high-speed ATM networks.

- While supporting a wide range of applications each with different bandwidth requirements and QoS constraints, *bandwidth utilization* has to be maximized.

- Due to high-speed transmission, the amount of data in transit will be huge. Smaller buffers will lead to data loss and large buffers will introduce large queuing delays. Therefore, sizes of buffers to be used in the intermediate nodes has to be optimized.

- Due to increased transmission and switching speeds, the traffic patterns will be more volatile making congestion control tougher.

- In order to keep up with the increased link speeds, it is desirable to have simple congestion control protocols so that the protocol processing time over each ATM cell is small.

- Fairness among the users, sharing the same network resources, should be ensured.

2.6.1 Overview of congestion control options

Congestion control is very important in high-speed networks in order to achieve network performance objectives and at the same time increase the network resource utilization. The congestion control schemes can be classified broadly into two categories: *preventive schemes* and *reactive schemes*. In the preventive schemes, the goal is to avoid congestion within the network. Reactive schemes control the congestion once it occurs in the network. Figure 2.9 shows the category of some of the congestion control schemes that are proposed [25].

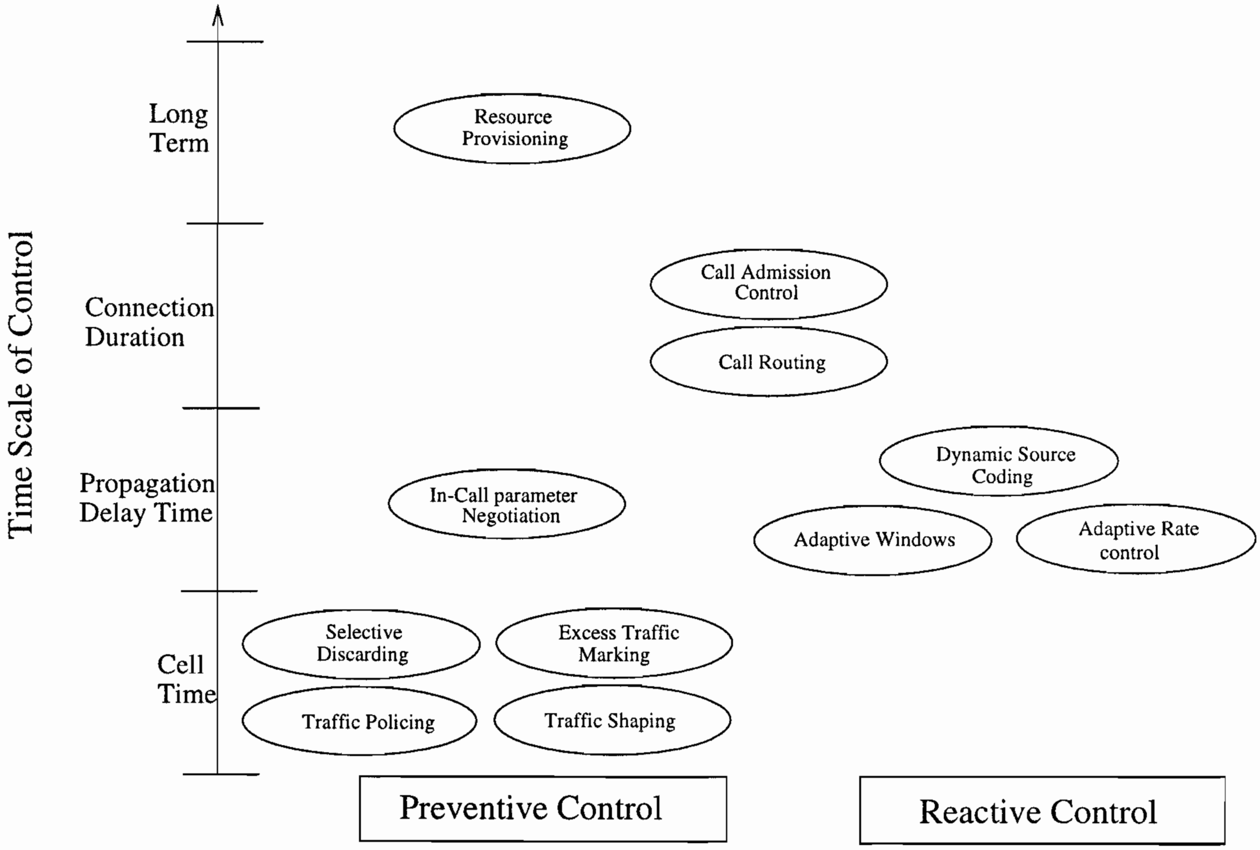


Figure 2.9: ATM congestion control options [25]

Preventive schemes - In these schemes, the connection's traffic is estimated a priori and allocated sufficient amount of bandwidth (and possibly buffers). This allocated bandwidth is constant for the duration of connection, and it is expected that the connection remains within the limits of its allocated bandwidth. Preventive schemes require accurate modeling of short and long term traffic patterns in order to satisfy the objectives of network. Thus, there are three stages to achieve congestion control with such schemes -

- Traffic characterization - Depending on the application, its short term and long term characteristics, traffic models need to be developed to carry out accurate design and performance evaluation of networks. Examples are on-off source model, Poisson model etc.

- Connection admission control (CAC) - Its function is to decide whether to accept or reject a new connection request. CAC estimates the amount of bandwidth required by the connection based on its traffic descriptors and QoS requirements. If the resources are available with the network, the connection is accepted; else it is rejected. Estimation of traffic descriptors in itself is a major research issue. Over the last few years, many CAC schemes have been proposed [2, 3, 12, 13, 5, 14, 15, 1], and [5] gives a nice overview of many of them. We will deal with a few of them in section 4.1.

- Bandwidth policing: Once bandwidth is reserved for a connection, the network must ensure that the source stays within the requested traffic parameters. This is known as traffic policing or usage parameter control (UPC). The policer resides at the user-network-interface (UNI). Examples of policing schemes are leaky bucket scheme, moving window, sliding window, jumping window etc.

Reactive schemes - With preventive control schemes, it is very difficult to eliminate the cell loss/delay due to buffer overflow. There are two types of congestion-related cell loss - momentary buffer overflows and sustained buffer overflows [25, 26]. Momentary buffer overflow occurs when many users submit traffic at their peak cell rate to the network simultaneously. Sustained buffer overflows occurs due to the correlations between user traffic for a sustained period of time. The reactive schemes attempt to ensure that momentary buffer overflow does not turn into sustained congestion. These schemes regulate the traffic from a connection at the access point based on feedback about the status of the network. This feedback information is sent by the network and/or end users.

Reactive schemes can be used as backup for preventive schemes, or as an aggressive control to maximize the network utilization. The primary disadvantage of most of the reactive schemes is their inefficiency in an environment with large delay-bandwidth product. In the work presented here (chapter 5), we have made an efficient use of this scheme in the context of voice and telephony over ATM (VTOA).

Many reactive congestion control schemes are proposed in the literature. Prominent of them are congestion control with

- Explicit congestion notification - A special bit in the packet header is reserved which informs the end systems about the congestion in the network. Each node monitors the traffic and sets the congestion bit on all the cells passing through it if congestion is detected.

- Estimation by end systems - Congestion is estimated by the end systems by measuring the round-trip response time between source and destination.The round-trip delay is calculated by the destination based on the time-stamp of cells (stamped at the source) and the cell arrival time. Destination sends a message to the source asking it to reduce its rate if the estimated one-way delay of few cells are greater than the expected delay.

- Adaptive rate control, adaptive windows and dynamic source coding all of which work on the same principle: reduce the source transmission rate in response to the congestion information from the network. Chapter 5 describes and evaluates a dynamic source coding scheme in an AAL-2 context.

Chapter 3

CBR vs. VBR virtual circuits for AAL2

As discussed in chapter 2, Asynchronous transfer mode (ATM) is considered to act as a universal bearer for broadband integrated services digital networks (B-ISDN) networks, which can carry voice, data and video with the same cell transport arrangement. For any network service provider (NSP), voice will continue, for some time, to be a majority traffic through the network. A big challenge facing NSPs is to satisfy a variety of user requirements and at the same time increase the utilization of network resources, which increases the profit made out of the business.

To achieve the above mentioned goal, a concept called Voice Telephony Over ATM (VTOA) was introduced. This concept has many advantages over other concepts like Voice Over IP (VOIP), prominent of which is quality of service (QoS) guarantees that can be provided to the users by VTOA. VTOA is done with the help of AAL-2 which provides bandwidth efficient transmission of low-rate, short and variable length packets in delay sensitive applications (like voice) by performing multiplexing at the VC level. Many issues, ranging from selection of parameters to introduction of new protocols, are associated with implementation of VTOA.

AAL2 is found to be very efficient in terms of bandwidth savings (resource utilization) in applications like voice over ATM. However, the NSP can go on exploiting the maximum resource utilization techniques only to a certain extent (as long as he is satisfying users QoS requirements).

Realizing the need for increasing efficiency while delivering the promised quality of service, we are starting with one of the basic questions surrounding VTOA, which is selection of the type of virtual circuit (VC) service to be used for VTOA. This is in contrast to the view that the VC should use real-time variable bit rate (rt-VBR) service in order to match the traffic characteristics and efficiently use network bandwidth.

In this report, by explaining the effect of VC type on QoS requirements of users, we take the position that the VC should be a constant bit rate (CBR) VC since it makes most sense in applications like VTOA.

3.1 Selection of VC nature

Our criteria for selecting the VC type are following:

- Satisfy the QoS guarantees (for example, delay and cell loss probability constraints for voice users) for all admitted users, even in the worst case traffic scenario.
- Fairness among all the admitted users.
- Maximize the network's resource utilization (for example, better utilization of VC and virtual path).
- Minimize the overall network complexity (able to manage the traffic through the network easily).

The need to select a particular type of VC is felt because of the following reasons:

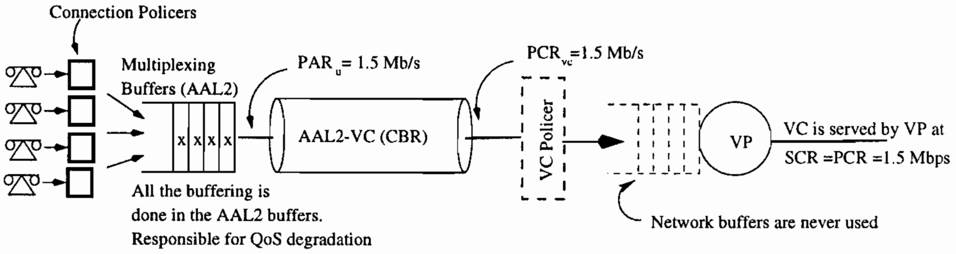
- The nature of the VC affects (more or less) the functionality of many network entities like policers, switches etc.,
- Many strategic decisions like the size of VC/VP, number of users to be supported on that VC/VP, size of buffers, signaling etc., depends on this.
- Dictates the Connection Admission Control (CAC) strategy of network,
- It also helps us in defining the functionality and interdependency of AAL-2 CAC (Connection level call monitoring) and ATM CAC (VC level traffic management) and their dependency on other network entities.

For the purpose of explanation and comparison, consider a scenario where we have a number of voice sources and all of them are sharing a common VC with an AAL2 *permit arrival rate* ($\mathbf{PAR_u}$) corresponding to 1.536 *Mbps* (T1). The type of VC could be constant bit rate (CBR), static variable bit rate (S-VBR) or dynamic variable bit rate (D-VBR). Each user is modeled as an on-off voice source with mean on-time equal to 0.42 seconds and and mean off-time of 0.58 seconds. When on, each user transmits at a rate of 32 *kbps*. The CPS-packet size from each user is 12 bytes [8]. Also assume that all of them have the same QoS requirements ($95^{th}\%ile$ delay inside AAL2 transmitter is less than or equal to 10 milliseconds).

3.2 Options Available

In order to define a particular CAC we need to have a reference bandwidth 'B' [1] for the VC and its nature (CBR, VBR, etc.). The nature of reference bandwidth 'B' could be of one of the three types

[1] What is the maximum number of connections we can support if we have a B amount of bandwidth



In CBR VC, VC policer and buffers before VP are not required.
Since there is no queueing at the VP Mux (because 1.5 Mbps bandwidth is always guartanteed), end-to-end QoS guarantees can be supported.

Figure 3.1: Illustration of CBR VC option

- CBR VC: The constant bit rate VC has a fixed amount of bandwidth corresponding to the AAL2 permit arrival rate ($\mathbf{PAR_u}$) (1.536 *Mbps*, in our case) dedicated to it throughout the network. This bandwidth is characterized by the peak cell rate ($\mathbf{PCR_{VC}}$) of the VC. Figure 3.1 illustrates this option and shows that $\mathbf{PCR_{VC}}$ (T1) amount of bandwidth is guaranteed by the network to this VC. AAL2 level CAC is done by taking peak cell rate ($\mathbf{PCR_{VC}}$) of the VC as the reference bandwidth. Figure 3.1 also shows the voice sources described above, multiplexed on to a single CBR VC. Due to statistical multiplexing of the connections (probability that all the connections are on at the same time is nearly zero), the maximum number of connections that can be admitted, with a guarantee of satisfying their respective QoS parameters, is greater than 48 ($\frac{1.536*10^6}{32*10^3} = 48$). Simulations [8] show that for the given source parameters, multiplexing buffer size (at VC) and QoS parameters, 72 sources can be supported simultaneously by a CBR VC serving at T1 rate. All these 72 users' traffic is first multiplexed in the buffers within the AAL2 transmitter. The buffer contents (or queue length) varies depending upon the instantaneous arrival rate [2] to these buffers. If the instantaneous rate is greater than the VC's service rate, buffering takes place, leading to delay of packets. Once served by the VC, these packets experience near-zero delay and zero loss while traversing the network. In other words, the *end-to-end QoS is essentially the same as the QoS provided by the AAL2 transmitter* (95% *of the packets delayed less than 10ms).*

  The policer before each connection makes sure that the connection is following the traffic contract agreed at the time of connection setup. If any source violates the traffic contract, the packets of the corresponding user are tagged and sent into the network. In case of congestion in the network (at the multiplexing buffers before the VC), these tagged packets may be dropped. This means that if a user is operating within the traffic contract, his QoS is assured, end-to-end. This is different from the case of a VBR VC, which we will see later in this

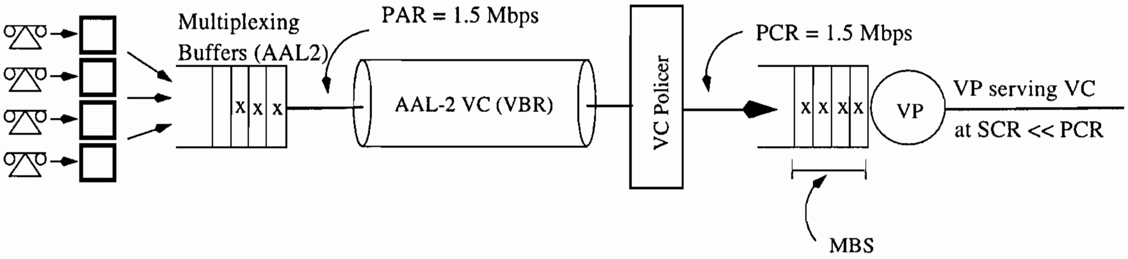
[2] $n * 32$ *kbps*, where n is number of active connections

section, where a user might experience QoS degradation even if he is operating within the traffic contract limits.

The policer and buffers before the VP are **not** required since the AAL2 permit arrival rate (**PAR$_u$**) effectively controls the VC's **PCR$_{VC}$** and **PCR$_{VC}$** amount of bandwidth is always guaranteed to the CBR VC by the network. QoS guarantees are met most easily with this option. The advantages and disadvantages of this option are summarized in section 3.4.

- Static VBR VC: A static VBR VC is characterized by triplet, peak cell rate (**PCR$_{vc}$**), sustainable cell rate (**SCR$_{vc}$**) and maximum burst size (**MBS$_{vc}$**) which is negotiated at the time of VC setup. The values of this triplet remain the same for the connection duration. Since the AAL2 permit arrival rate (**PAR$_u$**) effectively limits the VC's peak rate, the only sensible choice for **PCR$_{vc}$** is **PCR$_{vc}$** = **PAR$_u$** (or 1.536 *Mbps*) [3]. **SCR$_{vc}$** could be anything between (0,**PCR$_{vc}$**) depending on the traffic contract between the VC and network. **MBS$_{vc}$** depends on the buffers available in the network and values of other two parameters of the triplet. Network service provider promises to satisfy the QoS requirements as long as VC conforms to the triplet (**PCR$_{vc}$**,**SCR$_{vc}$**,**MBS$_{vc}$**), i.e., average rate over some interval must be less than **SCR$_{vc}$** and bursts at **PCR$_{vc}$** are limited to **MBS$_{vc}$**.

  Figure 3.2 illustrates this option. It is different from the CBR option in the sense that the VBR VC is served at its **SCR$_{vc}$** whenever the network is congested and has **PCR$_{vc}$** available only when the network is very lightly loaded.



Buffering of packets and cells takes place at VC buffers and VP buffers respectively.
This makes AAL2 CAC complex and dependent on ATM-CAC (CAC before VP).

Figure 3.2: Illustration of VBR VC option and the associated problems

The network is assumed to guarantee only **SCR$_{vc}$** bandwidth to the VC. Therefore, in order to satisfy the QoS of all the admitted users and maintain fairness among users, we need to do AAL2 CAC on the basis of **SCR$_{vc}$**.

[3] We express the cell rates using equivalent bit rates. (Bit rate = (cell rate)$*$(53)$*$(8))

One of the problems with VBR VC is to select the value of this $\mathbf{SCR_{vc}}$, on the basis of which AAL2 CAC has to be done. The answer to this question is one of the key points in deciding the type of VC. To answer this question, and for simplicity and comparison purposes, let's consider the same example where we have 72 users sharing a link with peak cell rate equal to 1.536 *Mbps*. With the help of source parameters described in [8], we can define three rates corresponding to the traffic coming to the AAL2 multiplexing buffers. They are

- ◇ Mean arrival rate ($\mathbf{MAR_u}$): It is the long term average equivalent cell arrival rate which in our example is appoximately equal to 1335.6 *kbps* $\{(72)(0.42)(32kbps)(\frac{53}{48})(\frac{15}{12})\}$.
- ◇ Permit arrival rate ($\mathbf{PAR_u}$): This is the AAL2 permit arrival rate, which controls the service rate of the AAL2 buffers. In our example it is equal to 1.536 *Mbps*.
- ◇ Maximum arrival rate ($\mathbf{XAR_u}$): The maximum possible equivalent cell rate at which traffic arrives to the buffers given our example by $\frac{\mathbf{MAR_u}}{0.42}$.
- ◇ Instantaneous arrival rate ($\mathbf{IAR_u}$): The equivalent cell rate at which the traffic is arriving at the AAL2 multiplexing buffers (before the VC) at a given instant of time. It can take any value ranging from zero to $\mathbf{XAR_u}$ depending upon the number of active on-off sources and their coding rates.

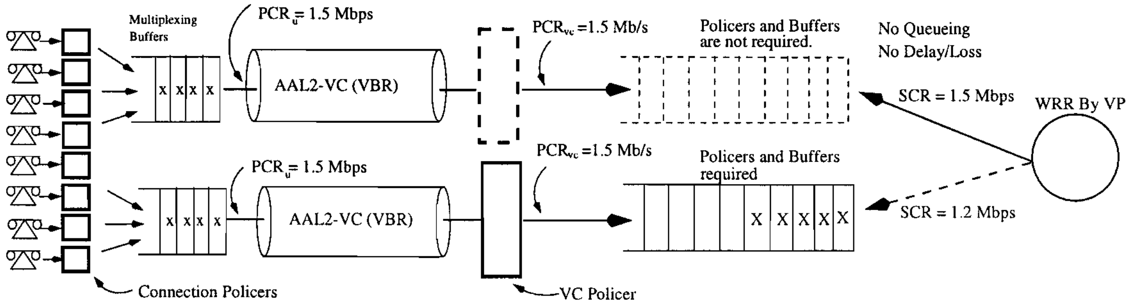
Coming back to our problem of selecting a value for $\mathbf{SCR_{vc}}$, the following two cases are possible.

1. The value of $\mathbf{SCR_{vc}}$ is equal to $\mathbf{PAR_u} = \mathbf{PCR_{vc}}$. This is equivalent to CBR VC.
2. The value of $\mathbf{SCR_{vc}}$ is less than $\mathbf{PAR_u} = \mathbf{PCR_{vc}}$. Assuming the network will occasionally be congested (network service rate is occasionally limited to $\mathbf{SCR_{vc}}$), $\mathbf{SCR_{vc}}$ should be greater than $\mathbf{MAR_u}$. For purposes of illustration, let's assume that $\mathbf{SCR_{vc}}$ is chosen as 1.2 *Mbps*.

Figure 3.3 shows these two cases. The first case is equivalent to choosing a CBR VC, and so requires no further explanation.

In the second case, the rate at which the VBR VC serves the AAL2 multiplexing buffers before it is constant ($\mathbf{PAR_u} = \mathbf{PCR_{vc}} = 1.536\,Mbps$) **but** the rate at which this VC is served by the VP varies with time and its guaranteed value is $\mathbf{SCR_{vc}} < \mathbf{PCR_{vc}}$. In this case, there are several choices for reference bandwidth for AAL2 CAC.

Suppose first that the reference bandwidth is taken to be the AAL2 $\mathbf{PAR_u}$ $(1.536Mbps)$. In this case, 72 voice sources would be admitted, but since the guaranteed network service rate is only $\mathbf{SCR_{vc}} = 1.2Mbps$, additional buffering (delay) could occur inside the network, so *the end-to-end QoS will depend on network load.* If the network is very lightly loaded, the end-to-end QoS will be essentially the same as the AAL2 QoS (95% of packets delayed less than 10 ms). If the network is very heavily loaded, the end-to-end QoS can be effectively



VC_ 1: PCR = 1.5 Mbps ; SCR = 1.5 Mbps;
VC_2: PCR = 1.5 Mbps ; SCR = 1.2 Mbps;
VC_3: PCR = 1.5 Mbps ; SCR = 0.968 Mbps; Maximum buffer size for all the VC are kept same = 1000 cells

In order to satisfy QoS guarantees end to end, we need to have a bandwidth >= the rate of traffic coming into the VC.

Figure 3.3: Issues associated with selection of different VC with different values of $\mathbf{SCR_{vc}}$

calculated as the QoS resulting from 72 users multiplexed on a link with bandwidth equal to $\mathbf{SCR_{vc}}$ (1.2 *Mbps*), which will be worse than the AAL2 QoS (which results from 72 sources multiplexed on a link with bandwidth equal to $\mathbf{PAR_u} = 1.536Mbps$).

If a conservative approach is taken to provide QoS guarantees, one would assume the latter (worst) case. *But note that doing so is essentially equivalent to using a CBR VC with* $\mathbf{PCR_{vc}} = \mathbf{PAR_u} = 1.2Mbps$. For this conservative approach, the VBR VC option has no advantage over the CBR VC option. If a less conservative approach is taken, the AAL2 CAC must somehow combine the QoS 'guaranteed' by the ATM-level CAC for the VBR VC (based on $\mathbf{SCR_{vc}} = 1.2Mbps$) and the AAL2 QoS (based on $\mathbf{PAR_u} = 1.536Mbps$) to estimate end-to-end QoS. This interdependency between AAL2 CAC and ATM CAC significantly complicates the CAC problem.

But suppose the reference bandwidth for AAL2 CAC is taken to be $\mathbf{SCR_{vc}}$ (1.2 *Mbps*). This is essentially equivalent to the conservative (worst-case) QoS assumption, (QoS resulting from 72 source multiplexed on a link with bandwidth of $\mathbf{SCR_{vc}} = 1.2Mbps$).And we have already concluded that in this case one may just as well use a CBR VC with $\mathbf{PCR_{vc}} = \mathbf{PAR_u} = 1.2Mbps$.

Furthermore, if the network is serving the VC at a rate *lower* than the rate at which the VC is serving its multiplexing buffers (at 1.536 *Mbps*), unfairness may result. In such a case, there will not be any buffering in network buffers or the AAL2 buffers as long as $\mathbf{IAR_u}$ is less than 1.2*Mbps*. But if $\mathbf{IAR_u}$ is greater than 1.2*Mbps* but less than 1.536 *Mbps*, buffering of *cells* will take place before the VP. In that case, the QoS degradation experienced by an individual packet is no longer solely dependent on the corresponding user's behavior; instead it depends on the behavior of all other users who are sharing the *cell* of which this packet is a part. In such scenario, users who are submitting traffic to the network according to the traffic contract as well as the users who are greedy and violating the traffic contract might experience QoS degradation at the VP level (fairness guarantees cannot be met by the network). This shows that there is no guarantee that the network can provide to the user packets even if the user conforms to the traffic contract.

Also, if $\mathbf{IAR_u}$ is greater than 1.536 *Mbps*, buffering of *packets* will take place in VC buffers and buffering of *cells* will take place in VP buffers. Again all the problems of QoS degradation will be applicable to this case as well. The second VC in figure 3.3 illustrates this.

Summary:
The QoS degradation experienced by the *cells* of the particular VBR VC depends on

- ◇ The difference between $\mathbf{PCR_{vc}}$ and $\mathbf{SCR_{vc}}$ of the VC (determines delay, lesser the difference lesser will be the delay and QoS degradation) ,
- ◇ Buffer size before the VP carrying this VC along with the other VCs, (decides cell loss probability),
- ◇ The traffic coming to the VP from other VCs because this is one of the factors which dictates the bandwidth available to the VC, above its promised $\mathbf{SCR_{vc}}$.

From the abovementioned arguments, we conclude that if we want end-to-end QoS to depend only on AAL2 QoS and not ATM (network) QoS, we require $\mathbf{SCR_{vc}} = \mathbf{PAR_u} = \mathbf{PCR_{vc}}$, i.e, the VC should be CBR. Choosing $\mathbf{SCR_{vc}} < \mathbf{PCR_{vc}}$ (i.e, a "real" VBR VC) will result in interdependencies between AAL2 QoS and ATM QoS, significantly complicates the AAL2 CAC problem.

- Dynamic VBR VC: Similar to static VBR VC but the triplet changes during the connection period depending on offered traffic and available bandwidth in the network. This could be done using dynamic renegotiation between user and network. Since it is practically very difficult to implement, we will not pursue this option.

3.3 Alternate Simple Explanation

The selection of CBR VC or VBR VC actually affects the point of constriction (buffering) in the network. The CBR VC does that at point A (shown in figure 3.4.). In order to avoid congestion, AAL2 CAC needs the information about the user traffic at AAL2 buffers and the size of these buffers. Since the AAL-2 CAC has access to this information, it is possible to develop an effective, simple scheme (which is independent of ATM CAC scheme) to maximize the link utilization and satisfy the QoS guarantees. CBR VC guarantees fairness in the network and the degree of QoS degradation experienced by individual user depends on the extent to which the user is violating his traffic contract with network.

As opposed to CBR VC, in the case of VBR VC, congestion (buffering of *packets* and *ATM cells*) occurs at the entry to the VC as well at the VP entry point (multiplexing buffers at points C and D in figure 3.4) depending upon the value of $\mathbf{SCR_{vc}}$. Even if the AAL2 CAC is done on the basis of $\mathbf{SCR_{vc}}$, the instantaneous arrival rate to the AAL2 multiplexing buffers could be more than $\mathbf{SCR_{vc}}$. In such a case, the incoming traffic will be served by the VC at $\mathbf{PCR_{vc}} = \mathbf{PAR_u}$ (and not at $\mathbf{SCR_{vc}}$) whereas the same traffic might be served by the VP at $\mathbf{SCR_{vc}}$ (depending on network congestion). This would result in delay/loss of *cells* which might contain *packets* coming from a user operating within the limits of the traffic contract. To provide QoS guarantees to all the users and maintain fairness in the network, it is required that the VC should be continuously served at the same rate at which this VC is serving the traffic coming to its multiplexing buffers. That is, the VC should be a CBR VC.

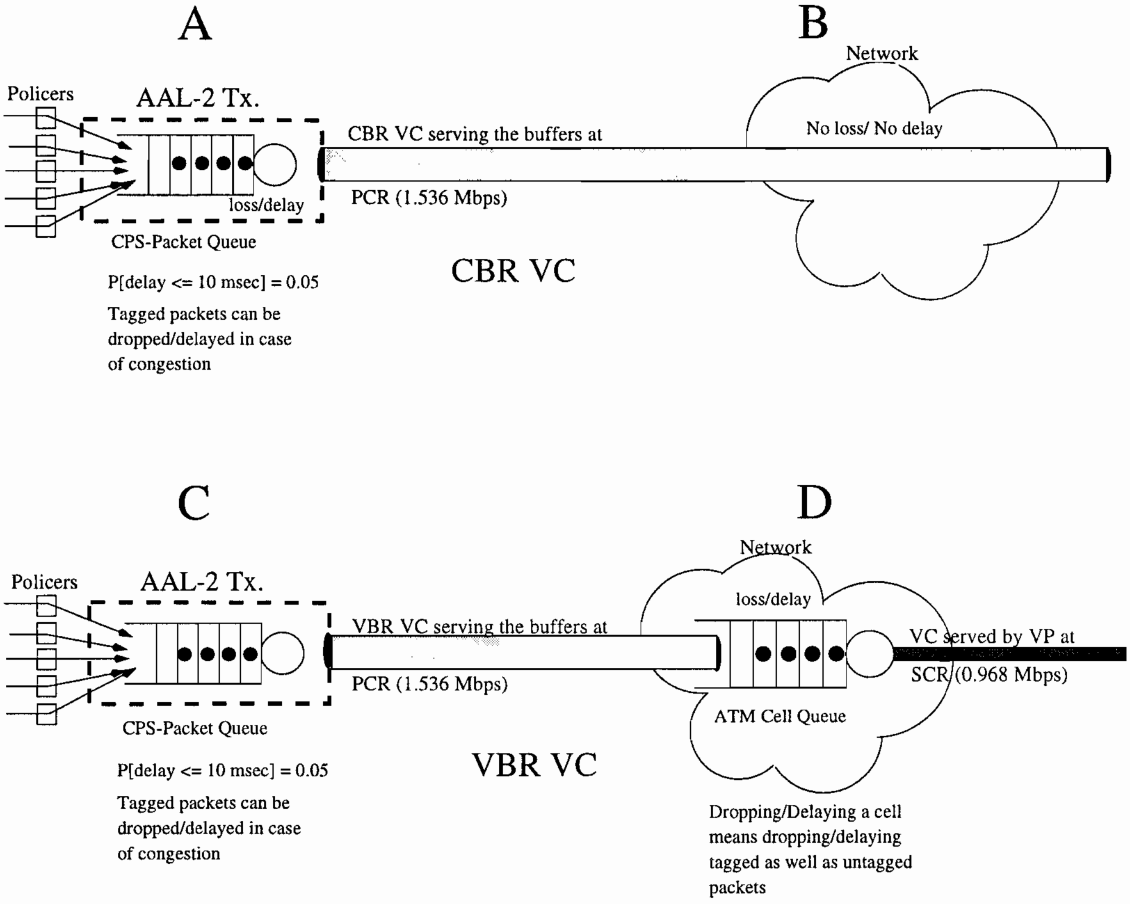

A
B
Network
Policers
AAL-2 Tx.
CBR VC serving the buffers at
No loss/ No delay
loss/delay
PCR (1.536 Mbps)
CPS-Packet Queue
CBR VC
P[delay <= 10 msec] = 0.05
Tagged packets can be dropped/delayed in case of congestion
C
D
Network
Policers
AAL-2 Tx.
loss/delay
VBR VC serving the buffers at
VC served by VP at
PCR (1.536 Mbps)
SCR (0.968 Mbps)
ATM Cell Queue
CPS-Packet Queue
P[delay <= 10 msec] = 0.05
VBR VC
Tagged packets can be dropped/delayed in case of congestion
Dropping/Delaying a cell means dropping/delaying tagged as well as untagged packets

Figure 3.4: Points of Buffering in CBR VC and VBR VC

3.4 Advantages and disadvantages of each option

CBR VC: **Advantages**:

- Individual user QoS can be guaranteed while maximizing the resource utilization with the help of an effective AAL2 CAC scheme and policers at connection level.
- Traditional CAC schemes used for ATM can be be mapped for AAL2 with slight modifications.
- The AAL2 CAC is totally independent of ATM CAC (ATM CAC ceases to exist if we assume that the VCs are always up) or other network entities or conditions like buffers at VP level or traffic on other VC sharing the same VP.
- Since the connection can be supported end to end if it can be supported at the AAL2 VC, the connection set-up time is small compared to VBR VC.

CBR VC: **Disadvantages**:

- Link utilization is not maximum, since the traffic is arriving to AAL2 buffers at $\mathbf{MAR_u}$ but served at $\mathbf{PAR_u}$. In our example, 11% of the bandwidth is wasted [4].
- Irrespective of the usage we have to pay for whole CBR bandwidth (T1 in our case).

Possible solutions:

- Develop a feedback based source rate control scheme (especially if they are all voice sources), which when coupled with an effective AAL2 CAC scheme will help in multiplexing more users on a single VC while reducing the degree of bandwidth wastage.

VBR VC: **Advantages**:

- On an average, in the homogeneous environment (the parameters and QoS requirements of all the users is same), bandwidth utilization at the VP level is better than it is in the case of CBR.
- Cheaper compared to CBR VC which has same PCR as VBR VC's PCR.

VBR VC: **Disadvantages**: See figures 3.3 and 3.4,

- QoS guarantees are difficult to determine. The ability to support user's QoS guarantees depends on various factors like the buffer size at VP level and traffic carried in other VCs sharing the same VP.

[4] $(1 - \frac{(72)(0.42)(32*10^3)(\frac{15}{12})(\frac{53}{47})}{1.536*10^6}) = 11\%$

- CAC is very difficult (or complex) and new questions associated with CAC arise as to:
  - ◇ What should be the traffic descriptor triplet for this VC .
  - ◇ How to map individual user QoS into VC's QoS (aggregated QoS). If this mapping is not done properly, we might lose whatsoever advantage we get in terms on bandwidth utilization with VBR VC.
  - ◇ The AAL-2 CAC has to have knowledge of ATM (network) QoS, and perhaps buffers and traffic at VP.
- Fairness among users admitted by AAL-2 CAC is very difficult to ensure.
- Individual user traffic cannot be monitored after multiplexing (first multiplexing is done before VC), so the behavior of every user affects other users' QoS guarantees (since only $\mathbf{SCR}_{vc}$ amount of bandwidth is guaranteed end-to-end but the traffic allowed into the network is on the basis of $\mathbf{PCR}_{vc}$ of VC *token rate*). See Figure 3.4 for better understanding.
- End-to-end connection establishment time may be large since in this case (VBR VC), the availability of bandwidth at VC doesn't necessarily mean the availability of bandwidth end-to-end.
- Complexity of the network is increased. With VBR VC, we need ATM-CAC and VC policers for traffic management.

Summary: In the real time applications like Voice/Video Over ATM, which are very delay sensitive, we should strive to guarantee the Quality of Service parameters while trying to maximize the link utilization. As explained in this section, VBR VC makes it more difficult to deliver the promised QoS, and it fails to ensure fairness amongst the users admitted into the network. Selection of VBR VC complicates the CAC schemes and introduces interdependencies between many network entities.

By selecting the VC as CBR VC, we can overcome the problems mentioned above and develop independent and effective CAC schemes. With the help of some reactive schemes like dynamic source-coding rate control to act as a back up for preventive CAC schemes, we can improve the link utilization with small and predictable degradation in user QoS.

Also due to the presence of other traffic classes like ABR and UBR, the wasted amount of bandwidth may be very small. In the next section we will look at different static CACs for AAL2.

Chapter 4

Static CAC for AAL2

An important functionality of traffic control in ATM is *Connection Admission Control (CAC)*. A connection can only be accepted if sufficient network resources are available to establish the connection end to end at its required quality of service. Also, the agreed QoS of pre-established connections in the network must not be adversely influenced by this new connection. Hence, we need to design appropriate resource allocation schemes which maximize network utilization (income) while guaranteeing the users' QoS. The uniform CAC framework should have an ideal resource allocation policy and should be robust and simple. The robustness requires that the control functions should be capable of accommodating future services and present heterogeneous sources. The need for simplicity means that the CAC must be applicable in real-time and implementable from a practical standpoint.

Call admission schemes are broadly classified into static and dynamic CACs. The static CAC has the property of assigning a fixed amount of bandwidth to every connection at the time of connection setup. The 'fixed amount' of bandwidth allocated is calculated using the traffic descriptors (PCR,SCR,MBS) of the connection. A resource allocation policy can also be based on other descriptors, for example, 'on time' and 'off time' of a voice source. At the onset of a connection request, a certain amount of bandwidth, calculated on the basis of the traffic descriptors, is allocated to the connection and removed from the 'total available bandwidth'. At the time of 'connection close', this bandwidth is restored to the 'total available bandwidth' pool. There is a separate 'total available bandwidth' pool for each link or virtual path (VP) in the network.

Unlike static CAC, dynamic CAC changes the amount of bandwidth allocated to any user during the call. This kind of CAC is particularly useful in cases where the rate changes slowly compared to the round-trip time. A related technique is feedback based source rate control, in which sources adjust their traffic in response to the network conditions. The scope of the present research includes a study of feedback rate control applied to Voice and Telephony Over ATM (VTOA) applications. This issue is addressed in more detail in the chapter 5.

Many studies/comparisons [5] have been done on CAC schemes. Those that use bandwidth as the decision criterion include the equivalent bandwidth based schemes [2], the Gaussian approximation approach [6] and heavy traffic approximations [7]. Their results indicate that they are all simple to be used in CAC, but they are all too conservative due to underlying approximations. There are other CAC schemes which base their CAC procedure on 'statistical bandwidth' based on closed-form expressions that use diffusion models [4, 3]. In this chapter, four CAC schemes are compared in the AAL2 scenario.

1. Fluid-flow analysis [1]
2. Equivalent Bandwidth, [2]
3. NEC Multi-class CAC [3]
4. Diffusion Based CAC [4]

The purpose of this chapter is five-fold:

1. To apply each of these CAC schemes to an AAL2 CAC environment.
2. To compare the accuracy of each of these CAC schemes with results obtained via simulation.
3. To find an optimum value of CPS packet size for given user bit rate, link rate and on-time and off-times for which number of users that can be supported on the link is maximized while maintaining the promised QoS of all the users.
4. To pick a particular CAC scheme which is well suited for this scenario (VTOA) while achieving the above mentioned goals.
5. To outline the need for and approach to a new feedback based voice coding rate control scheme.

4.1 Introduction to CAC Schemes

4.1.1 Equivalent Capacity Based Schemes

- Goal: To represent the effective bandwidth used by connections and the corresponding effective load on the network for desired QoS.
- Approach based on two complementary approximations
  - ◇ Fluid flow model
  - ◇ Stationary bit rate distribution
- Assumptions:

- ◇ Two state fluid flow model,
- ◇ Source behavior represented by connection metric vector (R_{peak}, ρ, b), where R_{peak} is the peak rate of the connection, ρ is utilization and b is the mean burst period.
- ◇ Finite buffer size and constant service rate C.

Background for Fluid Flow Model [1]: AMS method

Fluid-Flow Model

- Useful when impact of individual connection characteristics is critical.
- Steps involved:
  - ◇ Find distribution of buffer contents in terms of (R_{peak}, ρ, b) and 'C' (Link Capacity).
  - ◇ It is an iterative numerical technique to find the equivalent capacity for given a number of connections [1]. The various probabilities F_i(x), that the buffer does not exceed x given that i sources are on, are gathered into a vector **F**(x), which can be calculated from:

$$\mathbf{F}(x) = \sum_{i=1}^{N} a_i \mathbf{\Phi}_i e^{z_i x} \tag{4.1}$$

  where the z_i and $\mathbf{\Phi}$ are, respectively, generalized eigenvalues and eigenvectors associated with the solution of the differential equation satisfied by the stationary probabilities of the system, and the a_i's are coefficients determined from the boundary conditions [1, 2]. The distribution **F**(x) is completely determined from the values of the associated eigenvalues, eigenvectors, and corresponding coefficients. The equivalent capacity corresponding to a set of multiplexed sources can then be obtained again using iterative numerical techniques, where at each iteration a new solution to equation 4.1 must be computed [2].

- Problems involved:
  - ◇ Not robust (cannot be generalized for any type of source) and not computationally simple.
  - ◇ Due to lack of closed form solution to achieve our goal, it is incompatible with a dynamic and real time environment.
  - ◇ Need for approximations.
- Asymptotic approximation

- ◇ Approximate buffer overflow probability $G(x) = \beta e^{z_0 x}$ where β is a constant term independent of buffer size, x. z_0 is the largest negative eigenvalue corresponding to equation 4.1.
  Using $\beta = 1$ will yield $\frac{lnG(x)}{x} = z_0$
  Using z_0 and equation 4.1, we can obtain the value of the equivalent capacity $\hat{C}_{(F)}$ given by the flow approximation for N multiplexed sources, $\hat{C}_{(F)} = \sum_{i=1}^{N} \hat{c}_i$ where $\hat{c}_i$, equivalent capacity for individual connections, are determined from equation 4.2 below.
- ◇ Assumptions:
  Large buffer, small number of connections.
- ◇ Incompatible with dynamic and real time environments.

Equivalent Capacity By Guerin method [2]

- Case 1: Effect of statistical multiplexing is less.
  - ◇ Simplification of Eqn.(1) is obtained by assuming $\beta = 1$ where β is a function of C, ρ, R, and the overflow probability, ϵ.
  - ◇ Equivalent Capacity of Single Source is given by

$$\hat{c} = \frac{a - K + \sqrt{(a-K)^2 + 4Ka\rho}}{2a} R \ \ where \ \ a = ln(1/\epsilon)b(1-\rho)R \tag{4.2}$$

  - ◇ Large buffer, Small number of connections.
- Case 2: Effect of Statistical multiplexing dominant.
  Stationary bit rate model (SBR)
  - ◇ β is significantly different from 1 when a number of connections with equivalent capacity much larger than their mean bit rate are multiplexed which is essentially the case for connections with long burst periods and low utilization.
  - ◇ $G(x) < \epsilon$ is ensured but smoothing effect of the buffer is neglected.
  - ◇ SBR distribution is approximated by Gaussian.

$$\hat{C}_{(s)} = m + \alpha' \sigma \ \ where \ \ \alpha' = \sqrt{-2ln(\epsilon) - ln(2\pi)} \tag{4.3}$$

  where m and σ are mean and standard deviation for aggregate bit rate.
- Total bandwidth used by 'N' connections is given by:

$$\hat{C} = min\left[\hat{C}_{(s)}, \sum_{i=1}^{N} \hat{c}_i\right] \tag{4.4}$$

- Admission procedure
  - ◇ Origin node computes request vector $\mathbf{r_i} = (m_i, \sigma_i^2, \hat{c}_i)$ from $(R_{peak}^{(i)}, \rho_i, b_i)$.
  - ◇ Request vector is sent to all nodes on the route.
  - ◇ If $\hat{C}$ is available on all nodes, accept the connection.
  - ◇ Update the link metric vector.

4.1.2 Statistical Bandwidth Based CAC schemes

Features

- Based on closed form solutions that use diffusion models.
- Takes into consideration the user's cell loss requirements, their aggregate traffic characteristics and the available buffer size at the statistical multiplexers.
- Captures the interaction between individual traffic streams at the ATM multiplexers.

NEC's CAC scheme: Background [3]

Assumptions:

- M x M input/output buffered switch.
- Overall Buffer is partitioned between classes.
- Frame-Based scheduling
  Allocating C_k bandwidth to class k = assigning n_k time slots to this class out of each frame, where

$$n_k = \left\lceil N_f \frac{C_k}{C_{link}} \right\rceil$$

  N_f represents the total number of time slots available in a frame.
- QoS requirements are in terms of: CLR $< \epsilon_1$ and $Prob[CDV > t_{max}] < \epsilon_2$
- **Goal:** Find effective required bandwidth given UPC parameters of New + Existing connections, ϵ_1, ϵ_2 and t_{max}

NEC's CAC scheme for VBR sources [3]
Given the UPC parameters, $(\lambda_p^*, \lambda_s^*, B_s^*)$ of new connection, where $\lambda_p^*, \lambda_s^*, B_s^*$, are, respectively, peak cell rate, sustainable cell rate and maximum burst size.

- Construct fictitious "on/off" source model using

$$T^*_{on} = \frac{B^*_s}{\lambda^*_p}$$
$$T^*_{off} = \frac{B^*_s}{\lambda^*_p} \frac{(\lambda^*_p - \lambda^*_s)}{\lambda^*_s} \tag{4.5}$$

- *Lossless Multiplexer Model*
    - ◇ Calculate C^{new}_{vbr} using eqn.

$$C^{new}_{vbr} = max\left((\lambda^*_p + \sum_{i=1}^{n} \lambda^i_p)(1 - \frac{B_{vbr}}{B^*_s + \sum_{i=1}^{n} B^i_s})^+, \lambda^*_s + \sum_{i=1}^{n} \lambda^i_s\right) \tag{4.6}$$

    where B_{vbr} is the buffer allocated for VBR traffic class and $[x]^+$ means max(0,x).
    - ◇ If CDV is not satisfied, recalculate C^{new}_{vbr} replacing B_{vbr} by $C^{new}_{vbr} * t_{max}$
    - ◇ Additional bandwidth required to Support New Connection

$$\delta_1 = C^{new}_{vbr} - C^{old}_{vbr} \tag{4.7}$$

- *Statistical Multiplexer Model* To Find C^{new}_{vbr}
    - ◇ Construct a modified "on/off " source model

$$\lambda^*_H = min(1, \frac{T_{on}}{T_N})\lambda^*_p + [1 - \frac{T_{on}}{T_N}]^+\lambda^*_s;$$
$$\lambda^*_L = [1 - \frac{T_{on}}{T_N}]^+\lambda^*_s$$

    where T_N is time required to empty half-filled buffer.
    - ◇ Calculate C^{new}_{vbr} using eqn.

$$C^{new}_{vbr} = \bar{M}_{new} + \zeta * \sigma_{new} \tag{4.8}$$

    where $\bar{M}_{new}$ and σ_{new} are the mean and variance of the aggregate arrival process due to all admitted VBR connections after the new VBR connection is admitted. ζ is dependent on $min(\epsilon_1, \epsilon_2)$

- Find additional bandwidth required to support the new connection

$$\delta_2 = C^{new}_{vbr} - C^{old}_{vbr} \tag{4.9}$$

- Admit the new connection if capacity $\Delta_{vbr} = min[\delta_1, \delta_2]$ is available in the free pool of bandwidth (i.e. $\Delta_{vbr} \leq C_f$) and $C^{new}_{vbr} \leq C^{max}_{vbr}$.

Diffusion Based Statistical CAC [4]

Diffusion Approximations

- Continuous approximations to the discontinuous arrival and service process.
- Require first two moments of inter-arrival and service times. Methods are known to compute these.
- Accurate for traffic models using on-off sources and their generalizations.
- Can be obtained in simple closed forms.

Finite Buffer Diffusion Cell Loss Estimate (FBDCLE)

- $L_{FB} = \Psi \ e^{\frac{2(B-1)}{\alpha}\mu} Pr[R(t) \geq C]$
  where Ψ is function of instantaneous average rate of change of buffer contents(μ), instantaneous variance of change of buffer contents (α), and E[H] which is defined as average holding time at buffer limit. R(t) is the instantaneous arrival rate and B is the buffer size in cells.

Infinite Buffer Diffusion Cell Loss Estimate (IBDCLE)

- $L_{IB} = \gamma \ e^{(\frac{2B}{\alpha}\mu)} \frac{E[(R(t)-C)^+]}{\lambda}$
  where γ is function of μ, α and E[h] defined as average holding time at the lower boundary (buffer fill=0). λ is the aggregate cell arrival rate. $(R(t) - C)^+$ implies max(0,R(t)-C).

Approximations done to carry out computation of effective bandwidth in real time:

- $\Psi \ e^{\frac{-2\mu}{\alpha}} < 1$ and $\gamma < 1$
  so that $L_{FB_{bound}} > L_{FB}$, $L_{IB_{bound}} > L_{IB}$

Set of Equations:

- C_{df_1}: Statistical Bandwidth by diffusion model for finite buffer is obtained with the help of L_{FB} and the abovementioned approximations,

$$C_{df_1} = \lambda - \delta + \sqrt{\delta^2 - 2\sigma^2\omega_1} \tag{4.10}$$

  where δ is function of buffer size (B), α and variance of arrival rate (σ). ω_1 is a function of CLR.

- C_{df_2}: Statistical Bandwidth by diffusion model for infinite buffer is obtained with the help of L_{IB} and the abovementioned approximations,

$$C_{df_2} = \lambda - \delta + \sqrt{\delta^2 - 2\sigma^2\omega_2} \tag{4.11}$$

where ω_2 is a function of CLR, σ and λ.

Since α and μ can be calculated using peak rate, mean rate and burst length, C_{df_1} and C_{df_2} are functions of these standard parameters.

Call Admission Control Procedure

- Connection established only if enough bandwidth is available on every intermediate link along the selected path.
- Every link along the path has an information vector:

$$\mathbf{I} = \left\{ \sum_{u=1}^{N} \lambda_u, \sum_{u=1}^{N} \sigma_u^2, \sum_{u=1}^{N} \lambda_u c_u^2 \right\}$$

  containing status of current connections on each corresponding link.

- Let C_{df} be the statistical bandwidth, C_l be link capacity and $\mathbf{U} = \{\lambda_U, \sigma_U^2, \lambda_U c_U^2\}$ be the information vector of new connection.
- For each link along the selected path
  - ◇ Update **I** ← **I+U**
  - ◇ Calculate C_{df}
  - ◇ If $C_{df} \leq C_l$ **Accept**
    Else If $C_{df} > C_l$ **Reject and Restore I ← I − U**

4.2 CAC For AAL2: Methodology

The function of AAL2 CAC is to decide whether to accept or reject a connection depending on the connection identifiers and the resources available with in the network.The AAL2 CAC calculates the effective bandwidth required to support this connection by taking into account its descriptors like coding rate, CPS packet size, mean on-time and mean off time.

Due to the overhead involved per cell, the arrival rate as seen by the AAL2 CAC is more than the coding rate of the user. This increase in the arrival rate is incorporated in the analysis by multiplying the coding rate of the user with a factor called 'overhead factor'.This overhead factor (assuming that there are enough users to fill a cell) is calculated as follows:

$$Overhead\,Factor = \left(\frac{ATM\,Cell\,Size}{Max\,CPS\,PDU\,Size}\right) * \left(\frac{Packet\,Size + 3}{Packet\,Size}\right) \tag{4.12}$$

$$= \left(\frac{53}{47}\right) * \left(\frac{Packet\,Size + 3}{Packet\,Size}\right)$$

$$Effective\,Peak\,Rate = Overhead\,Factor * Coding\,Rate \tag{4.13}$$

For small number of users, the probability of transmitting filled cells is very small. Hence the overhead factor will be more than that calculated with the help of equation 4.12. The exact value of overhead factor is calculated by taking into account the total number of users, their average on time and the CPS packet size. Figure(4.1) indicates the overhead factor calculated for different number of users and different CPS Packet sizes. For illustration purposes, let us consider an example where two users each with a coding rate of 32 *kbps*, on with a probability of 0.42 (P_{on}) and CPS packet size 'C' bytes are sharing a common link. The overhead factor corresponding to different CPS packet sizes is as follows:

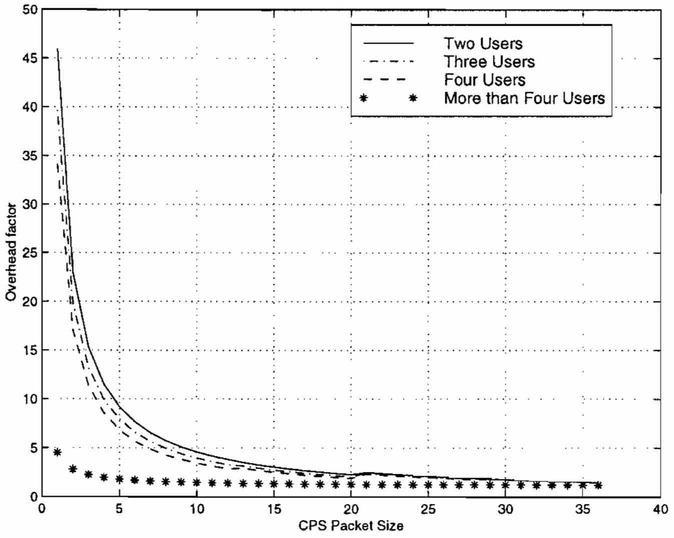


Figure 4.1: Overhead factor per CPS Packet Size for given number of users

- **Case 1:** ($C \leq 20$)

  a) Both the sources are on: Since the packet size is less than or equal to 20, we can pack the packets coming from both the sources into one single ATM cell. In this case the overhead factor is given by $\frac{53}{2C}$

  b) One of the two sources is on: Overhead factor is given by $\frac{53}{C}$.

Therefore the average overhead factor is given by

$$Overhead\, Factor = \frac{\frac{P(2,2)*53}{2C} + \frac{P(2,1)*53}{C}}{P(2,1) + P(2,2)} \tag{4.14}$$

where $P(n,k) = \binom{n}{k} P_{on}^{k} * (1 - P_{on})^{(n-k)}$

- **Case 2:** $(20 < C \leq 40)$

  a) Both the sources are on: Since the packet size is greater than or equal to 20, we cannot pack the packets coming from both the sources into one single ATM cell. In this case the overhead factor is given by $\frac{53*2}{2C}$. The difference between case (1a) and (2a) is that in (2a) we need to have two cells to accommodate two packets coming from users (this is taken care of by a constant '2' in the numerator) whereas in (1a) we need only one ATM cell fit two CPS packets.

  b) One of the two sources is on: Overhead factor is given by $\frac{53}{C}$. Note that there is no difference between case (1b) and (2b), because CPS packet of any size (maximum packet size allowed for a delay QoS of 10 m seconds is equal to 40 bytes) can be fit into an ATM cell. Therefore the average overhead given that one or two sources are on, is given by

$$Overhead\, Factor = \frac{\frac{P(2,2)*53*2}{2C} + \frac{P(2,1)*53}{C}}{P(2,1) + P(2,2)} = \frac{53}{C} \tag{4.15}$$

Continuing with our example, for a CPS packet size of 20, equation (4.12) results in an overhead factor of 1.2968 where there are many users, whereas equation (4.14) yields 2.2978 when there are only two users.

As we can see from Figure (4.1) as the number of users increases, the overhead factor approaches the limit given by equation (4.12). We have done detailed overhead calculations using the approach mentioned above only up to four users. For any number of users greater than four, we approximated the overhead factor by equation (4.12). As the number of users increases, the number of breakpoints (the point C=20 in our example) in the analysis increases and for five users we saw that calculation becomes very cumbersome. The change in the way of overhead calculation (from detailed to approximation) results in a discontinuity in one of our results Figure 4.5(b).

In analysis as well as the simulation, the QoS parameter requirement for each user is taken to be identical. QoS requirement of $95^{th}\%ile$ delay $\leq 10msec$ means that 95 percent of the total transmitted packets will suffer a delay (within the transmitter) of less than or equal to 10 milliseconds.

The maximum delay any packet can suffer in the transmitter is:

$$Delay\, Bound = Packetization\, Delay + Max\, Queueing\, Delay$$

$$or \qquad Delay\,Bound = \left(\frac{CPS\,Packet\,Size}{Coding\,Rate}\right) + Max\,Queueing\,Delay$$

In our case,

$$Max\,Queueing\,Delay = Delay\,Bound(10\,msec) - Packetization\,Delay \tag{4.16}$$

We also know that

$$Max\,Queueing\,Delay = \left(\frac{Queue\,Size}{Link\,Rate}\right) \tag{4.17}$$

Equating the above two equations, we have,

$$\left(\frac{Queue\,Size}{Link\,Rate}\right) = Delay\,Bound - Packetization\,Delay \tag{4.18}$$

Therefore, the maximum queue size "x" is

$$Maximum\,Queue\,Size\,(x) = (Delay\,Bound - Packetization\,Delay) * Link\,Rate \tag{4.19}$$

So the QoS requirement as used in the analysis is

$$Prob[Queue\,Fill \leq x] = 0.95 \tag{4.20}$$

or

$$Cell\,Loss\,Probability \approx Prob[Queue\,Fill > x] = 0.05 \tag{4.21}$$

4.2.1 Fluid Flow Method for AAL2

In order to estimate the bandwidth required to support a connection, the fluid flow method requires the knowledge of peak rate, load, mean burst size, buffer size and link capacity. Peak rate can be obtained from equation 4.13. Load and mean burst size can be calculated using the following identities.

$$Load\ \ (\rho) = \frac{Mean\,On\,Time}{(Mean\,On\,Time + Mean\,Off\,Time)} \tag{4.22}$$

$$Mean\,burst\,Size = (Mean\,On\,Time) * (Effective\,Peak\,Rate) \tag{4.23}$$

$$Mean\,cell\,rate = P_{on} * (Effective\,Peak\,Rate) = (Load) * (Effective\,Peak\,Rate) \tag{4.24}$$

The buffer size can be obtained from equation 4.19. The required bandwidth as calculated using fluid flow model can be obtained using equations 4.1, 4.13, 4.19, 4.22, 4.23.

4.2.2 Equivalent Capacity Method for AAL2

To estimate the bandwidth required to support a connection and satisfy its QoS, the equivalent capacity method requires:

- Effective peak rate(P), load (ρ), mean burst size (b), link capacity(C)
- Buffer size(x)
- Mean(m) and variance (σ^2) of arrival rate.

First two components can be obtained with the help of equations 4.13, 4.19, 4.22, 4.23 where as the last component is obtained as follows:

$$Mean(m) = \rho P \tag{4.25}$$

$$Variance(\sigma^2) = \rho(P-m)^2 + (1-\rho)m^2 \tag{4.26}$$

The required bandwidth is obtained with the help of equations 4.2, 4.3, 4.4, 4.13, 4.19, 4.22, 4.23, 4.25 and 4.26.

4.2.3 NEC's CAC scheme for AAL2

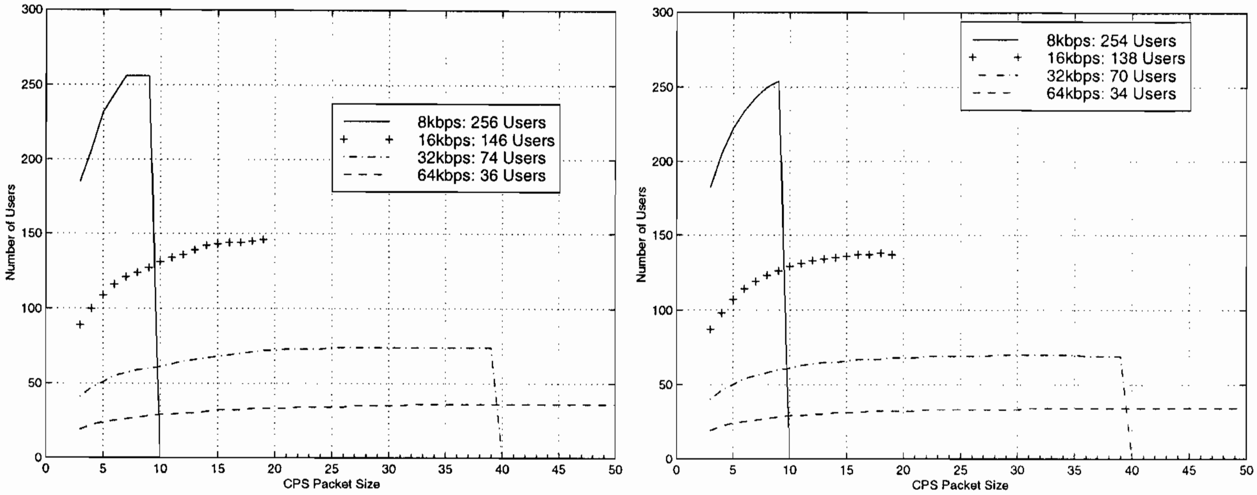
This CAC scheme requires the UPC triplet i.e. (PCR,SCR,MBS). In the literature [3], ways are described to map a source defined by UPC parameters to a worst case source defined by mean on time and mean off time (eq.4.6). This mapping is illustrated in equation 4.5. This CAC is in general defined for multi class traffic, but in our analysis we are making use of VBR traffic class alone. C_{vbr}^{new} is calculated with the help of eqn. (4.7) in which λ_p^*, B_{vbr}, B_s^*, λ_s^* can be obtained using equations 4.13, 4.19, 4.23, 4.24 respectively. Then the required bandwidth is obtained with help of equations , 4.7, 4.8, 4.9, 4.22, 4.25 and 4.26.

4.2.4 Diffusion Based CAC

This CAC scheme requires all the parameters required by the equivalent capacity CAC plus an extra parameter called squared coefficient of variation of inter arrival times, c^2 [9, 4]. For our analysis purposes, where we are considering sources to be on-off, transmitting at peak rate in the on period and transmission rate to be zero in off period, the squared coefficient of variation of inter arrival times of cells is given by

$$c^2 = \frac{1-(1-\alpha T_u)^2}{(\alpha T_u + \beta T_u)^2} \tag{4.27}$$

$$where \quad T_u = \frac{1}{P}, \quad \alpha = \frac{1}{T_{on}} \quad and \quad \beta = \frac{1}{T_{off}} \tag{4.28}$$



(a) As obtained from simulations

(b) As obtained from Anick et. al. analysis

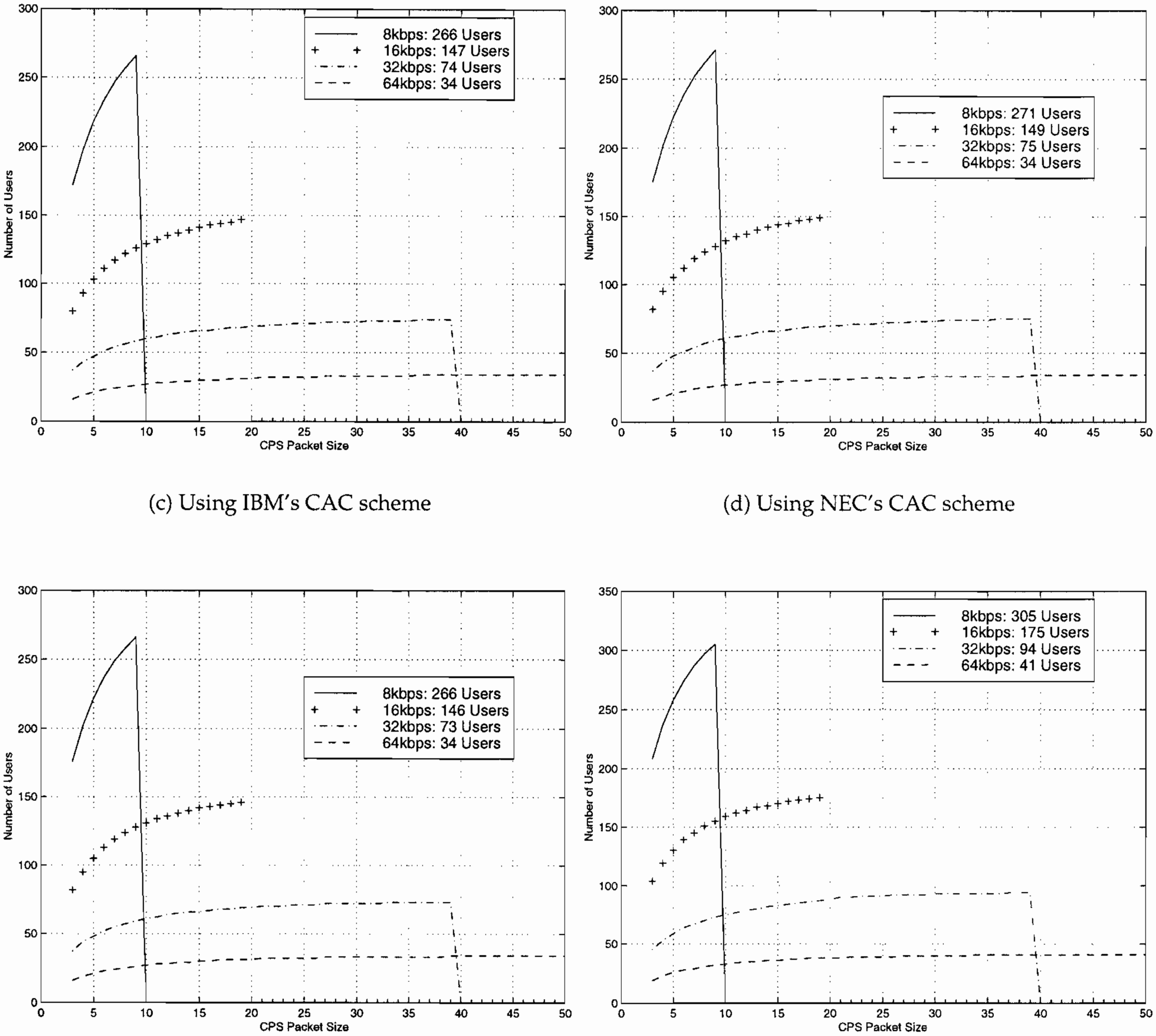
The required capacity is then obtained with the help of equations 4.10, 4.11, 4.13, 4.19, 4.22, 4.23, 4.25, 4.26, 4.28.

4.3 Comparison and Results

In this section we will compare different CAC schemes with simulation results [8]. For the purpose of comparison, all the traffic sources are assumed to be homogeneous, on-off type sources with on time equal to 420 milliseconds and off time equal to 580 milliseconds. The QoS parameters are as defined in equations (19) and (20).All the curves are generated with the help of MATLAB and the algorithms mentioned in section 4.2.

Figure 4.2(a) shows the maximum number of users that can be supported on a link serving at a rate of 1.536 Mbps (for a source of type described in [8]), for different CPS packet sizes and for different coding rates, as obtained from simulations [8].

Figure 4.2(b) plots the same but the results are obtained using the Anick, Mitra and Sondhi analysis [1]. This analysis is very elegant and is able to capture the effect of individual connections very well (when the impact of individual connection is very important, this analysis works best) but due to the absence of a closed form solution, it is impossible to use this in real-time or implement on a switch. For example, in figure 4.3(d), where each connection is transmitting traffic at a rate of 64 kbps so that the number of users is relatively small, this analysis tracks the simulation results best. Figure (4.2(c)) shows the maximum number of users that can be supported on a



(c) Using IBM's CAC scheme

(d) Using NEC's CAC scheme

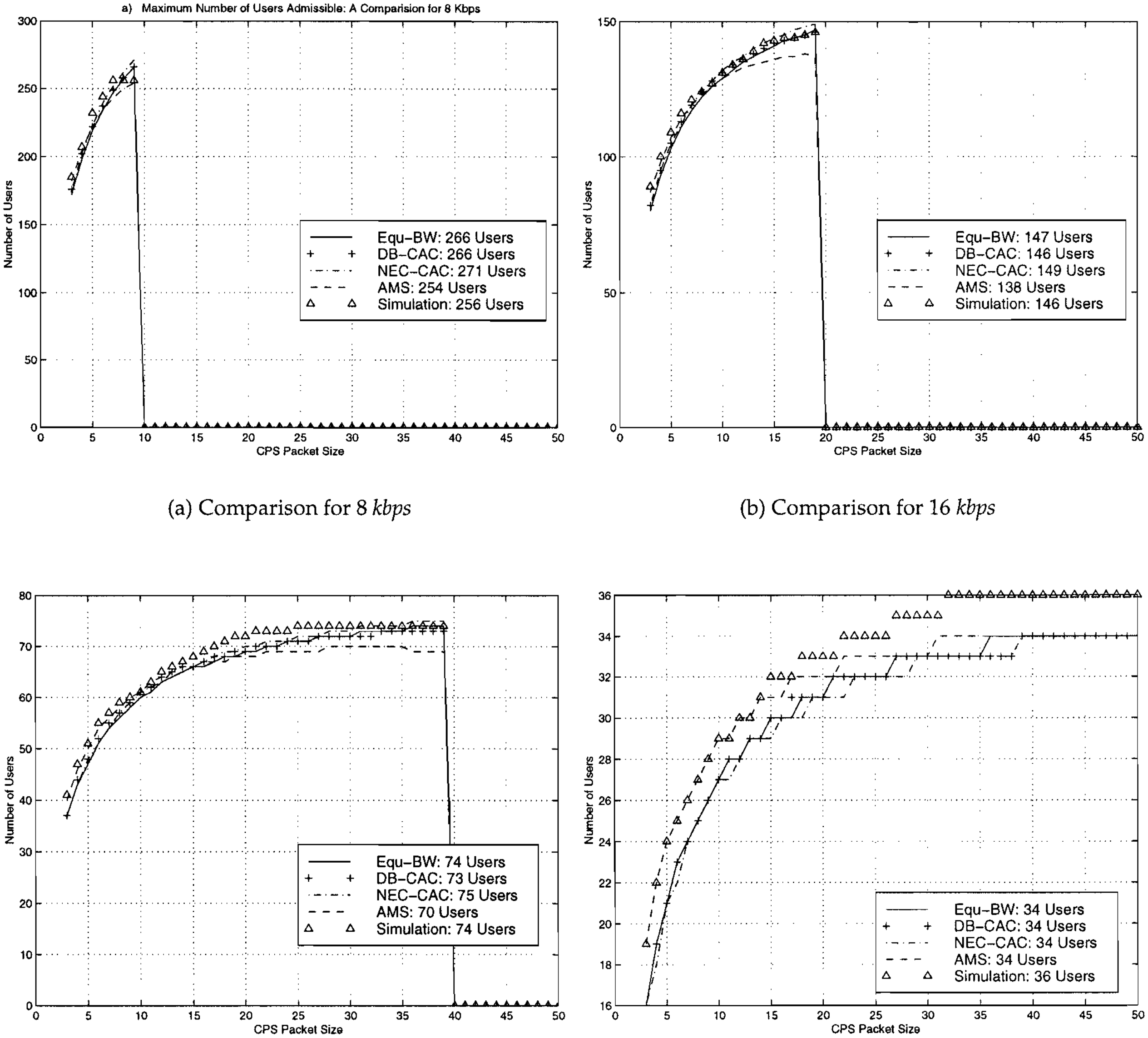
(e) Using finite buffer, diffusion based CAC scheme

(f) Using infinite buffer, diffusion based CAC scheme

Figure 4.2: Maximum number of sources admitted: Obtained from simulations and different CAC schemes link as obtained by IBM's CAC schemes discussed in subsection 4.1.1. Figure 4.2(d) plots the same but for the results obtained using NEC's CAC scheme described in subsection 4.1.2. Figure 4.2(e)) and 4.2(f) represents the maximum number of sources that can be admitted using FBDCLE and IBDCLE described in subsection 4.1.2. Figures 4.3(a), 4.3(b), 4.3(c) and 4.3(d), are comparison of the four CAC schemes and simulation results for different user coding rates (8 *kbps*,16 *kbps*,32 *kbps* and 64 *kbps* respectively). We can see that for large number of users all the CAC schemes perform more or less equally (Diffusion Based CAC is always closer to the simulation results). One important point to mention about figure 4.3(a) is that, in simulation, a cap was introduced on the maximum number of users supported on that VC (256 users). That is the only reason the results seem not to match in that region. If that cap is removed, we can say that all the CAC's admit nearly same number of sources.

All simulations were conducted with a $95^{th}\%ile$ delay bound of 10 ms, but our analysis allows us to quickly explore the effect of changing the percentile or the delay bound. Figures 4.4(a) and 4.4(b) indicate maximum number of users that can be supported by different CAC schemes for different $95^{th}\%ile$ and $99^{th}\%ile$ delay constraints, respectively. An important point to be noted is that the maximum number of connections that can be supported on the link (calculated by CAC schemes specified earlier) are calculated for optimum CPS packet size related to each delay value. As can be seen from the graph, at 10m sec we are near saturation (in terms of maximum number of users).

Our analysis also allows us to quickly determine the required link bandwidth for a given number of connections. With simulation this would require multiple simulation runs to "zero-in" on the required bandwidth. Of particular interest for residential and small business applications of AAL-2 is the question of required bandwidth for a small number of connections. As mentioned in section 4.2 equation (4.12) does not produce an accurate value for the overhead factor in such cases. Figure 4.1 indicates the overhead factor per CPS packet size for small number of users. It also shows the asymptotic overhead factor curve. Figure 4.5(a) then shows the total amount of bandwidth required to support a given number of connections. In this case the number of users are varied from 1 to 50 and corresponding "required bandwidth" (specified by each CAC scheme) is plotted. Figure 4.5(b) which plots additional bandwidth required for each new connection, is just a demonstration of the ability of the CAC schemes to exploit the statistical multiplexing gain property. In figure 4.5(b) there is no data point corresponding to five users. At five users we changed from one method of overhead calculation to the other (from detailed calculation to approximation with equation(4.12)) as discussed in section 4.2. It can be seen from Figures 4.5(a) and 4.5(b) that all the three CAC schemes allocate the same amount of bandwidth to a new connection *asymptotically*. In AMS analysis, IBM's equivalent bandwidth approach or NEC's CAC scheme, the sources are considered to be on-off with exponentially distributed on and off times. This may not be a particularly accurate source model, but it is probably adequate when many on-off voice sources are aggregated. For *small number of sources* scenario, any inaccuracy in the assumed on-period distribution may be more important. The three CAC schemes did not propose any solution for this. However, NEC's Multiclass CAC scheme seems to be close (but conservative) to the sim-



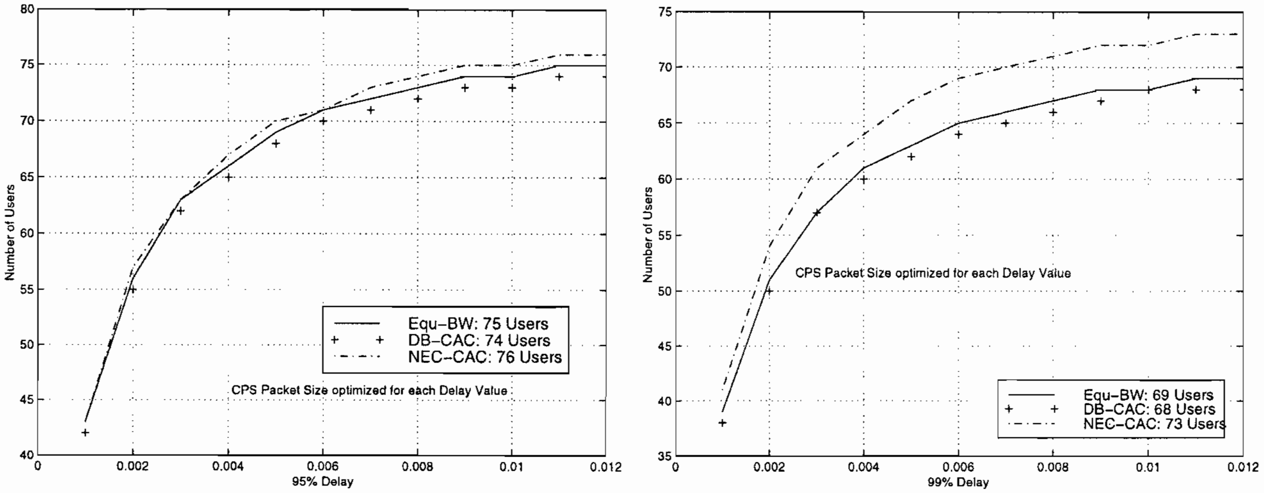
(a) Comparison for 8 *kbps*

(b) Comparison for 16 *kbps*

(c) Comparison for 32 *kbps*

(d) Comparison for 64 *kbps*

Figure 4.3: A comparison of different CAC schemes with simulation results for different coding rates



(a) Maximum number of users supported as a function of delay QoS and for corresponding optimum packet size

(b) Maximum number of users supported as a function of delay QoS and for corresponding optimum packet size

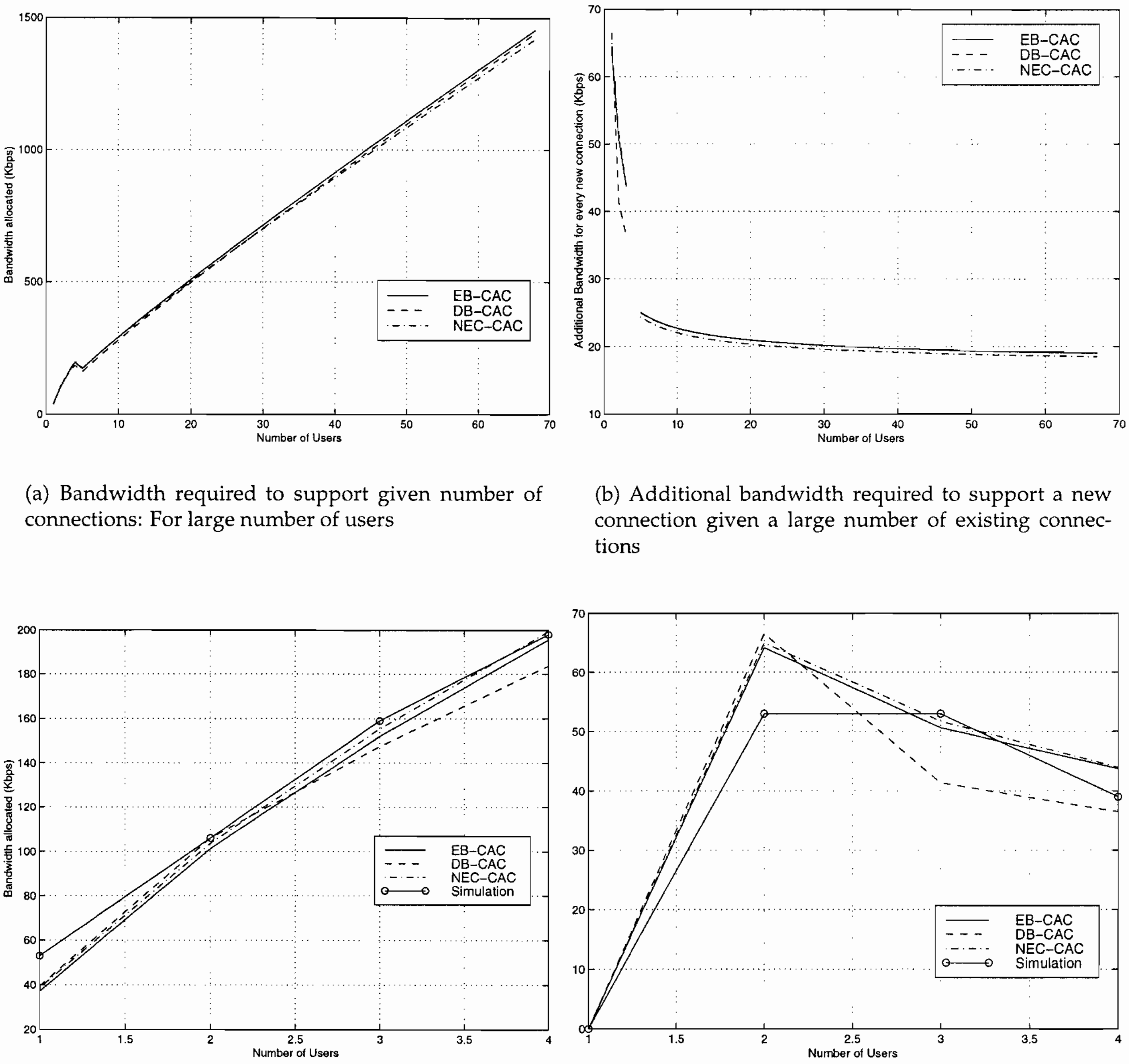
Figure 4.4: Illustration of effects of different delay QoS on CAC schemes

ulation results. One of the reasons could be because of its ability to capture the burstiness factor effectively (due to *modified worst case on-off source*) compared to other CAC schemes (figures 4.5(c) and 4.5(d)). Figures 4.5(c) and 4.5(d) take a closer look at Figures 4.5(a) and 4.5(b), varying the number of users from 1 to 5. This closer look greatly highlights the abovementioned property. Note that the results here assume that there is enough load on the link so that there is no affect of timer-CU (discussed in chapter 5) in our calculations and this assumption will effect our calculations when the number of users are very small. However, if our goal is to find the maximum number of sources that can be accommodated on a link and the sources on-period is such that $mean\_on\_period >> \frac{1}{LinkRate}$, the results and analysis will hold good.

4.4 Next step

As can be seen from Figures 4.5(c) and 4.5(d), there is a limit to the bandwidth conservation with the help of any static CAC. Since the on period of the voice source is very large, significant amount of gain (better link utilization) can be obtained with the help of feedback control of voice coding rate. Steps to be followed involve

- Introduce a technique in which the voice coding rate changes dynamically in response to a



(a) Bandwidth required to support given number of connections: For large number of users

(b) Additional bandwidth required to support a new connection given a large number of existing connections

(c) Bandwidth required to support given number of connections: For small number of users

(d) Additional bandwidth required to support a new connection given a small number of existing connections

Figure 4.5: Total bandwidth required to support a given number of connections and Additional bandwidth required to support a new connection: An illustration for statistical multiplexing gain feedback signal.

- Defining the instances when this feedback signal is triggered, for example, in response to congestion at the multiplexing buffers etc.
- Defining the approach that will be adopted in response to the feedback signal, for example, in the case of congestion, different approaches can be tested. 1) Up/Down coding. 2) Embedded Coding, where LSBs are dropped in response to congestion.
- Identifying the scenarios where this kind of scheme is best suitable.
- Identifying the optimum values for the parameters required for the above mentioned scheme, for example, queue fill thresholds which will trigger feedback signals.
- Finding, with the help of simulations, the efficiency achieved by using feedback control, i.e. how many extra users we can support without violating the QoS ?

Chapter 5

Dynamic Source-Coding Rate Control

It has been proved by simulations [8] that AAL-2 performs better in terms of bandwidth savings compared to other AALs. In the previous chapters, we have seen CAC schemes and underlying framework (in the context of AAL-2) which helps in bandwidth savings while satisfying the user guarantees. Though the bandwidth savings are significant when compared with other AALs, still there is an underutilization of approximately 11% of the link bandwidth. This underutilization has to be borne with in order to satisfy users' QoS. In this chapter, we study a feedback based dynamic rate control scheme, which adjusts the coding rate of the variable bit rate coder residing inside the AAL-2 transmitter based on packet-queue-length information. This scheme is found to be appropriate in relieving congestion and increasing the link utilization at the expense of "graceful degradation" of user QoS.

The earliest research involving reactive congestion control schemes for relieving congestion in voice networks were embedded coding schemes. In these schemes, the information bits were divided into most significant bits (MSBs) and least significant bits (LSBs) [17]. These code bits may be placed in separate packets or arranged in sub-blocks within the same packet. While traversing the network, if the packet finds any congested nodes or links, the LSBs were discarded thereby reducing the effective coding rate of the source. An important advantage of such schemes based on embedded coding is the ability to quickly and accurately reduce the coding rate of the arriving traffic at the points of congestion. There are some disadvantages associated with such schemes.

- In order to achieve the embedded property, one has to compromise with certain set of coding schemes which fits the constraints of embedded algorithms.

- Carrying the packets up to the point of congestion and then discarding them reduces the resource utilization and in case of long duration congestion, this disadvantage becomes very prominent.

- Excessive discarding of packets/sub-blocks in response to a long duration congestion results in a processing burden on high-speed switching of traffic.

To overcome these drawbacks, research was focused on developing schemes in which the source coding rate is adjusted at the source node in response to the congestion information. Bially *et al.* [29] suggested an end-to-end feedback technique in which the packets were dropped at the intermediate nodes resulting in a lowered received bit rate at the receiver. Every voice terminal will report its received bit rate to the other terminals. The encoders respond to this congestion information by discarding the low-priority sub-blocks before submitting the traffic to the network. This scheme still has the disadvantages of the embedded coding schemes.

Yin *et al.* [13] suggested a dynamic rate control in which the congestion information is carried by a single bit in the voice packets and returned to the source by the destination. The source only switches to a higher rate after the congestion in network subsides. The main advantages of this scheme are that unlike the other control algorithms, this control mechanism does not need involvement of intermediate nodes for dropping sub-blocks or packets and thereby reducing the burden on intermediate nodes. It is highly desirable to keep the intermediate nodal processing independent from the source coding process because it allows an easy migration to new coding schemes without the need to change the subnetwork. The disadvantages associated with this technique are

- There is a delay in sending the information on traffic loads on the intermediate nodes in the network back to the source node. Due to this delay, a source cannot react to the congestion immediately and continues sending the data at higher rate and thereby worsening the congestion situation.
- In case the feedback information is lost in the network, it will escalate the congestion.

In the work presented here, the bandwidth required to support a given number of users is calculated using their coding rate, packet sizes and the QoS constraints. The analytical results obtained in 4.2 are compared with the simulation results obtained using the modeling done in BONeS. Furthermore, the bandwidth efficiency achieved using dynamic source coding rate control technique in the context of voice over ATM using AAL2 is evaluated by incorporating the feedback based rate control in the AAL2 transmitter model. The model is discussed in detail in next section. It is found that with DRC, the number of users on a link can be increased but while doing so the received voice quality is degraded. This results due to the transition from high coding rate to low coding rate in the event of congestion and also due to the total number of transitions from high-to-low and low-to-high coding states. In order to reduce the high-to-low transitions, the packet queue thresholds (discussed in subsection 5.1.3) have to selected carefully. To overcome the problem of *total* number of state transitions, the decision to change the state is taken on the basis of *filtered* queue length information instead of *instantaneous* queue information. The nature of the filter, its parameters and their effects are discussed further in subsection 5.1.3.

5.1 System Description

The simulation of an AAL2 network with variable bit rate voice sources is used to evaluate the performance of *dynamic rate control* (DRC) scheme in VTOA. Modeling and simulations are done using Block Oriented Network Simulator (BONeS) [30]. A brief description of network models, location and nature of few network components and the simulation parameters is presented in this section. In order to do a fair comparison between the system with and without DRC, we have carried out simulations with same set of parameters. The performance metric used is $95^{th}\%ile$ delay, average bit rate of the source and number of transitions experienced by the source from state of higher coding rate to a state of lower coding rate. The block diagrams of DRC scheme are shown in figure 5.1 and 5.4.

5.1.1 Voice source model

All the voice sources used in this project [1] are homogeneous on-off sources. All the voice calls are assumed to be independent of each other. Each voice source alternates between talk-spurt and silence periods, which are assume to be exponentially distributed with mean values 0.42 seconds and 0.58 seconds. In the on state the voice source transmits data at its peak cell rate [2], whereas in the off state the voice source remains idle. This results in a speech activity factor given by

$$\rho = \frac{0.42}{0.42 + 0.58} = 0.42$$

5.1.2 AAL2 CPS Procedure

The CPS consists of distinct transmission and reception state machines that function independent of each other. The transmission state machine needs to multiplex the various channels into as few ATM SDUs as possible, while still maintaining the time requirements of the CBR traffic, while the reception state machine needs to demultiplex channels that can be spread over multiple ATM SDUs.

AAL2 Transmitter

The multiplexing function in the CPS to merge several different sized streams into a single ATM SDU requires a method for scheduling these streams so that none of the streams suffer any more

[1] Most of the network models and parameters used are reused from [8] for purposes of fair comparison and verification of system performance. Though most of them are modified to take into account the feedback property, the basic design and logic remained the same.

[2] In our system, we have selected the packet size is 32 bytes and peak cell rate of each source to be equal to 39.468 *kbps* (coding rate of 32 *kbps* plus an overhead of $(\frac{53}{47})(\frac{Packetsize+3}{Packetsize})$

than acceptable delays. The nature of traffic on AAL2 channels require a CPS SDU to be transmitted within a certain time frame after it is generated. In algorithmic form, the CPS transmitter has the following procedure [10, 24]

1. AAL2 transmitter has four states IDLE, PART, FULL and SEND.
2. The CPS transmitter starts in the IDLE state. When the transmitter receives a CPS PDU, its builds the CPS Packet Header, STF and start Timer_CU.
3. If a previous part or whole CPS SDU is waiting for transmission (state is SEND or FULL), then append the current SDU to the waiting ATM SDU for transmission.
4. If no data is waiting to be transmitted (State is IDLE), then start a new ATM SDU.
5. If the STF needs to be built, use the end of the waiting data as the start pointer in the STF.
6. If Data Queued < ATM SDU size, set state as PART, set the Part variable for future STF calculations (if needed), wait for new CPS SDU. When the SDU arrives, jump to step 2.
7. If the ATM SDU is filled, stop the Timer_CU, queue it for transmission.
8. If data remains to be transmitted, start Timer_CU, build the STF, jump to step 2.
9. Set State as IDLE, jump to step 1.
10. If the Timer_CU expires, build an STF if none is built (using the start of the padding as the start pointer in the STF), pad the remaining part of the ATM SDU with 0's, jump to step 6.

An important and implementable assumption which we made in our scheme of DRC is the AAL2 transmitter and coder are **co-located**. The reasons for which this architecture is worth considering are:

1. Co-location of coder and AAL2 transmitter is a feasible and reasonable design choice.
2. Due to co-location of the AAL2 transmitter (and hence the packet queue which is responsible for feedback signal) and coder, there is essentially no feedback delay.
3. Since the AAL2 transmitter is located at the edge nodes, DRC has all the advantages of traditional edge-based coding rate control mechanisms [13].
4. If the connected sources are all analog sources or constant bit rate sources, it is prudent to have a common coder, because this results in simplicity of network architecture (for example, transcoding has to be done for only two bit rates).
5. Our model is accurate for coder and AAL2 not co-located so long as coder uses embedded coding and AAL2 can discard the LSB sub-blocks or packets in response to the congestion (traditional embedded coding scheme advantage).

AAL2 Receiver

The Reception state machine is simpler because the time dependence of the channels is taken care of while transmitting. However, packet discard is an important step in the receiver, since entire ATM SDUs cannot be discarded if any one channel has errors. Brief summary of AAL2 receiver is presented [10, 24]

1. The CPS receiver starts in IDLE State.
2. When a new ATM SDU arrive, obtain the offset value, check the STF parity.
3. Discard erroneous SDUs, and any waiting CPS SDU. If the offset is not zero, append data up to the offset from the ATM SDU to the waiting CPS SDU.
4. If the CPS SDU is > maximum data size for the channel, discard CPS SDU.
5. If more data remains in the ATM SDU, start a new CPS SDU, populate channel number field. If the channel number is 0, the rest of the ATM SDU is a pad, discard the ATM SDU, jump to step 1. If no more data, jump to step 1.
6. If data remains in the ATM SDU, find the length of the channel. If length is greater than Max channel length, discard the CPS SDU. Jump to step 4.
7. If data remains in the ATM SDU, get the HEC, perform Header Error Correction. If error detected, discard CPS PDU, jump to step 5.
8. If data remaining in the ATM SDU >= length of channel, transfer channel information to the CPS SDU, hand the SDU to the higher layer, jump to step 5.
9. If no more data remains, and new channel number exists, jump to step 1.
10. If data remaining in the ATM SDU < length, transfer remaining data from ATM SDU to CPS SDU, set state = PART, jump to step 2.

The AAL2 transmitter is developed in detail in [8] and is in accordance to the specifications in [10]. Figure 5.1 shows the high level block diagram of the dynamic source coding rate control scheme implemented in the context of AAL 2.

5.1.3 Simulation parameters

1. QoS parameters

    - Maximum delay bound: Delay is a main performance metric for any voice application. In packetized voice, two main components of delay (other than fixed transmission delay for a given link) are packetization delay and queuing delay. In order to avoid the

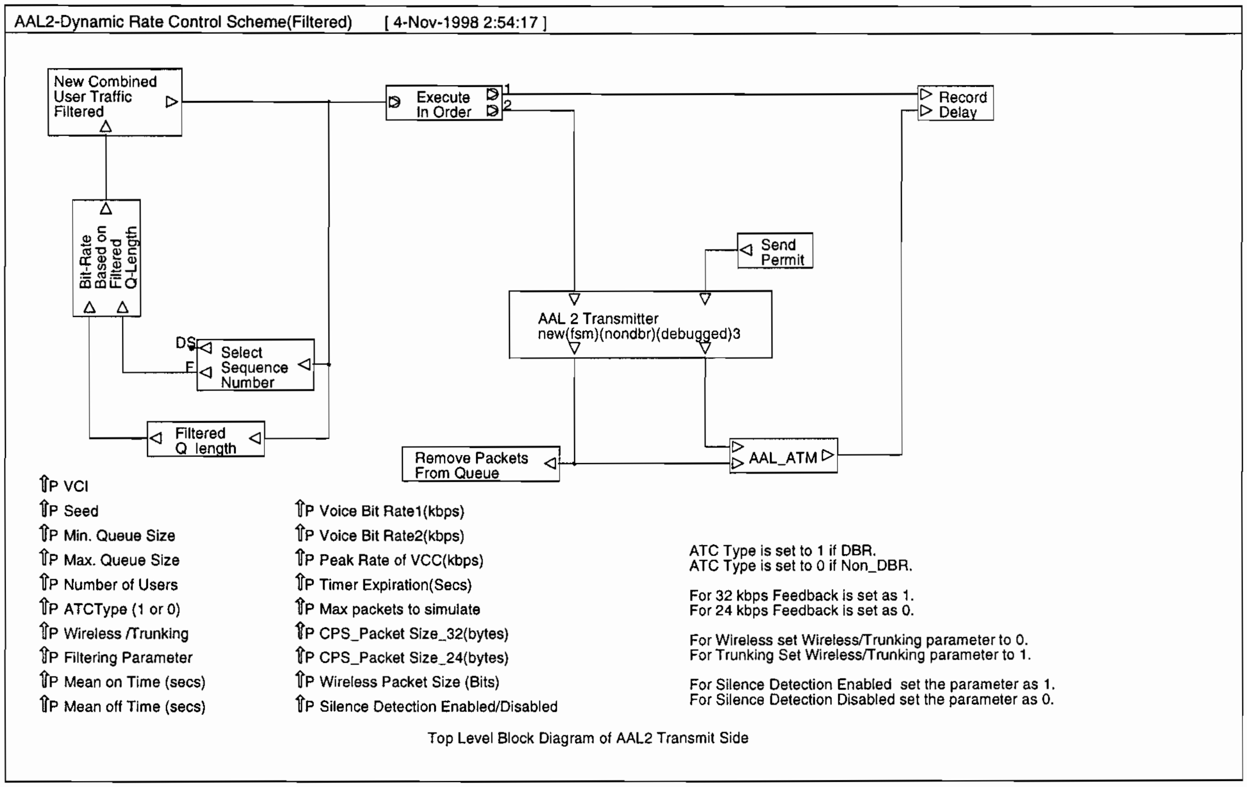

AAL2-Dynamic Rate Control Scheme(Filtered) [4-Nov-1998 2:54:17]
New Combined User Traffic Filtered
Execute In Order
1
2
Record Delay
Bit-Rate Based on Filtered Q-Length
Send Permit
AAL 2 Transmitter new(fsm)(nondbr)(debugged)3
DS
F
Select Sequence Number
Filtered Q length
Remove Packets From Queue
AAL_ATM
P VCI
P Seed
P Min. Queue Size
P Max. Queue Size
P Number of Users
P ATCType (1 or 0)
P Wireless /Trunking
P Filtering Parameter
P Mean on Time (secs)
P Mean off Time (secs)
P Voice Bit Rate1(kbps)
P Voice Bit Rate2(kbps)
P Peak Rate of VCC(kbps)
P Timer Expiration(Secs)
P Max packets to simulate
P CPS_Packet Size_32(bytes)
P CPS_Packet Size_24(bytes)
P Wireless Packet Size (Bits)
P Silence Detection Enabled/Disabled
ATC Type is set to 1 if DBR.
ATC Type is set to 0 if Non_DBR.
For 32 kbps Feedback is set as 1.
For 24 kbps Feedback is set as 0.
For Wireless set Wireless/Trunking parameter to 0.
For Trunking Set Wireless/Trunking parameter to 1.
For Silence Detection Enabled set the parameter as 1.
For Silence Detection Disabled set the parameter as 0.
Top Level Block Diagram of AAL2 Transmit Side

Figure 5.1: High-level block diagram of AAL2 transmitter with DRC

excessive usage of echo-cancellers, the total round trip delay should be kept below 50 ms. In this study we have restricted a maximum delay of 10 ms inside the AAL2 transmitter [8] leaving 15 ms for other delays like network queuing delay, propagation delay, coder/decoder delay etc. A connection is supported by AAL2-CAC if it can provide less than or equal to 10 ms delay inside the transmitter.

- Quality of voice: In this project we are not doing any subjective test on the quality of received voice. In order to estimate the quality degradation experienced, we are using metrics like average bit rate of a connection, percentage of time a source stays in the state of lower coding rate, number of transitions from one state to other.

2. Voice bit rate: Two bit rates are considered in the experiments: 32 *kbps* and 24 *kbps*. The bit rates present a trade-off between number of connections supported on the link and the quality of the received voice. Taking lower bit rate (for example 24 *kbps*) during congestion might increase the number of connections on the VC at the expense of poorer voice quality.

3. CPS Packet size: The CPS packet size is selected such that the packetization delay remains constant (8 ms). In our experiments CPS packet size is 32 bytes and 24 bytes for 32 *kbps* and 24 *kbps* coding rates respectively.

4. Timer setting: The CU-timer dictates the maximum time the AAL2 transmitter waits before transmitting a partially filled cell. At higher link rates and higher loads the effect of CU-timer is negligible. In the earlier chapter, where we focused mainly on the maximum number of connections that can be supported (representing higher load region), and therefore our results were not very much dependent on timer value. In the experiments carried out in this project the timer value is set to the maximum possible value which is the difference between maximum allowed delay and packetization delay (2 ms) at higher loads and at lower loads we decreased the value of CU-timer to 1 ms. As a part of the work done here, we tried to investigate the affect of different values of CU-timer on the users' 95^{th} percentile delay in different load scenarios. The results of the simulations are presented in figure 5.2 (or tables 5.1 and 5.2). The results can be interpreted as follows: in low load scenarios (small number of users), the 95^{th} percentile delay is very sensitive to the values of CU-timer whereas for higher loads (approximately ≥ 0.5) the delay is relatively insensitive to CU-timer (supporting the assumption made in section 4.2 and stated earlier in this para). Another interesting point to note is that for CU-timer value of 1ms, the 95^{th} percentile delay is approximately same for every load scenario. The estimation of optimum range of CU-timer values to suit every is a subject of futher study.

No. of users	CU Timer Values						
	0.5 ms	1 ms	1.5 ms	2 ms	2.5 ms	3 ms	3.5 ms
5	8.5010e-03	9.0010e-03	9.5010e-03	1.0001e-02	1.0501e-02	1.1001e-02	1.1501e-02
10	8.5010e-03	9.0010e-03	9.5010e-03	1.0001e-02	1.0501e-02	1.1001e-02	1.1501e-02
20	8.5010e-03	9.0010e-03	9.5010e-03	9.9990e-03	1.0501e-02	1.0999e-02	1.1501e-02
30	8.5490e-03	9.0010e-03	9.5010e-03	9.9990e-03	1.0245e-02	1.0301e-02	1.0319e-02
50	8.6110e-03	9.0010e-03	9.3550e-03	9.3850e-03	9.3890e-03	9.3870e-03	9.3910e-03
65	8.7230e-03	8.9990e-03	9.0690e-03	9.0790e-03	9.0810e-03	9.0810e-03	9.0810e-03
72	8.7950e-03	8.9850e-03	8.9890e-03	8.9910e-03	9.0010e-03	8.9950e-03	8.9950e-03
76	8.8190e-03	8.9710e-03	8.9570e-03	8.9570e-03	8.9570e-03	8.9570e-03	8.9570e-03
80	8.8090e-03	9.1090e-03	8.9550e-03	8.9490e-03	8.9530e-03	8.9510e-03	8.9510e-03
85	8.8830e-03	8.9610e-03	8.9510e-03	8.9450e-03	8.9470e-03	8.9450e-03	8.9450e-03

Table 5.1: Effect of CU-timer on $95th$ $\%ile$ delay of the users. Note that link rate is kept constant at 1.536 Mbps; Timer value varied from 0.5 ms to 3.5 ms.

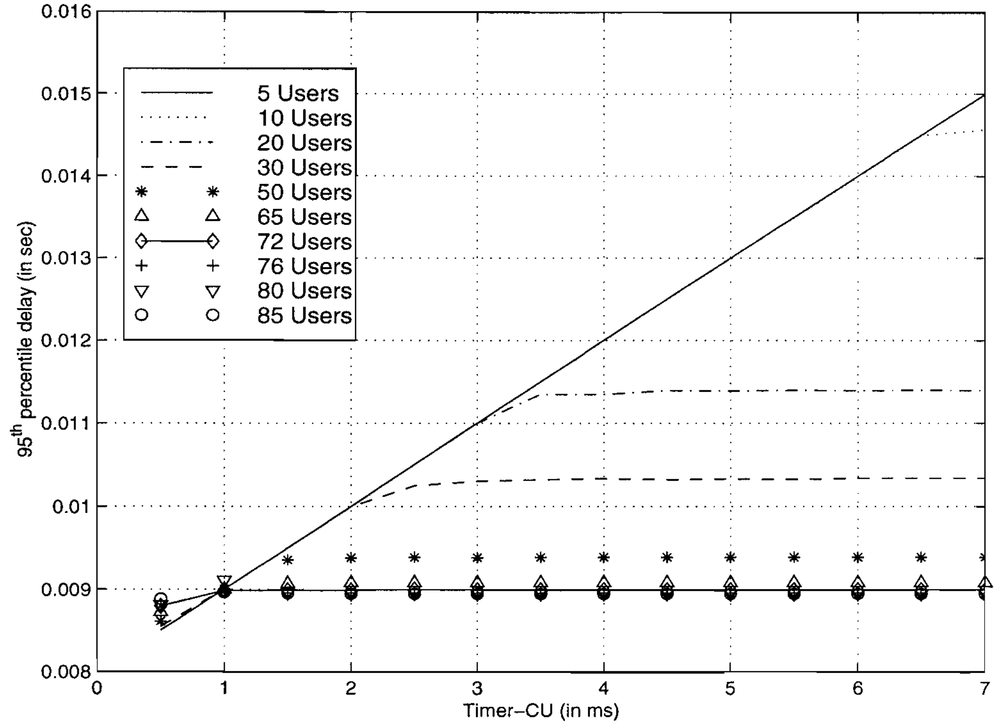


Figure 5.2: Illustration of impact of CU-timer on 95^{th} percentile delay in different load scenarios

5. Maximum and minimum queue thresholds: The packet queue in the AAL2 transmitter is monitored and depending on the queue fill the rate of the coder is adjusted. There are two thresholds defined for the queue as shown in figure 5.3, Q_{min} and Q_{max}. If the queue fill is more than Q_{max} the coder will always operate at 24 *kbps* and if the queue fill is less than Q_{min}, the coder will operate at 32 *kbps*. For queue fills between Q_{max} and Q_{min}, the coder remains in the earlier state. This gives rise to hysteresis and the efficiency of the system is estimated by looking at the percentage of time the coder operates at the higher level (32 *kbps*) of hysteresis curve. Value of queue threshold Q_{max} was selected using equation 4.19 which is

   $$Q_{max} = (Delay\, Bound - Packetization\, Delay) * Link\, Rate$$

   Later it was realized that the queue fill rarely reaches this limit (less than 5% of the time) and it stays there only momentarily. If we use a filter, the percentage of time the queue fill reaches this limit is nearly equal to zero. For the DRC scheme to be useful, the queue thresholds must be selected such that queue *fill* oscillates between Q_{max} and Q_{min}. Our

No. of users	CU Timer Values						
	4 ms	4.5 ms	5 ms	5.5 ms	6 ms	6.5 ms	7 ms
5	1.2001e-02	1.2501e-02	1.3001e-02	1.3501e-02	1.4001e-02	1.4501e-02	1.5001e-02
10	1.2001e-02	1.2501e-02	1.3001e-02	1.3501e-02	1.3999e-02	1.4499e-02	1.4569e-02
20	1.1347e-02	1.1395e-02	1.1389e-02	1.1397e-02	1.1395e-02	1.1397e-02	1.1395e-02
30	1.0331e-02	1.0319e-02	1.0327e-02	1.0321e-02	1.0337e-02	1.0337e-02	1.0337e-02
50	9.3910e-03	9.3910e-03	9.3910e-03	9.3910e-03	9.3910e-03	9.3910e-03	9.3910e-03
65	9.0810e-03	9.0810e-03	9.0810e-03	9.0810e-03	9.0810e-03	9.0810e-03	9.0810e-03
72	8.9950e-03	8.9950e-03	8.9950e-03	8.9950e-03	8.9950e-03	8.9950e-03	8.9950e-03
76	8.9570e-03	8.9570e-03	8.9570e-03	8.9570e-03	8.9570e-03	8.9570e-03	8.9570e-03
80	8.9510e-03	8.9510e-03	8.9510e-03	8.9510e-03	8.9510e-03	8.9510e-03	8.9510e-03
85	8.9450e-03	8.9450e-03	8.9450e-03	8.9450e-03	8.9450e-03	8.9450e-03	8.9450e-03

Table 5.2: Effect of CU-timer on $95th\ \%ile$ delay of the users. Note that link rate is kept constant at 1.536 Mbps; Timer value varied from 4 ms to 7 ms.

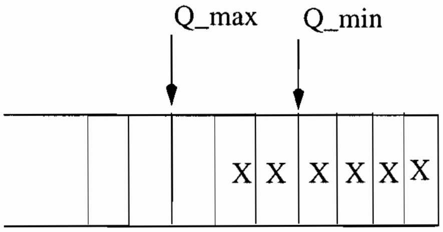


Figure 5.3: Illustration of queue thresholds

approach for the solution was to include a *Load* factor into the Q_{max} calculation, i.e.,

$$Q_{max} = \frac{(Delay\,Bound - Packetization\,Delay) * Link\,Rate}{(Packet\,size * 8)} * Load \tag{5.1}$$

$$where\,Load = \frac{(Number\,of\,users) * (0.42)(32)}{Link\,Rate}$$

where 0.42 is mean on-time, 32 kbps is source coding rate and link rate is 1.536 Mbps. The resulting Q_{max} in terms of *packets*. Q_{min} has to be selected such that the coder is in low-coding rate state as little as possible; also, the frequency of state changes can be minimized. By trial and error this optimal Q_{min} was found to be *two packets* less than Q_{max}. Note that the optimal value might change for different parameters like link rate, QoS constraints, etc.

One of the extensions to the current work could be to find an optimal value of Q_{min} which depends on Q_{max} and all the factors on which Q_{max} depends.

6. Time constant or filtering constant: One of the problems associated with DRC is the number of transitions between the two coding-rate states. It is desirable to have a small number of transitions from one state to another. To minimize the number of transitions, we are basing our decision to change states on the *filtered* queue-length information. This filtering is done by a first-order recursive filter

   $$\hat{Q_t} = \tau.Q_t + (1-\tau)\hat{Q_{t-1}}$$

   where $\hat{Q_t}$ and Q_t are the filtered and instantaneous queue lengths, respectively. The value of τ is chosen to be between 0 and 1. For $\tau = 1$, filtered and instantaneous queue lengths are identified. In our experiments we chose $\tau = 0.05$. Figure 5.4 represents how the filter is incorporated into the DRC scheme for AAL2.

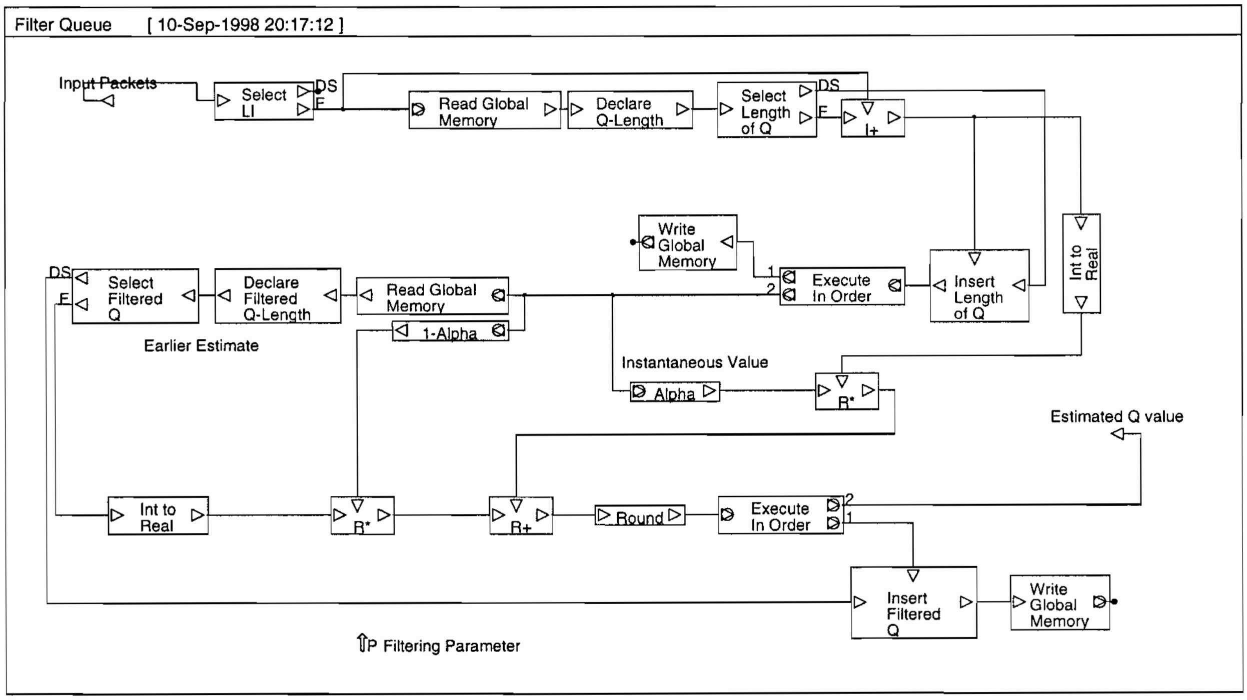


Figure 5.4: Deciding the coding rate of coder in AAL2 transmitter on the basis of filtered queue length information

5.2 Results

5.2.1 Small number of users

The first set of experiments is to investigate the efficiency of DRC in small number of users scenario. The bandwidth required to support different number of connections using AAL2 with and without DRC is tabulated in table 5.3. All the users are transmitting 32 byte packets at a rate of 32 *kbps*. From the simulations it is inferred that for small number of users, bandwidth savings cannot be achieved by coupling AAL2 with any kind of scheme. It can be argued that with proper selection of Q_{min} and Q_{max}, one can get improvement in the bandwidth required to support certain number of connections. However, this improvement can be achieved only at the expense of QoS degradation. This trade-off between bandwidth improvement and QoS degradation is very high for small number of users. In these scenarios (a small number of connections), the queue fill remains constant or it changes by a factor of 1 packet. If the selected Q_{max} and Q_{min} are such that the distance between them is more than one packet, the coding rate of the source remains in the low-bit-rate state for long time (once the coder goes to low bit-rate state, it remains there forever) which affects the QoS. If we try to selected Q_{max} and Q_{min} such that the difference between Q_{max} and Q_{min} is small (say, 1 packet), the total number of transitions between the states are very high and thereby the QoS again gets affected. In our work, we are giving more importance to QoS over bandwidth utilization.

	BW required w/o feedback		BW required with feedback	
No. of users	Total (kbps)	Avg (kbps)	Total (kbps)	Avg (kbps)
1	53	53	53	53
2	106	53	106	53
3	159	53	159	53
4	198	49.5	196	49
5	215	43	210	42
6	255	42.5	248	41.3
7	265	37.5	258	36.9
8	289	36.12	282	35.25
9	310	34.4	300	33.3
10	334	33.4	325	32.5
25	712	28.48	698	27.92
72	1536	21.33	1230	17.08

Table 5.3: Bandwidth required to support given number of users with different systems; Note: Filtering parameter of 0.05 is used

5.2.2 Conventional AAL2 vs. DRC coupled AAL2

For large number of users ($\approx$ 25 in our case), the delay experienced by each packet can be reduced considerably with the help of the feedback based DRC scheme. This also allows us to increase the number of users that can be supported on the link. Figures 5.5(a) and 5.5(b) illustrate the significant reduction in the delay that can be obtained with the help of DRC scheme over conventional AAL2 transmitter without any form of feedback.Figure 5.6 compares the 95^{th} percentile delay experienced by different number of users in conventional system and system with DRC. As can be seen from figure 5.6, we can support more number of users ($\approx$ 25% more users) on the given link for same delay [3] constraint (10 ms). We have to realize that we are achieving this increase in link utilization at the expense of QoS degradation. As mentioned earlier, this QoS degradation is measured using metrics like effective average bit rate of all the users in system,transitions between the two states and percentage time in the lower state.

Figure 5.7(a) shows the average bit rate of each user in the system with DRC compared to a straight line at 32 *kbps* in the case of conventional AAL2. With the increase in number of users, the transitions between the two states as well as the percentage time in the lower state will increase as shown in figure 5.7(b) and figure 5.7(c), respectively. In the case of AAL2 without DRC, they will be straight horizontal lines at zero.

5.2.3 DRC based AAL2 with fixed queue thresholds vs. variable queue thresholds

With DRC scheme, the average bit rate and $95^{th}\%ile$ delay are within the acceptable limits but the only metric of concern is the number of transitions. In order to reduce the number of transitions, we have tested two methods:

1. Varying the minimum and maximum thresholds in correspondence to the load (as opposed to fixed[4] thresholds)

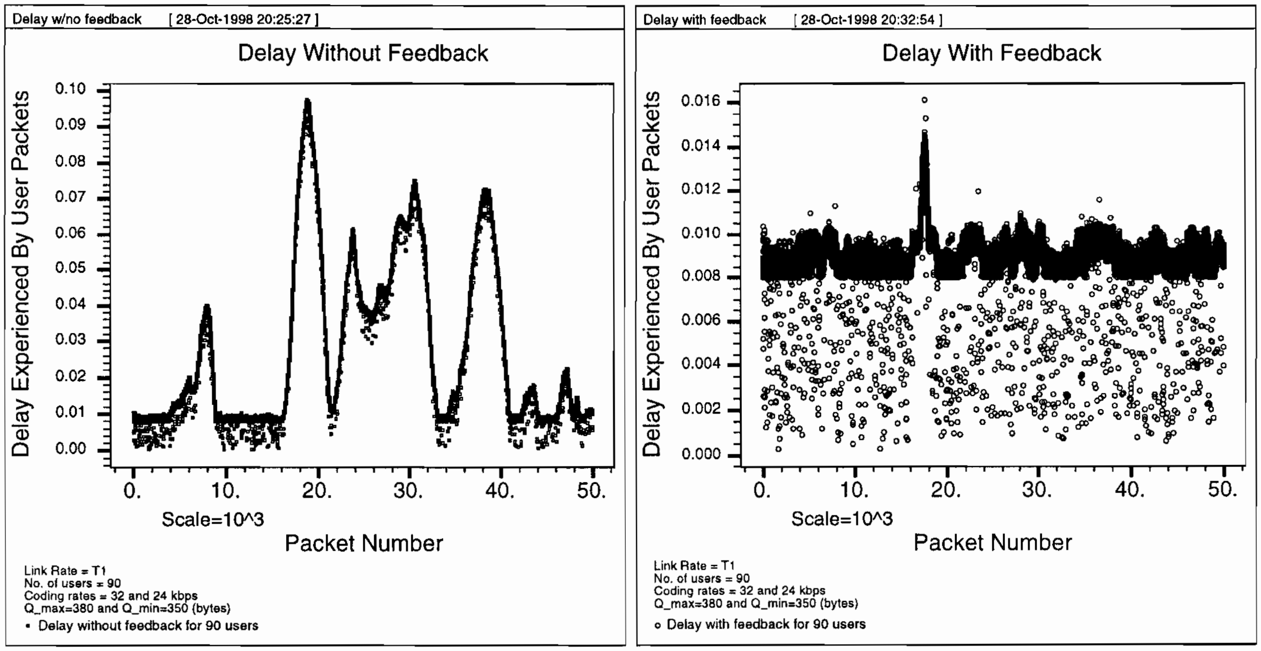
2. Introducing a first-order recursive filter through with the packet-queue length information is passed. In next subsection, we will look at this method in detail.

Figure 5.8(a) shows the number of transitions that are experienced by the first option which is varying the queue threshold[5]. It is found that there is no significant change in the number of transitions. The gain achieved (reduction in the number of transitions for number of users in the

[3] An important point to note in figure 5.6 is that the $95^{th}\%ile$ delay curve for 'system without feedback' remains at 15 ms for any number of users greater than 80. This is due to the modeling assumption (any delay greater than 15 ms is taken as 15 ms) which was done for ease of implementation. This assumption does not have any effect on the experiments we have performed. In reality, this curve will rise.

[4] The term *fixed* means that the queue thresholds are independent of load and they remain same for any number of users on a given link

[5] Remember that the queue threshold is dependent on the load and hence the threshold will change with number of users.



(a) Delay experienced by user packets in a system without feedback

(b) Delay experienced by user packets in a system with feedback

Figure 5.5: Demonstration of improvement in delay experienced by user packets using DRC scheme

range of 65 to 80) is at the expense of extra delay experienced (but still within the limits) for same number of users as shown in figure 5.8(b). An accurate way of defining queue thresholds and its relationship with load has to be found.

5.2.4 DRC coupled AAL2 system with filter vs. without filter

A first-order recursive filter was used. The inputs to the filter are *instantaneous queue length* and *filtered queue length at the previous instant.* The output of the filter is based on these inputs and the *constant τ (or filtering constant).* The sampling period of the queue is equal to one packet transmission time. The value of filtering constant(τ) is chosen to be 0.05.

With the filter option, a significant reduction in the number of transitions ($\approx$ 80%) is experienced. Figure 5.9(a) illustrates this point and we can see that the slope of the filtered curve is much smaller compared to the curve without filter. Figures 5.9(b), 5.9(c), 5.9(d) show that this reduction in total number of transitions is achieved with no degradation of other metrics.

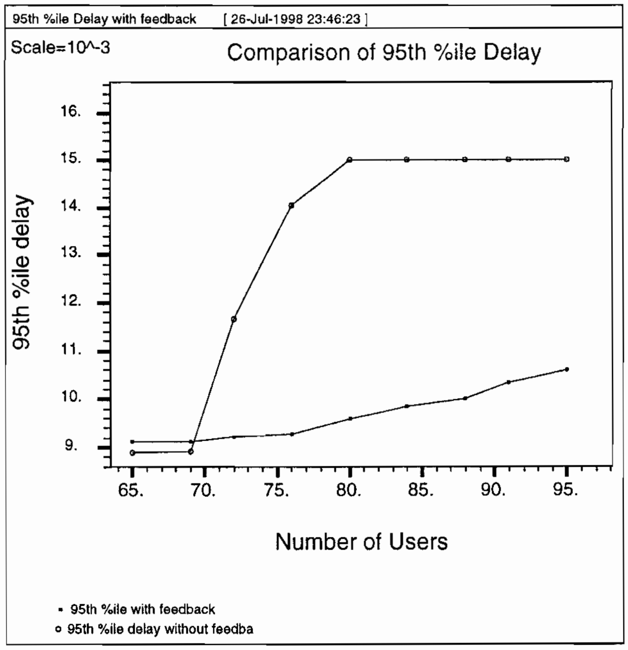
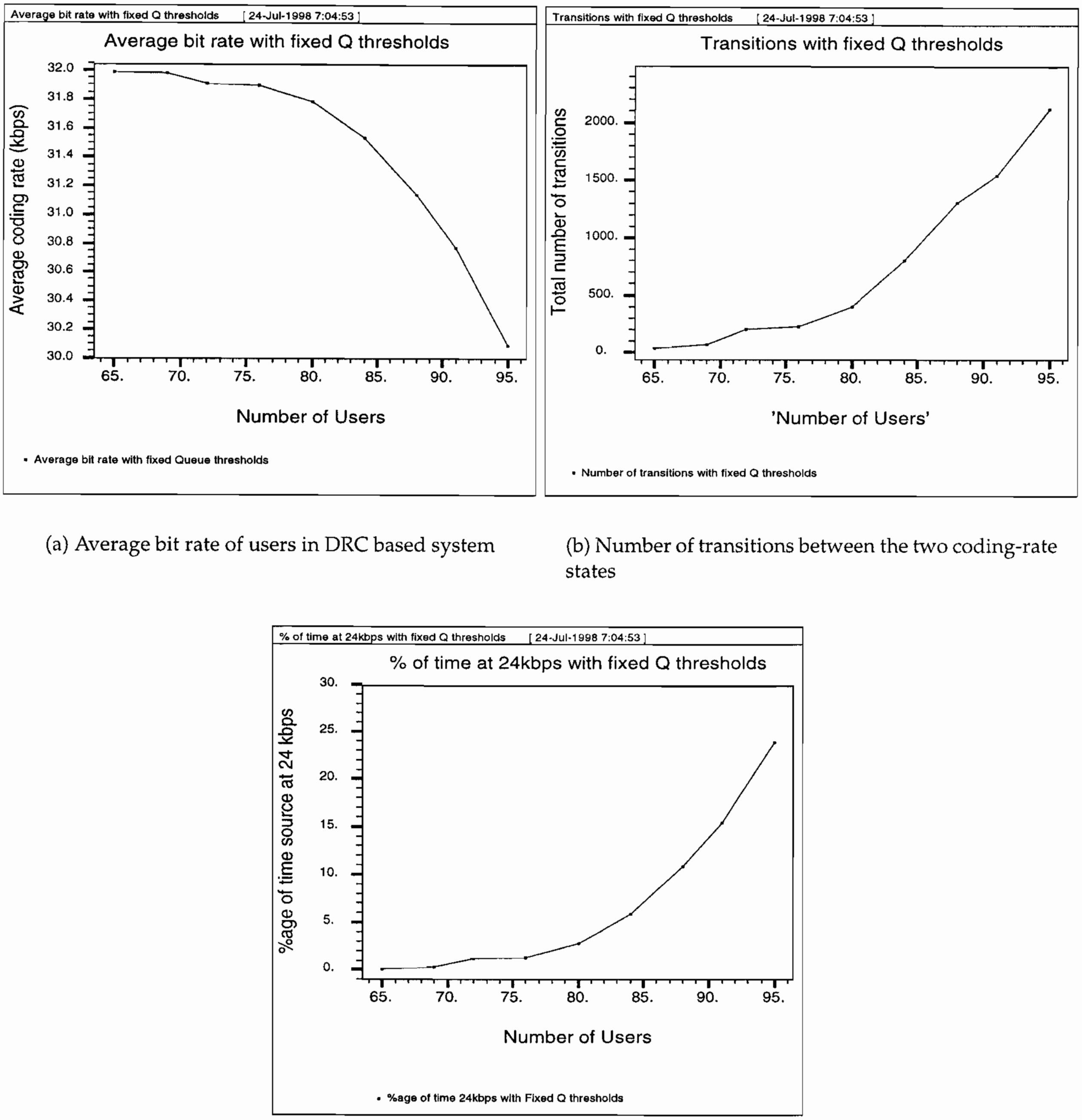


Figure 5.6: Comparison of $95^{th}\%ile$ delay for the packets of system with DRC and system without DRC

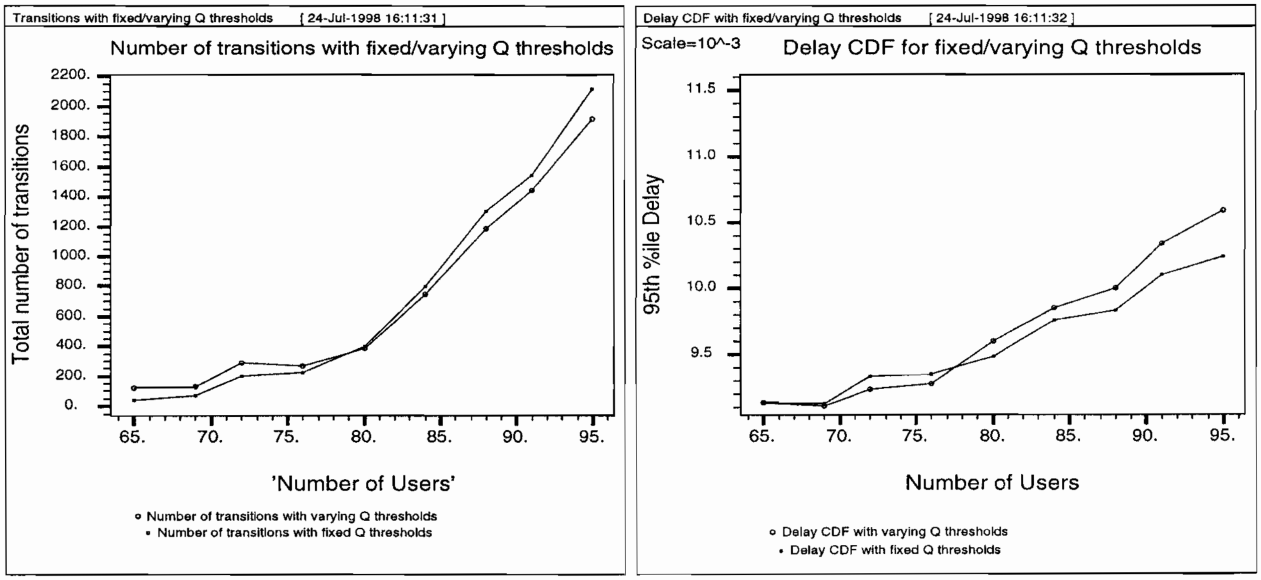


(a) Average bit rate of users in DRC based system

(b) Number of transitions between the two coding-rate states

(c) Percentage of time, coder stay in 24 kbps state. Greater the percentage, more is the QoS degradation

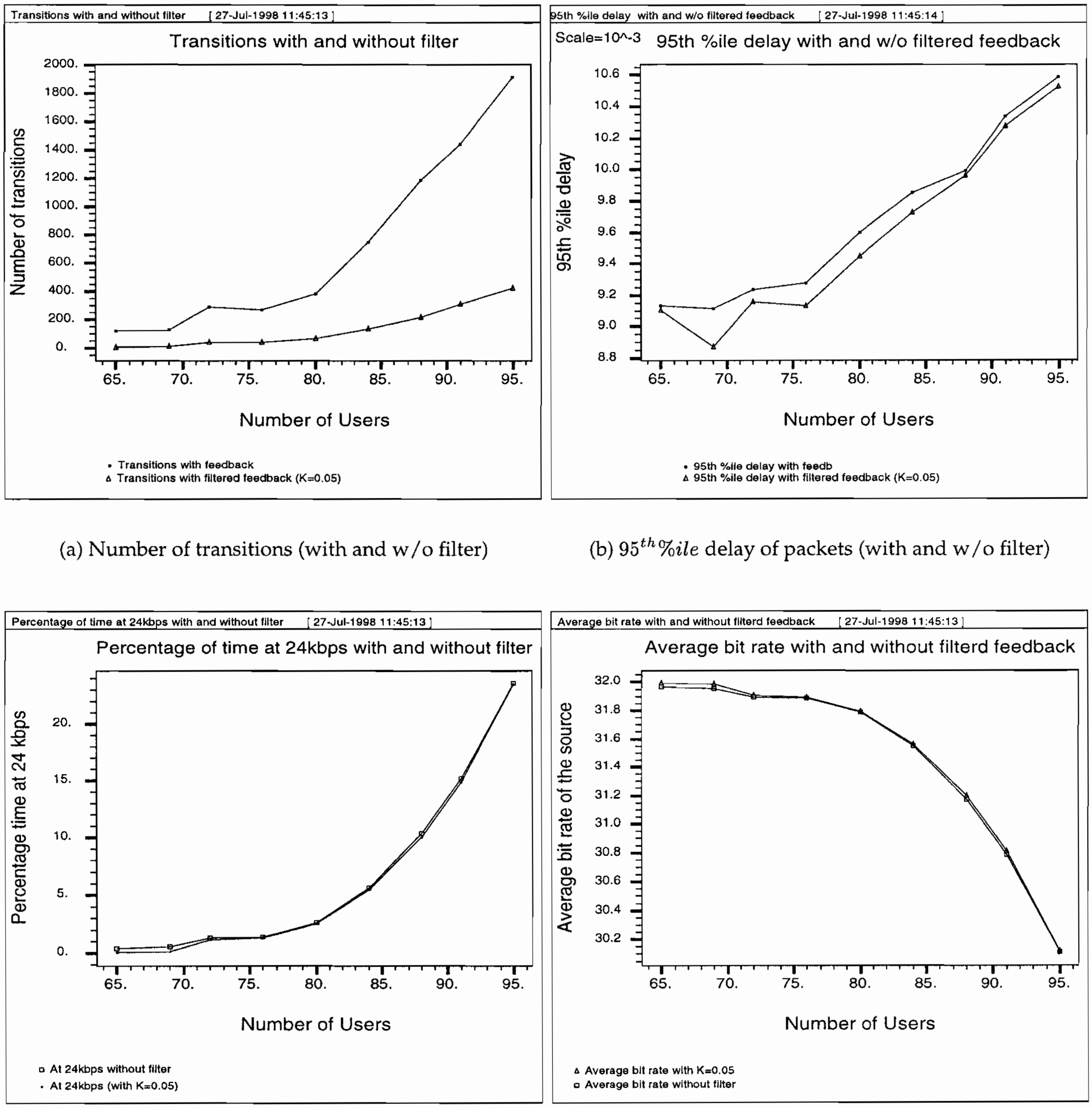
Figure 5.7: Performance evaluation of AAL2 system in the presence of DRC: With fixed Q_{min} and Q_{max} thresholds (independent of load).



(a) Comparison of total number of transitions in the system

(b) Comparison of $95^{th}\%ile$ delay for the packets of system

Figure 5.8: Comparison between performance of DRC based AAL2 system with fixed and varying queue thresholds (dependent of load).



(a) Number of transitions (with and w/o filter)

(b) $95^{th}\%ile$ delay of packets (with and w/o filter)

(c) Percentage of time the system is in 24 kbps state

(d) Average bit rate of users in system

Figure 5.9: Comparisons between performance of system in which feedback is based on filtered-queue information and a system without this filter.

Chapter 6

Conclusions and future work

AAL2 results in better bandwidth savings compared to other ATM adaptation layers. It is an ideal candidate to carry low-bit rate, delay sensitive applications like voice. In the work presented here, we have developed and investigated the static schemes for AAL2 CAC with the help of existing ATM CACs. It is shown that, in order to preserve fairness and avoid the complexity of interactions between AAL2 CAC and ATM CAC, the nature of VC on to which we are multiplexing users should be CBR. The selection of CBR VC might result in underutilization of link bandwidth. A dynamic rate control (DRC) mechanism is proposed to increase the bandwidth utilization without affecting the user QoS to a great extent. The bandwidth savings achieved by this mechanism is preferred to bandwidth savings by VBR VC because the QoS degradation in the former case is controllable and fair. With the selection of parameter values like queue thresholds and time-constants we can minimize the QoS degradation. The usage of AAL2 and DRC in AAL2 makes most sense when there are a large number of users ($\geq$ 25 voice users operating at 32 *kbps*). For small number of users, transmission of voice by AAL2 requires higher bandwidth than the voice bit rate. This is due to the overhead associated per packet. For example, in order to support two users with coding rates equal to 32 *kbps* with CPS packet size of 32 bytes, at least 106 *kbps* bandwidth is required. Due to the lack of any kind of statistical multiplexing and higher probability of transmitting partially filled cells in small number of user scenario, it seems prudent to use AAL1. The use of AAL1 will give the user better QoS guarantees and lesser network complexity to the service provider. It is found that even with the help of a dynamic coding rate control scheme we cannot achieve much advantage. The packet queue length in these cases is so low that minimum and maximum queue thresholds are nearly equal. This results in uncontrollable transitions or saturating in lower bit rate state for most of the time.

When a small number of users are multiplexed on a low capacity link, the required bandwidth (in AAL2) to satisfy the connection's QoS is highly dependent on the CU-timer value. Further studies are required to be carried out in order to find optimum value of CU-timer based on load and link conditions.

Further the studies show that the NEC's multiclass CAC is well suited for the AAL2 environment. Multi-class CAC, with the help of modified worst case on-off source, captures the burstiness effect of voice sources better than any other CAC schemes. Also, if the UPC parameters of the source are known, using multi-class CAC scheme, we can find mean on and off periods. This enables us in allocating the network resources for a wide variety of sources. When the number of users sharing a common link is high, AAL2 along with multi-class CAC and DRC scheme achieves better statistical multiplexing gain. Due to the co-location of AAL2 transmitter and the coder, the feedback delay is zero which results in immediate congestion control and better capability of avoiding momentary buffer overflow becoming sustained buffer overflow. In order to make DRC coupled with AAL2 most effective, we need to investigate following areas:

- Find the relation of CU-timer with network load and source and link parameters.
- Find a relation of queue thresholds with load on link, source parameters and QoS constraints.
- Optimum value of time constant in the recursive filter which can filter out all the high frequency fluctuations in every load scenario.
- An important issue in congestion control or network management is the measurement or identification of congestion. In the work presented here, we used queue length information as congestion level indicator. A different parallel study can be carried out by taking *number of sources on in a talk-spurt* as a measure of congestion.
- Investigating into different methods to improve the bandwidth requirement for small number of users while satisfying the QoS guarantees and maintaining low overall complexity of the network, could be a useful future work.
- The work done here involves *homogeneous* traffic environment. It would be interesting if studies can be done with *heterogeneous* traffic environment.
- Investigating into possibilities of constructing an analytical model which will help to make 'Effective bandwidth' calculation for above mentioned scheme in real-time, to be used for CAC purposes.

Bibliography

[1] Anick, D., et al., "Stochastic Theory of a Data-Handling System with Multiple Sources," *Bell System Tech. J.*, 61, 8 (Oct. 1982): 1871-1894.

[2] Guerin, R., et al., "Equivalent Capacity and Its Application to Bandwidth Allocation in High-Speed Networks," *IEEE JSAC*, 9, 7 (Sept. 1991): 968-981

[3] G. Ramamurthy and Qiang Ren, "Multi-class Connection Admission Control Policy for High Speed ATM Switches," *IEEE INFOCOM*, Vol.3, 963-972.

[4] Gelenbe, E., et al., "Diffusion Based Statistical Call Admission Control in ATM," *Perf. Eval.*, vol. 27, no.28, 1996, 411-436.

[5] H. G. Perros and K.M. Elsayed, "Call Admission Control Schemes: A review," *IEEE JSAC* November 1996, 82-90.

[6] Guerin, R., et al., "A Unified approach to bandwidth allocation and access control in fast packet-switched networks", *Proc. INFOCOM'*92 (1992) 1-12.

[7] Sohraby K., "On the Asymptotic Behavior of Heterogeneous Statistical Multiplexer with Applications," *Proc. INFOCOM'*92, 839-847.

[8] RaghuShankar Vatte and David W. Petr, "Performance Comparison between AAL1, AAL2 and AAL5," *Technical Report ITTC-FY98-TR-13110-03*, Department of EECS, University of Kansas, March 98.

[9] Harry Heffes and David M. Lucantoni, "A Markov modulated characterization of packetized voice and data traffic and related statistical multiplexer performance," *IEEE JSAC* Vol. SAC-4, No. 6, September 1986, 856-868.

[10] Schicker, P (editor), "B-ISDN ATM Adaptation Layer Type 2 Specification", ITU-T *Recommendation I.363.2*, Feb 1997.

[11] ITU-T, "Broadband aspects of ISDN ITU-T recommendation I.121," 1991

[12] S.-Q. Li, "A general solution technique for discrete queuing analysis of multimedia traffic on ATM," *IEEE transactions on communications*, vol. 39, 1991, pp. 1115-32.

[13] N. Yin and M. G. Hluchyj, "A dynamic rate control mechanism for source coded traffic in fast packet network," *IEEE Journal on Selected Areas of Communication*, 9(7): 1003-1012, Sept 1991.

[14] A. Elwalid and D. Mitra, "Effective bandwidth of general Markovian traffic sources and admission control of high speed networks," *IEEE/ACM Trans. Networking*, vol. 1, 1993, pp. 329-43.

[15] G. L. Choudhury, D. M. Lucantoni, and W. Whitt, "On the effective bandwidths for admission control in ATM networks," *Proc. 14th Int'l teletraffic congress (ITC '94)*, 1994, pp. 329-43.

[16] Mischa Schwartz, "Broadband Integrated Networks", Prentice Hall, UpperSaddle River, NJ 1996.

[17] K. Sriram, "Dynamic bandwidth allocation and congestion control schemes for voice and data multiplexing in wideband packet technology", *IEEE* International Conference on Communications (ICC), April 1990, pp. 1003-1009.

[18] Wing Cheong Lau, "Traffic Characterization, Quality of Service and system design in multimedia broadband networks", Ph.D Dissertation, The University of Texas at Austin, December 1995.

[19] R.Nagarajan, "Quality-of-Service issues in high speed networks," Ph.D thesis, University of Massachusets, Amherst, 1993.

[20] S. Q. Li, "Study of information loss in packet voice systems," *IEEE Trans. Commu.*, vol. 37, No. 12, Nov. 1989, pp. 1330-1339.

[21] Zbigniew Dziong, "ATM network resource management", New York: McGraw-Hill, 1997.

[22] G. Karlson and M. Vetterli, "Packet video and its integration in the network architecture," *IEEE JSAC*, vol. 7, No. 5, June, 1989, pp. 739-751.

[23] R.J. Gibbens, F.P. Kelly and P.B. Key, "A Decision-Theoretic approach to call admission control in ATM networks," *IEEE JSAC*, vol. 13, No. 6, Aug. 1995, pp. 1101-1114.

[24] Vishal Moondra, "Implementation and performance analysis of ATM adaptation layer type 2", M.S. thesis, University of Kansas, Lawrence, 1998.

[25] A. Periyannan, "A reactive congestion control scheme for gateways in high-speed networks", M.S. thesis, North Carolina State University, Raleigh, 1992.

[26] H. Gilbert, O. Aboul-Magd, and V. Phung, "Developing a cohesive traffic management strategy for ATM Networks" *IEEE Communications Magazine*, 29(10):36-45, Oct 1991.

[27] CCITT 1992c, Recommendation I.363, Broadband ISDN Adaptation Layer (AAL) Specifications, Geneva, 1992.

[28] R. Yelisetti, D. W. Petr, "Development of simulation models for AAL1 and AAL5", *Technical Report ITTC-FY98-TR-13110-02*, Department of EECS, University of Kansas, February '98.

[29] T. Bially, B. Gold, and S. Seneff, "A technique for adaptive voice flow control in integrated packet networks," *IEEE trans. Commun.*, vol. COM-28, no. 3, pp. 325-333, Mar. 1980.

[30] Comdisco Systems, Inc., *BONeS Designer Modeling Reference Guide*, June 1993.