Order Number 9510713

Study of congestion control schemes in multimedia traffic networks

Saleh, Mohsen A., Ph.D.

City University of New York, 1994

U·M·I
300 N. Zeeb Rd.
Ann Arbor, MI 48106

STUDY OF CONGESTION CONTROL SCHEMES IN MULTIMEDIA TRAFFIC NETWORKS

by

MOHSEN SALEH

A dissertation submitted to the Graduate Faculty in Engineering in partial fulfillment of the requirements for the degree of Doctor of Philosophy, The City University of New York.

1994

This manuscript has been read and accepted for the Graduate Faculty in Engineering in satisfaction of the dissertation requirement for the degree of Doctor of Philosophy.

9/22/94
Date

Chair of Examining Committee

9/22/94
Date

Executive Officer

Prof. Joseph Barba

Prof. Michael Conner

Prof. Myung Jong Lee

Prof. B. Kraimeche

Supervisory Committee

THE CITY UNIVERSITY OF NEW YORK

ABSTRACT

STUDY OF CONGESTION CONTROL SCHEMES IN MULTIMEDIA TRAFFIC NETWORKS

by

Mohsen Saleh

Advisor: Professor Tarek N. Saadawi

In our study, we are interested in the access flow control and the bandwidth control window which are applied at the cell level during the progress of the call. The flow control action, at the cell level, is implemented at the input access of the multiplexer. Using a suitable congestion measure, namely the multiplexer buffer length, the scheme dynamically controls the arrival rate by switching the coder to a different compression ratio (i. e.,Changing the coding rate). VBR coding methods can be adaptively adjusted to transmit at a lower rate with very little degradation in the service quality. Comparisons between statistical, deterministic and bandwidth control window allocations to voice and video traffic are presented. Our objective, here, is to control the input bursty traffic at heavy congestion states and to be able to traffic shape the input arrival process such that we can smooth down its characteristics and force it to be well behaved.

Our study analyzes the performance of an access node in ATM network with and without flow control using simulation methods, and includes the following, multiplexing of homogeneous types of traffic as voice or video sources, multiplexing of heterogeneous types of traffic as voice and video sources and bandwidth management for different classes of traffic using deterministic, statistical and bandwidth control window allocation to voice and video traffic.

The contribution of this study is to provide a performance evaluation of the access flow control algorithm which is applied at the access nodes to the integrated network, specifically at the input voice and video multiplexers. Also it provides a performance analysis of the bandwidth assignment to voice and video traffic. The results of the simulation study include the optimum values of the multiplexer buffer thresholds that satisfy an objective cell blocking probability of 10^{-4} for voice traffic and 10^{-7} for video traffic and the effects of the access control window (the minimum period of time at which the controller stays in a given state before it is being triggered to another action) on the controller triggering rate and mean bits per sample. Also the optimum values of the bandwidth control window are presented.

ACKNOWLEDGEMENT

I would like to take this opportunity to express my appreciation for all the members of my doctoral committee for their guidance and support during my graduate study. It is a great pleasure to acknowledge my deep appreciation to my thesis advisor: Professor Tarek Saadawi for his guidance, encouragement, and numerous suggestions and discussions. I would like to dedicate this thesis to my wife and my daughter, whose patience, great support and understanding during the course of this undertaken encouraged me to attain my goals.

Table of Contents

List of Tables

List of Figures

1. INTRODUCTION

1.1 Background

Recently, the asynchronous transfer mode (ATM) has been widely recognized as a promising solution for the implementation of broadband ISDN (B-ISDN) [1-3]. ATM is an attractive communication transport technique for carrying future broadband services with a broad spectrum of different traffic characteristics, such as peak and average bit rate, statistical distribution of cells interarrival time, etc. In ATM networks, all information such as voice, video and data is conveyed using a fixed-size block called cell and statistical multiplexing is employed. Since ATM involves a statistical multiplexing scheme, queueing delay, cell loss, or other degradation of quality of service (QOS) may easily occur in cases of network congestion. Bursty traffic, such as image data transfers, has a particularly serious impact on communication quality because of statistical load fluctuations. Statistically multiplexing bursty traffic is a very efficient technique to maximize the utilization of the link capacity. Also it exploits the variance of the source statistics to achieve an efficiency gain measured by allowing a number of sources whose peak rates are greater than the link capacity. Thus ATM networks have the flexibility to support various services and introduce new services easily, and are efficient due to the high utilization of the network resources [4]. The gained flexibility in accommodating traffic sources with wide range of bit rates may cause serious congestion problems that will lead to severe buffer overflow, cell loss and degradation in the required service quality. Therefore, bandwidth management and traffic flow control are required to meet the quality of service (QOS) for various service types and mixes (homogeneous and heterogeneous).

Quality features of multiplexed traffic with various bit rate and burstiness are being studied to estimate the necessary bandwidth allocated to each call. Various congestion control approaches have been proposed for ATM networks in order to ensure that the cell transport performance (cell loss and delay) remains at a satisfactory level for all the users. Many of the congestion control schemes developed for existing networks may not be applicable to ATM networks. Therefore, a new concept is required for congestion control in an ATM network. All proposed schemes fall in the class of preventive or reactive control. Preventive control tries to prevent congestion before it happens. The objective of preventive control is to ensure a priori that network traffic will not reach the level which causes unacceptable congestion. Reactive control reacts to the congestion after it happens and tries to bring the degree of network congestion to an acceptable level. However, reactive control is not suitable for a use in ATM networks [5]. In order to traffic engineer integrated voice/video networks, it is first necessary to understand the characteristics of the packet traffic generated by voice and video sources. In particular, it is of interest to know a way to characterize the superposition of packet arrival streams generated by a number of voice and/or video sources.

The most important element of a packet-switched speech network is the packet-speech multiplexer in which, packets arriving once the queue has reached a certain limit are discarded, or if embedded encoding has been used, shortened. All previous methods for the analysis of a packet-speech multiplexer have drawbacks. The simple one-dimensional models proposed by Minoli [6],[7], Coviello [8],[9] and Seguel et al. [10] are too inaccurate to be of use, and conclusions based upon them can be very misleading. These models fail because they assume packet arrivals are uncorrelated, whereas in fact they are highly correlated. Other models take these correlated arrivals into account and produce more accurate results [11],[12]. However, these models can only accurately predict the mean and variance of queueing delays, which are of no help in estimating packet loss. The most accurate models look beyond the packet-arrival process to the speech-generation, and therefore the packet-generation, process [13]-[17]. However, only one of these [13], [14] has been used to analyze a limited queue. Unfortunately, this model requires very large array sizes and large amounts of computation, and in practice, the technique is limited to low-capacity links with small queue limits. Recent studies have shown that the superposition of packet sequences generated by packetized voice sources with speech detection exhibit high burstiness (relative to a Poisson process) due to inherent correlations between successive interarrival times in the superposition stream. Using the indexes of dispersion for counts and intervals, Sriram and Whitt [18] and Heffes and Lucantoni

[19] have investigated the burstiness due to correlations. These correlations tend to cause significantly larger queueing delays and packet losses than would be predicted by a Poisson model.

When voice constitutes a significant portion of the network traffic, compression makes it possible to have significant additional capacity in the network. The methods of voice compression include adaptive differential pulse code modulation (ADPCM) algorithms and digital speech interpolation (DSI). Many new low bit-rate voice coding techniques are being developed and submitted for standardization(e.g., 16 kb/s low delay-code excited linear predictor) which can be potentially considered for use in B-ISDN in the future. Voice traffic becomes statistical and is bursty when it is multiplexed using DSI and other methods for compression. Therefore, appropriate voice congestion control methods are needed. In conventional data packet switching networks, flow control is reserved for the network layer functions. In ATM, flow control is reserved for all control functions at the network, the call and the cell levels. Some of flow control methods employ block dropping techniques to decrease the voice packet size and thereby increase the queue service rates during congestion periods [20],[21]. A variation of these techniques, called cell discarding (CD), may be used in ATM networks. In the CD scheme, voice cells are categorized as high-priority (more significant) and low-priority (less significant) information, so that the low-priority cells may be discarded during congestion [22],[23].

In addition to voice and data, video is becoming an increasingly important communication medium. In video communication, a variable rate video transmission is employed which keeps the picture quality constant by transmitting variable bit rates according to the amount of information generated in a picture. Coding algorithms and implementation technologies will evolve further in the future, so more efficient codecs will be applicable. Variable-bit-rate compression algorithms transmit at a higher rate during high-activity (motion) scenes and at a low rate when there is less motion. It is possible to multiplex statistically several independent video sources at a speed lower than the aggregate peak coding rate. The law of large numbers indicates that as the number of independent sources increases, the aggregate rate approaches the average, without adjustment of individual source rates by varying the picture quality.

Two different bit rate allocation policies seem applicable to ATM networks which are the average and the peak rate allocation. The basis of the peak rate allocation is that the sums of maximum rates of all sources should not exceed the link capacity. Then congestion due to traffic overload is very unlikely. Peak rate bandwidth allocation (deterministic) is not efficient. The usage of the peak rate criterion for call acceptance lowers the utilization factor of the link to the one of a simple time sharing scheme met in a circuit switching network. This eliminates the statistical advantage of the Asynchronous Transfer Mode. The basis of the average rate allocation (statistical allocation) is that the sums of average rates of all sources should not exceed the link capacity. Obviously a call, which would violate this criterion, should not be accepted.

If however some sources are bursty, long delays may occur or cell loss rate will increase. A combination of peak and average rate allocation might be used. However the exact mix is very difficult to define. It suggested that bandwidth for constant bit rate (CBR) services should be allocated according to their peak rate, whereas variable bit rate (VBR) services might be given less than peak bandwidth. In [24] an effective bandwidth assignment is presented, which assigns a value of bandwidth between the peak and the average bit rate of the source. Effective bandwidth allocation could be used for efficient bandwidth management, thus leading to a more effective utilization of the link. The effective bandwidth depends mainly on the burstiness of the source, the burst length and the size of buffer at multiplexer. In [25] a bandwidth allocation scheme called the (T1,T2) is presented. It does not consider the issue of the optimal bandwidth allocation to each type of traffic, hence the results could not be used for ATM networks. A simulation study of a bandwidth allocation scheme called Class Related Rule (CRR) is presented in [26]. The study does not consider the characteristics of real time voice and video traffic, thus the results reported there could not be applied to the real ATM traffic case.

1.2 Contributions

In our study, we are interested in the access flow control and the bandwidth control window which are applied at the cell level during the progress of the call. The flow control action, at the cell level, is implemented at the input access of the multiplexer.

Using a suitable congestion measure, namely the multiplexer buffer length, the scheme dynamically controls the arrival rate by switching the coder to a different compression ratio (i. e.,Changing the coding rate). VBR coding methods can be adaptively adjusted to transmit at a lower rate with very little degradation in the service quality. The access scheme controls the input bursty traffic at heavy congestion states and it is able to traffic shape the input arrival process such that we can smooth down its characteristics and force it to be well behaved. At the cell level, the statistical behavior of the bursty traffic, even after the mixing of many sources, is far from Poisson, making the analysis very difficult [27]. Hence, our analysis is carried out by simulation. Also simulation is used to observe the history of some processes over a period of time in addition to estimating certain parameters. We define an access control window which is the minimum period of time at which the controller stays in a given state before it is being triggered to another action. Also we define a bandwidth control window to control the scheduling of different types of traffic such as voice and video traffic. Voice and video traffic are being multiplexed into the outgoing link. Each type of traffic is supported on a separate buffer and then multiplexed on a separate virtual path. Using proper values of the bandwidth control window, a path bandwidth to each traffic can be allocated such that it guarantee the traffic required QOS in terms of its cell loss rate.

Our study analyzes the performance of an access node in ATM network with and without flow control using simulation methods, and includes the following:

1. Multiplexing of homogeneous types of traffic as voice or video sources.

2. Multiplexing of heterogeneous types of traffic as voice and video sources.

3. Bandwidth management for different classes of traffic using deterministic, statistical and bandwidth control window allocation to voice and video traffic.

The contribution of this study is to provide a performance evaluation of the access flow control algorithm after including the access control window. Also it provides a performance analysis of the bandwidth assignment to voice and video traffic. The results of the simulation study include the optimum values of the multiplexer buffer thresholds that satisfy an objective cell blocking probability of 10^{-4} for voice traffic and 10^{-7} for video traffic and the effects of the access control window on the controller triggering rate and mean bits per sample. Comparisons between statistical, deterministic and bandwidth control window allocations to voice and video traffic are presented. Also the optimum values of the bandwidth control window are presented.

1.3 ATM Transport Networks

ATM is a cell based transfer mode using asynchronous time division multiplexing technique. An ATM cell stream consists of a continuous sequence of cells, some of which are designated empty and others of which have been dynamically allocated to variable or fixed rate sources. The principle characteristics of asynchronous transfer mode (ATM) networks are that they can absorb bit rate variations so that links are suited to cell traffic generated by variable rate sources and that they can offer high efficiencies because numerous variable bit rate sources can share network resources. In an ATM network, the user information is carried in cells which are fixed-size blocks and consist of a header and an information field. The header contains a label which uniquely identifies a channel and is used for multiplexing and routing. The information field is transported transparently by the ATM layer. Since the interface structure contains a set of labeled channels (logically bound rather than physically positioned), an ATM can dynamically adapt to the changes in the service demand and allocate the resources such as channel bandwidth accordingly [28].

An ATM network model based on the concept called "Virtual Path" (VP) has been proposed, and described in [29],[30]. In this model two nodes are connected by a logical link called virtual path. A virtual path consists of a number of virtual circuits and has its own bandwidth which gives the upper limit of that of the virtual circuits enclosed. Virtual paths are multiplexed in a physical link on a cell multiplexing basis. They are established or released dynamically based on: 1) long term service provision; 2) demand for alternate routing in case of network failure; and 3) short term demand. The capacity of the virtual paths can be dynamically allocated in the same fashion.

The cell header contains two subfields which are the virtual path identifier (VPI) and the virtual channel identifier (VCI). The separation of a VPI from a VCI eliminates call-by-call basic processing at each transit node during call establishment. During the call set-up process, the access node identifies the appropriate path, or creates one if necessary, and makes the decision whether to accept or reject the call based on the current path bandwidth utilization conditions. The cell header functions are shown in Figure 1.1. The Header size is 5 bytes long, both at the User Network Interface (UNI) and Network Node Interface (NNI). The following fields are identified across the UNI, [31],[32],[33],[34]:

1) Generic Flow Control: This is a 4 bits field which is defined at the UNI to assist the users in the flow control of their traffic according to a certain class of service. It specifies the medium access control functions, and across the UNI this field is replaced with a label field.

2) Virtual Channel Identifier: This field is 12 to 16 bits at the UNI, and 16 bits at the NNI. This field identifies a particular end to end switched connection. It relates to the switching functions of cells belonging to a certain logical connection. The value of the VCI may change as the cell traverses the network.

3) Virtual Path Identifier: This field is 8 to 12 bits at the UNI and 12 bits at the NNI. It consists of a bundle of virtual channels that are carried on the same physical media, from one end to the other. It relates to the cross-connection functions of the cells. It emulates the functions of the trunk concept in circuit switching.

4) Payload Type: It is a 2 bits field. It is used to distinguish network information from user information. In network information cells, it provides in-band control message. In user information cells, it provides service adaptation functions. For example, it can identify low priority cells or cells that have violated a certain traffic characteristics.

5) Header Error Check: One byte field used for error detection and correction on the header. It is important to perform this function on the header in order to avoid misdelivery of cells.

The functions of the header can be summarized in the following:

1) To identify the characteristics of each virtual channel, which in turn provides the basis for categorizing virtual channels (with similar traffic characteristics and class of service) to be cross-connected as a single entity (virtual path).

2) To provide the network management functions, such as bandwidth assignment and distributed control, with a simple tool to employ bandwidth control and enforcement, which is accomplished via the virtual path concept. The VPI concept plays an important role in bandwidth allocation.

3) To implement real time physical connection control without the excessive overhead, which is a crucial design point in the ATM transport principle.

1.4 Real-Time Traffic Source Models

Recently, there has been interest in supporting real-time communication applications such as interactive voice and video applications in an ATM networks. Such real-time traffic differs from traditional data traffic in several ways. Most importantly, real-time traffic is delay sensitive and less sensitive in loss while data traffic is loss sensitive and less sensitive in delay. Hence, it is natural to engineer communication networks that support real-time traffic, so that delays are bounded at the expense of some loss. However, the magnitude of this loss determines the quality of service and, hence, it is critical to predict this loss accurately in order to provide an acceptable grade of service. In ATM network, most traffic sources are bursty such as voice and video sources. A bursty source may generate cells at a near-peak rate for a very short period of time, and immediately afterwards, it may become inactive, generating no cells. Burstiness is a parameter to describe how densely or sparsely cell arrival occur. Some of the definitions of burstiness include: (1), it is the ratio of peak bit rate to average bit rate [35],[36],

(2), it is the average burst length (i. e., The mean duration of time during which the traffic source transmits at the peak rate) [26]. As the burst length gets longer, the worse the network performance becomes, namely the cell loss probability becomes larger. With longer bursts, statistical multiplexing becomes less effective, and thus, fewer active sources can be supported for a given amount of bandwidth. The burst length is important when it is comparable to the size of the multiplexer buffer. However, in most real-time application, the buffer length will be in the order of several microseconds and the burst length will be greater than the size of the buffer. If the input source peak rate is decreased, then it will be possible to deliver the required class of service without depleting the network resources. At the same time, the statistical multiplexing will be more effective and the traffic characteristics get smoother.

The traffic generated by a single user for a single service can be either Constant Bit Rate (CBR) or Variable Bit Rate (VBR). A CBR results in an ATM cell stream which is associated with a particular Virtual Channel Identifier (VCI) and has a predictable and uniform cell rate, while the cell stream resulting from a VBR service generally has a nonuniform cell rate. A more accurate characterization of VBR traffic can be given as follows: the interarrival time between successive cell bursts is denoted by the random variable I, the cell burst length is denoted by the random variable B, and the time interval between successive cells during a burst is denoted by D. In general, the probability distributions of I and B depend strongly on the particular application being modeled. A single connection has a variable bit rate bounded by the maximum bit rate of its physical attachment. In order to characterize the effective bit rate, we need to select an appropriate model to specify its characteristics in terms of known parameters or metrics. The two-state model captures the basic behavior of the source associated with a connection. The rationale for such a model is that a source is either in an "idle state", transmitting at zero bit rate, or in a "burst state" and transmitting at its peak bit rate. Such a source model has the advantage of being both simple and flexible as it can be used to either represent connections ranging from bursty to continuous bit streams or approximate more complex sources. Based on this two-state model, the idle and burst periods are defined as the times during which the source is idle or active, respectively. The peak rate of a connection R_{peak} and distribution of idle and burst periods completely identify the traffic statistics of a connection. Assuming the parameters of a connection are stationary, its peak rate R_{peak} and utilization ρ, i.e., fraction of time the source is active, completely identify other quantities of interest such as mean and variance of the bit rate. For exponentially distributed burst and idle periods, the source is furthermore completely characterized by only three parameters, namely R_{peak}, ρ, and b, where b is the mean of the burst period. The mean burst period b gives some indications on how data is being generated by the source. Two sources, with identical mean and peak bit rates but different burst periods, have different impacts on the network [37]. It is our belief that the connection metric vector (R_{peak},ρ,b) represents the most significant aspects of a source behavior. This model can be extended to handle sources with nonexponential burst and idle periods, through the use of simple approximation techniques.

1.4.1 Input Traffic Models for Voice Sources

The arrival process of new voice calls and the distribution of their durations can be characterized by a Poisson process and an exponential distribution, respectively. Within a call, talkspurts and silent periods alternate . During talkspurts, voice cells are generated periodically; during silent periods, no cells are generated. The correlated generation of voice cells within a call can be modeled by an Interrupted Poisson Process (IPP) [17],[38],[39],[40],[41].In an IPP model, each voice source is characterized by ON (corresponding to talkspurt) and OFF (corresponding to silence duration) periods, which appear in turn. As seen in Figure 1.2, the transition from On to OFF occurs with the probability β, and the transition from OFF to ON occurs with the probability

α. ON and OFF periods are exponentially distributed with the mean $1/\beta$ and $1/\alpha$, respectively. Cells are generated during the ON period according to a Bernoulli distribution with the rate λ, and no cell is generated during the OFF period.

The input process of superposition of N voice sources to the statistical multiplexer is a fairly complex process and can possess correlations, in the number of arrivals in adjacent time intervals, which can significantly affect queueing performance. These correlations result from the fact that the aggregate voice arrival rate is a modulated process obtained by modulating the individual voice source rate by the number of voice sources in their talk spurt, which is itself a correlated process. Even if a component voice process is approximated as a renewal process, with deterministically spaced cells during a talk spurt followed by an exponentially distributed silence period, the superposition process is a complex nonrenewal process. It turns out that the aggregate cell arrival process resulting from the superposition of the streams from the voice sources is quite complicated, possessing a burstiness (high variability) that leads to a large cell delays in the multiplexer under heavy load. Exact analysis of systems to which this superposition process is offered is intractable, especially when such systems contain finite buffers and overload control mechanisms. When N independent voice sources are multiplexed, aggregated cell arrivals are governed by the number of voice sources in the ON state. Assuming a discrete time system, the probability P_n that n out of N voice sources are in the ON state (n voice cell arrivals in a slot) is given by

$$P_n = \binom{N}{n}\left(\frac{\alpha}{\alpha+\beta}\right)^n\left(\frac{\beta}{\alpha+\beta}\right)^{N-n} \quad for\ 0 \le n \le N.$$

The continuous time analogy represents the number of voice sources in the ON state as a birth-death process with birth rate $\lambda(n)$ and death rate $\mu(n)$, as seen in Figure 1.3, where

$$\lambda(n) = (N-n)\alpha,\ \mu(n) = n\beta\ for\ 0 \le n \le N$$

For this continuous time case, the probability P_n that n out of N voice sources are in the ON state is also given by the above equation [40].

Another common approach to model aggregate arrivals from N voice sources is to use a two-state Markov Modulated Poisson Process (MMPP) [13],[42]. The MMPP is a doubly stochastic Poisson process where the rate process is determined by the state of a continuous-time Markov chain [13]. In the two-state MMPP model, an aggregate arrival process is characterized by two alternating states. It is usually assumed that the duration of each state follows a geometrical (discrete time case) or an exponential (continuous time case) distribution, and cell arrivals in each state follow a Bernoulli (or a Poisson) distribution with different rates. Therefore, four parameters are necessary to describe an MMPP: the mean duration of each state and the arrival rate in each state. Note that an IPP, a process used to describe a single voice source, is a special case of the MMPP in which no cell arrives during an OFF period.

To determine the values of these four parameters, the following MMPP statistical characteristics are matched with the measured data [13]:

(1) the mean arrival rate,

(2) the variance-to-mean ratio of the number of arrivals in a time interval $(0,t_1)$,

(3) the long term variance-to-mean ratio of the number of arrivals, and

(4) the third moment of the number of arrivals in $(0,t_2)$

1.4.2 Input Traffic Models for Video Sources

Video traffic requires large bandwidth. For instance, in TV applications, a frame of 512 x 512 resolution is transmitted every 1/30 second, generating 512 x 512 x 8 x 30 bits per second (approximately 63 Mbits/s), if a simple PCM coding scheme is used. Therefore, video sources are usually compressed by using an interframe variable-rate coding scheme, which encodes only significant differences between successive frames. This introduces strong correlation among cell arrivals from successive frames.

Like a voice source, a video source generates correlated cell arrivals. However, its statistical nature is quite different from a voice source. Two types of correlations are evident in the cell generation process of a video source: short-term correlation and long-term correlation. Short-term correlation corresponds to uniform activity levels (i. e , small fluctuations in bit rates), and its effects last for a very short period of time (on the order of a few hundred milliseconds). Long-term correlation corresponds to sudden scene changes, which cause a large rate of arrivals, and its effects last relatively long (on the order of a few seconds) [43]. In subsection 1.4.2.1, models which consider only short-term correlation (i. e., Models for video sources without scene changes) are examined. In subsection 1.4.2.2, models which consider both short-term and long-term correlation (i. e., Models for video sources with scene changes) are examined.

1.4.2.1 Models for Video Sources Without Scene Changes

These models are applicable to video scenes with relatively uniform activity levels such as videotelephone scenes showing a person talking. Two models have been proposed. The first model approximates a video source by an autoregressive (AR) process [44],[45]. This model describes the cell generation process of a video source quite accurately. However, because of its complexity, queueing analysis based on this model is very complicated and may not be tractable. This model is more suitable for use in simulations. The second model approximates a video source by a discrete-state Markov model [48]. This model is more tractable in queueing analysis than the first model, and yet describes the cell generation process of a video source well.

Model (1): Continuous-State Autoregressive Markov Model

The definition of an AR process is as follows:

$$\lambda(n) = \sum_{m=1}^{M} a_m(n-m) + bw(n)$$

where $\lambda(n)$ represents the source bit rate during the nth frame; M is the model order; w(n) is a Gaussian random process; and a_m (m = 1,2,...,M) and b are coefficients. It is shown that the first-order autoregressive Markov model,

$$\lambda(n) = a_1 \lambda(n-1) + bw(n)$$

is sufficient for engineering purposes. Assuming that w(n) has the mean η and the variance 1, and that a_1 is less than 1, the values of coefficients a_1 and b are determined by matching the steady-state average $E(\lambda)$ and discrete autocovariance C(n) of the AR process with the measured data. $E(\lambda)$ and C(n) of the AR process in the previous equation are given by [46],

$$E(\lambda) = \frac{b}{1-a_1}\eta, \quad C(n) = \frac{b^2}{1-a_1^2} a_1^n \quad n \geq 0$$

This model provides a rather accurate approximation of the bit rate of a single video source without scene changes. However, analysis of a queueing model with the above arrival process can be very complex and may not be tractable. Therefore, this model is suitable for use in simulations.

Model (2): Discrete-State, Continuous-Time Markov Process

The process $\lambda(t)$ describing the bit rate of a video source at time t is a continuous-time, continuous-state process. In this model, process $\lambda(t)$ is sampled at random Poisson time instances and the states are quantized at these points as seen in Figure 1.4. The process $\lambda(t)$ is approximated by a continuous-time process $\overline{\lambda}(t)$ with discrete jumps at random Poisson times. This approximation can be improved by decreasing the quantization step A and increasing the sampling rate.

The state transition diagram for $\overline{\lambda}(t)$ is shown in Figure 1.5. This process $\overline{\lambda}(t)$ can be used to describe a single source as well as an aggregation of several sources. The aggregated arrival process from N video sources can transit between M+1 levels. The label in each state indicates the data rate in that state. (A is a constant). To determine values of the quantization step A and the transition rates α and β, the steady-state mean $E(\overline{\lambda}_N)$, variance $\overline{C}_N(0)$, and autocovariance function $\overline{C}_N(\tau)$ of the process $\overline{\lambda}(t)$ (describing an aggregate of N independent sources) are matched with the measured data. (τ is a time parameter.) $E(\overline{\lambda}_N)$, $\overline{C}_N(0)$ and $\overline{C}_N$ are given by

$$E(\overline{\lambda}_N) = MA\frac{\alpha}{\alpha+\beta}, \overline{C}_N(0) = MA^2\frac{\alpha}{\alpha+\beta}\left(1 - \frac{\alpha}{\alpha+\beta}\right), \overline{C}_N(\tau) = \overline{C}_N(0)\exp-(\alpha+\beta)\tau$$

The number of quantization levels M is chosen arbitrarily, but it should be large enough to cover all likely bit rates.

The process in Figure 1.5 can be decomposed into a superposition of simpler processes. It can be thought of as a superposition of M independent identical ON-OFF minisources, each being modeled as in Figure 1.6. Each minisource alternates between ON and OFF states. Transition from ON to OFF state occurs with the rate β, and transition from OFF to ON state occurs with rate α. (Thus, both ON and OFF periods are exponentially distributed). The data rate of a minisource in ON state is A. A minisource does not generate bits during OFF state (data rate is 0). (Note that, in Figure 1.5, a label associated with the state represents the data rate of a minisource in that state). The state of the aggregated arrival process can thus be represented by the number of minisources which are in ON state.

1.4.2.2 Models for Video Sources With Scene Changes

These models capture both short-term and long-term correlations. Thus, they are suitable to describe a cell generation process from video scenes with sudden changes such as videotelephone scenes showing changes between listener and talker modes or scene changes in broadcast TV [43]. Two models have been proposed . The first model is an extension of Model (2) explained above, and the second model approximates a video source by the discrete-state continuous-time Markov process ,Model (2), with batch arrivals.

Model (3) An Extension of Model (2)

The state transition diagram of the cell generation process from an aggregation of N video sources is shown in Figure 1.7. This process can also be used to describe a single video source with scene changes. The label in each state indicates the data rate in that state. There are two basic data rate levels: a high data rate A_h, which represents a sudden scene change, and a low data rate A_l, which represents a uniform activity level. If scene changes do not exist, the process in Figure 1.7 reduces to the one used in Model (2). The aggregated process of N video sources can transit between $(M_1+1)(M_2+1)$ levels, where $M_1 = NM$, $M_2 = N$. Here, M is chosen arbitrarily.

To determine the values of system parameters c and d (the transition probabilities between uniform activity level and high activity level), the fraction of the time spent in the high activity level $\left(\frac{c}{c+d}\right)$ and the average time spent in the high activity level (1/d) are equated with the actual measured data. To determine the rest of the parameters in the model. i.e.,the transition probabilities within the uniform activity level (a and b), and the two basic data rates (A_l and A_h), the first and second order statistics are matched with the actual measured data.

As in Model (2), the process described in Figure 1.7 can be decomposed into a superposition of simpler processes. This process can be thought of as a superposition of M_1 independent identical ON-OFF minisources of the type shown in Figure 1.8(a) and M_2 of the type shown in Figure 1.8(b). The state of the aggregated arrival process can thus be described as the number of each type of minisource which is in the ON state.

Model (4) Discrete-State Continuous-Time Markov Process with Batch Arrivals [47],[48]

The uniform activity level is represented by a discrete-state continuous-time Markov process as in Model (2). This M-state Markov process can be decomposed into M independent identical ON-OFF minisources. Scene changes (high activity levels) are represented by a batch arrival process. The interarrival times between scene changes (between batches) are assumed to be exponentially distributed. The batch size is assumed to be constant.

1.5 Statistical Multiplexing of Bursty Sources in ATM Networks

The effects of statistical multiplexing of bursty sources in an ATM network are investigated in [49, 50, 51, 52]. They investigate how the performance (the cell loss probability and the average delay time) varies as a function of various parameters such as the number of sources, the peak bit rate and burstiness of the sources. Some of the common observations in these papers are:

1) The average burst length is a very important parameter. As the average burst length increases, the performance degrades, i.e., the cell loss probability and delay time increase significantly [49, 50, 51, 52].

2) As the peak rate of each source is increased, the loss probability increases [50, 51]. This should be intuitively clear.

3) In the case where homogeneous sources are multiplexed, if the offered load (i.e., the number of sources x mean bit rate of each source) is kept constant, the cell loss probability decreases as the number of sources multiplexed increases. The reason is that when the number of sources multiplexed increases (keeping the offered load constant), the mean bit rate of each source decreases. The mean bit rate is a product of peak bit rate and the fraction of time in which a source is in the active-state (i.e., the state in which a source is transmitting at the peak rate). Therefore, the reduction in the mean bit rate means the reduction in either the peak bit rate or the burst length (or both). In either case, the cell loss probability decreases [49, 51].

4) In the case where heterogeneous sources are multiplexed, high-bit-rate sources dominate the performance; increase in high-bit-rate traffic causes more significant increases in the cell loss probability than an increase in low-bit-rate traffic does [52]. Similar observation is made in the case when homogeneous sources are multiplexed; when high-bit-rate sources are multiplexed, the fluctuation in the cell loss is larger than it is when low-bit-rate sources are multiplexed [38]. This is due to the fact that, because

of the high-bit rate, the number of traffic sources which can be multiplexed on one link is rather limited and not large enough to smooth out the bursty nature of each call.
5) The cell loss probability decreases as the offered load decreases [49,52]. Therefore, a very efficient way to lower the cell loss probability is to decrease the offered load by providing larger bandwidth. However, this is only possible under the assumption that bandwidth is negligibly cheap.

1.6 Congestion Control and Policing Function in ATM Networks

Recent advances in fiber optics and high-speed network technology have opened the possibility of providing integrated switching and transport for a variety of services in single integrated broadband network. Integration of such services poses serious challenges because of their widely different traffic characteristics and performance or quality of service (QOS) requirements. New congestion control strategies are now necessary to make such integration possible. In existing low-speed networks, data sources have traditionally been statistically multiplexed by taking advantage of their low activity factor (ratio of active to idle period) and short burst lengths. Any overload resulting from statistical multiplexing could be successfully smoothed out with a moderate amount of buffering at each node. During severe overload, link-by-link flow control mechanisms could easily throttle the sources. However, integration of high-speed data sources with long burst lengths relative to the time constants of the network pose a serious problem. Since such sources are likely to have a low activity factor, a large number of sources must be multiplexed together for efficient use of the network resources. If several such sources become active simultaneously, their combined arrival rate can exceed the capacity of the link, leading to blocking and consequent loss of cells, unless there are adequate control mechanisms to regulate these sources effectively [53].

1.6.1 Reactive Congestion Control Schemes

When congestion occurs at the network, reactive control instructs the source nodes to throttle their traffic flow by giving feedback to them. A major problem with reactive control in high-speed networks is slow feedback. Therefore, by the time that feedback reaches the source nodes and the control is triggered, it may be too late to react effectively.

There is a possible improvement technique to overcome the difficulty caused by slow feedback. If reactive control is performed between network users and the edge of the network as in [54], the effect of propagation delay may not be significant since the distance feedback information propagates is short. However, this limits the reactive control to the edge of the network. In general, the reactive control in an ATM environment is not effective.

1.6.2 Preventive Congestion Control Schemes

Preventive control tries to prevent the network from reaching an unacceptable level of congestion. The most common and effective approach is to control traffic flow at entry points to the network (i.e., at the access nodes). This approach is especially effective in ATM networks because of its connection-oriented transport. With connection-oriented transport, a decision to admit new traffic can be made based on knowledge of the state of the route which the traffic would follow [55].

Preventive control for ATM can be performed in two ways: admission control and bandwidth enforcement (policing). The admission control determines whether to accept or reject a new connection at the time of call set-up. This decision is based on traffic characteristics of the new connection and the current network load. The bandwidth enforcement monitors individual connections to insure that the actual traffic flow conforms with that reported at call establishment. Bandwidth enforcement schemes may be used with traffic shaping (smoothing). The purpose of traffic shaping is to throttle cell inputs into a network in order to avoid burst cell transmissions. Burst cell transmissions are avoided, for example, by separating successive ATM cells by an idle time. The shaping function could either be performed by the access control at a user-network interface or at a data source by buffering and injecting cells into a network at a slower speed. Since traffic shaping reduces network congestion by suppressing inputs to the network, it may be able to support a greater number of calls than a network without one. With traffic shaping, the service quality is degraded in a graceful way. However, entire transmission of traffic may be unnecessarily slowed down since cells are injected into a network at a slower speed even when the network load is light.

Holtzman has proposed a new and very different approach to preventive control [56] and applied his approach to admission control. In admission control, the decision to accept a new connection is made based on the predicted network performance. If there is some uncertainty in the parameter values of the incoming traffic, the network may underestimate the impact of accepting a new call, and congestion may result. Holtzman's approach tries to prevent the network congestion by taking uncertainties in traffic parameter values into account.

1.6.3 Policing Function

In cases where statistical cell multiplexing characteristics are evaluated based on parameters declared by the user, networks must make sure cells from the source are transmitted according to the declared parameters. What is more, if the declared parameters are violated, the network must restrict cell transmission in order to assure cell transfer quality for the other sources. A traffic shaper/enforcer monitors (or polices) each virtual connection to ensure that its traffic flow into the network conforms to the traffic descriptor (including parameters such as average and peak transmission rates, and maximum burst lengths), which could be specified at call setup. If the user's traffic does not conform to the traffic descriptor, some action has to be taken against the violating traffic. Generally, the accuracy of evaluating cell multiplexing characteristics increases as the number and types of parameters used in evaluation increases. Such proliferation of parameters, however, not only complicates the implementation of a policing function, it also imposes difficulties on parameter declaration. To be cost-effective, the policing function must be simple because it will be applied to every virtual circuit. To minimize the number of parameters that must be declared, it should be clarified what parameters are needed to effectively estimate the multiplexing delay. Moreover, as already mentioned, the parameters should be easy to declare. In summary, the parameters should satisfy the following minimal set of requirements. They should:

1) be sufficient to estimate the statistical characteristics, such as moments of cell arrival process, that are necessary to evaluate the multiplexing delay;

2) be easy to declare;

3) permit easy monitoring and restriction as necessary.

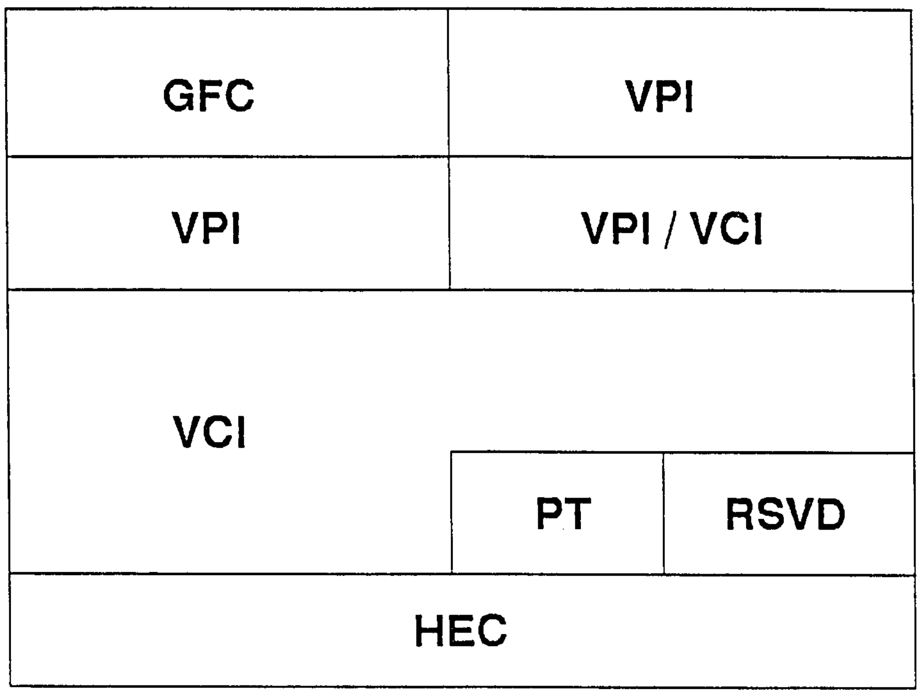


Figure 1.1 ATM Cell Header

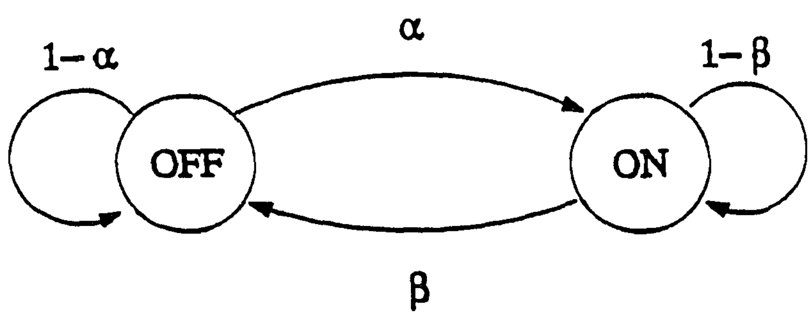


Figure 1.2 IPP Model

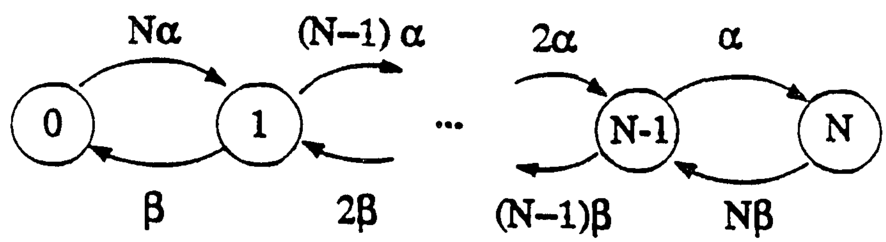


Figure 1.3 Birth-Death Model for the Number of Active Voice Sources

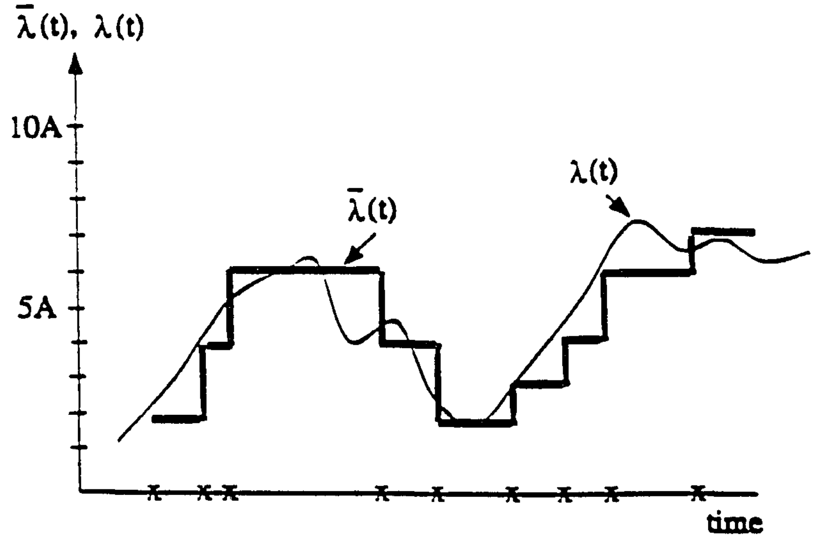


Figure 1.4 Poisson Sampling and Quantization of the Source Rate

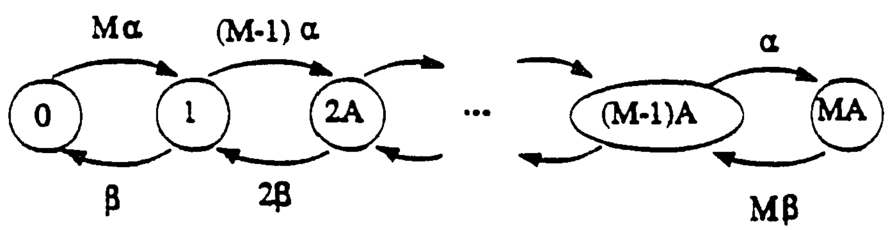


Figure 1.5 State Transition Diagram, Model (2)

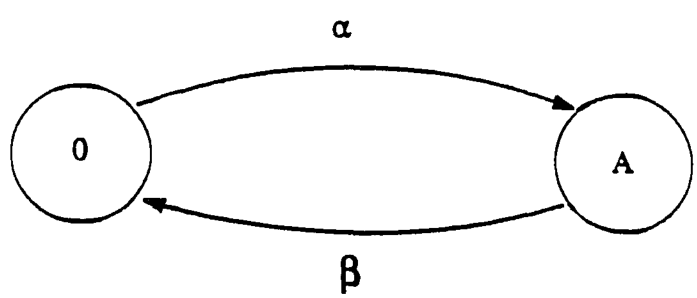


Figure 1.6   Minisource Model

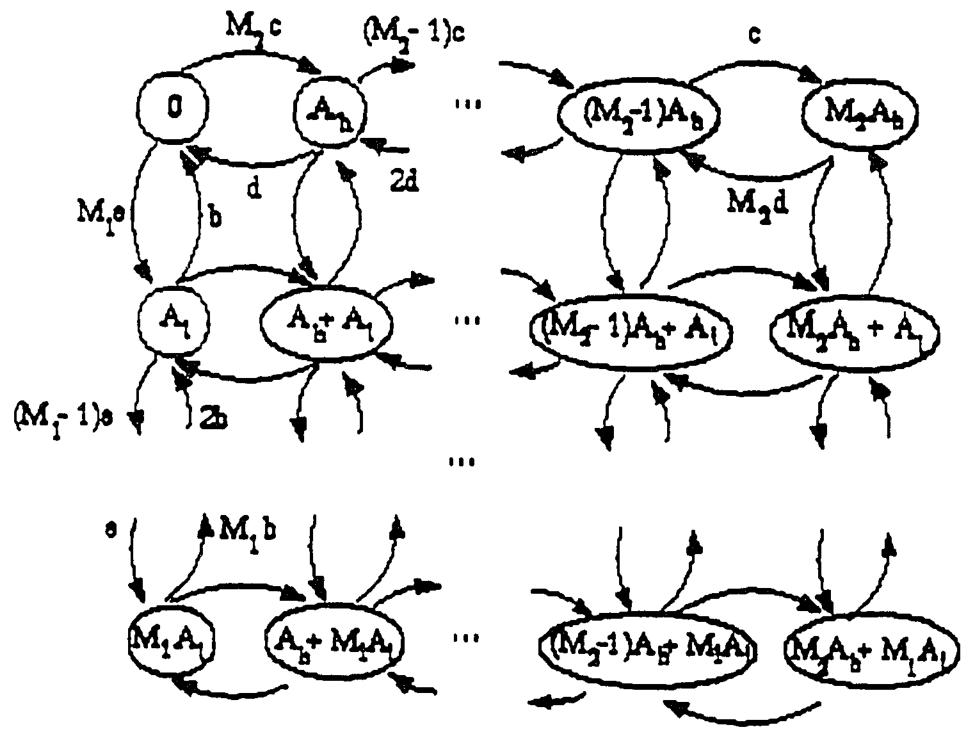


Figure 1.7

State Transition-Rate Diagram for the

Aggregate Source Model, (with Scene Change)

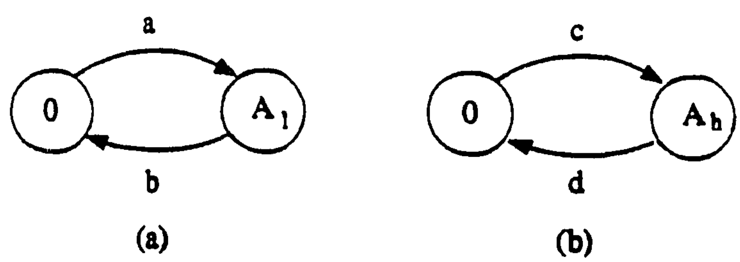


Figure 1.8 Miniprocess Models

2. Simulation Analysis of A Communication Link with Statistically Multiplexed Bursty Voice Sources

2.1 Introduction

Recent advances in fiber-optics, switching and coding compression techniques have made it possible to accommodate integrated multimedia traffic on a high speed communication link. The integrated multimedia traffic (such as video, voice, image and data) is transported and switched using the same interfaces and switching fabrics [57-60]. The expected multimedia traffic covers a very wide spectrum of bit-rate and time-domain characteristics. It ranges from low bit-rate , random (Poissonian), non-real time data-file transfer up to very high bit-rate, bursty and correlated real time video and voice interactive sessions [61-65].

Statistical multiplexing exploits the variance of the source statistics to achieve an efficiency gain measured by the number of sources whose peak bit rate is greater than the link capacity. Since the link capacity is highly utilized, there is a potential for serious congestion problem and cell loss especially at high loads. The characteristics of the bursty traffic complicate the congestion problem even more. Bursty traffic tends to generate bit-rate at or near its maximum for a short durations (burst) followed by longer periods of low or no transmission. A bursty voice source can be characterized by three main parameters: peak bit-rate, average bit-rate and duration of the burst. The bit-rate process resulting from the superposition of a number of voice sources exhibits very high fluctuations over short time periods. This scenario leads to major congestion problems unless an appropriate flow control scheme is maintained. Also, the implicit periodicity in video or voice traffic patterns causes strong dependency and correlations among the cells interarrival times arriving at the multiplexer and queued for transmission over the link. The cumulative correlations among the queued cells cause the queueing process over the multiplexer buffer to deviate substantially from that of the simple Poisson-like one leading to the formation of very long queues even at moderate loads (60% to 70%).

The inherent correlations in the aggregate arrival process from R input voice sources, make it a complex nonrenewal process, hence approximations are used. There has been a considerable amount of work in the literature in this area, and several mathematical models were proposed and analyzed. One approximation is to represent the aggregate arrival process by a continuous time phase process, where the state of the chain represents the number of active source [66]. Another possible representation of the aggregate arrival process is by a two state Modulated Markov Poisson Process (MMPP). The MMPP is a doubly stochastic Poisson Process where the rate process is represented by the state of a two state continuous time Markov chain. The parameters of the MMPP is then matched to some of the statistical moments of the arrival process. Recently, there has been a significant development in the area of stochastic modeling of multimedia traffic sources. A Compound-Phase type Markovian Renewal Process was introduced which models a wide class of phase type processes with batch arrivals, and of which the MMPP is a special case. A Batch Markovian Arrival Process was introduced to model a variety of the versatile Markovian Point Process.

In [67],[68],a mathematical analysis of such arising queueing problems with access control was presented and alluded on the efficacious of the different possible approximations.

The aggregate arrival process is fed to the buffer with fixed size N. The arising process is of the form of $\sum$GI/D/1/K with state dependant arrivals which is too general to solve. Three different approximations were used, one is to model the arrival process with the MMPP with state dependant arrivals and approximate the deterministic service time by an exponential distribution which will lead a continuous time Quasi Birth Death queueing process that can be solved using Matrix-Geometric techniques [68]. In the ATM multiplexer, the buffer size is limited to small values (in the order of microsecs.) and feedback control signal is used to throttle the input peak rate, hence no build up of lengthy queues are allowed and the correlation effects of the cells interarrival times becomes less significant, thus this approximation overestimates the probability of cell loss . Another more closer approximation, is to model the deterministic service time by an Erlang distribution with r-stages, In [68] six stages were used to limit the size of the resulting matrices. Therefore, the Poisson approximation was used based upon the above mentioned reasons.

Using the $\hat{M}$/D/1/K model, the probabilities of number of arrivals during the deterministic service time D were given as:

$$a_i = (\lambda_1 D)^i e^{-\lambda_1 D} / i!, \qquad i \geq 0 \qquad (II.1)$$

$$b_i = (\lambda_2 D)^i e^{-\lambda_2 D} / i!, \qquad i \geq 0 \qquad (II.2)$$

$$c_i = (\lambda_3 D)^i e^{-\lambda_3 D} / i!, \qquad i \geq 0 \qquad (II.3)$$

Where λ_i denotes the average arrival rate from N sources.

The threshold levels at which the flow control is activated were presented where K2 > K1. When the number of cells in the buffer reaches K1, (i.e. congestion level K1), the number of bits per sample drops from 4 to 3 bits per sample. While at K2, the feedback control signal causes the number of bits per sample to drop from 3 to 2 bits per sample.

The steady state transitional probability matrix was given in terms of the three probabilities that were given in (II.1), (II.2) and (II.3) and the mean bits per sample was given as:

$$\overline{B} = 4\sum_{n=0}^{K1} P_n + 3\sum_{n=K1+1}^{K2} P_n + 2\sum_{K2+1}^{N-1} P_n \qquad (II.4)$$

Using the MMPP/M/1/K model, the arrival process is the two state MMPP where the arrivals are state dependant. The infinitesimal generator matrix Q is of size (2Nx2N) where N is the buffer size. The matrix was solved using matrix-geometric techniques, which were modified to suit the overload control case.

Using the MMPP/Er/1/K Model, the service time is approximated by a 6-stage Erlang distribution, Also the infinitesimal matrix was solved using matrix-geometric techniques.

The approximations encountered with the mathematical analysis, often limits the scope of the results rendering simulation tools more powerful and accurate to use.

2.2 Feedback Access Flow Control of Voice Traffic

Figure 2.1 shows the schematic diagram of the access control scheme. It controls the bursty arrival process at the input multiplexer via a feedback control signal that throttles the arrival process during over-load periods. The scheme uses the buffer-length as an indication of possible congestion. The motivation behind this choice is that the voice-cells are blocked only when the buffer is full. Therefore, it is a good apriori indication of possible congestion. As the number of cells in the buffer reaches a first threshold level (K1), a feedback control signal switches the input source-coders to a higher coding rate (i.e., decreasing the number of bits/sample). When another congestion threshold (K2) is reached a similar action is taken thus reducing the input bit-rate further more. For example, consider ADPCM with original coding rate of 4 bits/sample. At K1 threshold level, the coding rate is switched down to 3 bits/sample. While at K2 threshold level, it is switched further down to 2 bits/sample only.

In real-time, it is important that the access-control scheme be robust against the instantaneous and rapid statistical fluctuations of the buffer-length queueing process. Otherwise, there will be frequent control signals triggered by those instantaneous fluctuations. To reduce the controller triggering rate, a parameter called control window(W_c) is introduced. The buffer-length is monitored, and a control signal is generated accordingly, only at the time-epochs that mark the beginning of each control window W_c. The control signal is then fed-back to the coders and maintained unchanged for the duration of the control window irrespective of the buffer-length process. At the following time-epoch (marking the next W_c), the buffer-length is monitored and the process is repeated. For example, if the buffer occupancy at one of those time-epochs is greater than K1, then the controller is activated and as a result the coding rate is reduced to 3 bits/sample. Thus, the control window is the minimum period of time at which the controller stays in a given state before it is being triggered to another action. The set of threshold levels (K1, K2) depends upon the traffic characteristics and the user required Quality Of Service (QOS). The (QOS) is defined by a set of user's service requirements such as cell blocking probability, mean delay, and voice quality. The voice quality is defined as the mean number of bits/sample over the multiplexer buffer. For example a 4 bits/sample is considered to be an "excellent" quality for the voice traffic using ADPCM, while it is regarded as a "poor" quality below 3.5 mean bits/sample.

In our work, we use simulation analysis to gain insight to the performance of the communication link under the above-mentioned conditions. The simulation analysis allowed us to study the system performance at specific time-epochs that mark the beginning of a control time-window (W_c). The rationale behind the introduction of the parameter (W_c) is to simulate the performance of the access control scheme in real-time life. In real-time, it is too difficult to employ an access control scheme that can follow the instantaneous and very rapid changes of the buffer-length stochastic process.

The advantages of this scheme are several. First, it prevents congestion from happening since it decreases the number of cells waiting in the buffer. Second, it is applicable regardless of the type of coder used. It is good for Variable Bit Rate (VBR) coders as well as Fixed Bit Rate (FBR) ones, although the trend is to employ VBR codes due to their perceived gain over FBR codes. Third, it provides the means for the maximum possible shaping of the input arrival process without the need to design complicated time-wave shaping algorithms to regulate the traffic burstiness. It follows then that the bandwidth allocated to the input call can be decreased. Hence, the statistical multiplexing gain is enhanced and a significant improvement in the throughput is achieved. The price to be paid is a slight degradation in the voice quality measured by a decrease in the mean bits/sample. As we will show in latter section, this degradation is very graceful versus the input load.

2.3 Simulation Model

The arrival process for each voice source is represented by a simple two state discrete time Markov chain, where the time that the process spends in each state is approximated by an exponential distribution. Figure 2.2 represents a single voice source model. The voice source alternates between talkspurt (ON) periods and idle (OFF) periods. During the talkspurt period, a fixed number of cells are generated whereas no cells are generated during the idle period. Each period is exponentially distributed with means T_A and T_I respectively. The ratio of the ON period to the sum of both ON and OFF periods is called the activity factor. The average bit-rate of the voice source equals the result of the product of its peak bit-rate by its activity factor. Three different sets of voice traffic statistics are designated in order to study the effect of different burstiness parameters on the link's performance. Bursty sets (1) and (2) have equal activity factors yet different burst lengths, namely set (2) has longer ON, OFF periods than set (1). Bursty set (3) has larger activity factor than burstiness set (1) and set (2). All sets have the same peak bit-rate but different averages.

The communication link is represented by a server with service time equal to L/C where L is the cell length in bits and C is the communication link capacity in bits per seconds. The bursty voice sources access the link via asynchronous statistical multiplexing on a first come first serve (FCFS) basis. The server serves one cell per transmission time which is the unit time. Cells arriving during busy server intervals are queued in the buffer of length M cells. The length of the buffer (M) is chosen to limit the maximum permissible delay per voice-cell.

The simulation program used in this study was written using Simscript II.5 and implemented on Sun workstation. The program contains two main parts, the first is the preamble at which, all global variables and processes are defined, the second is the main at which, the program starts simulation by calling all routines and executing processes in a proper sequence using the timing routine. The flow chart of the program is shown in Figure 2.3. As shown in the figure, the program reads the input data parameters that include the number of input voice sources, the mean talkspurt period, the mean silence period, three different cell interarrival times (T1, T2 and T3) that incorporates the three possible sampling rates for each source (i.e., 4, 3 or 2 bits/sample), the time length for simulation, the cell service time and the buffer length. Once the input data have been read, the program initializes all variables to their appropriate values, creates a server along with a buffer associated with the server and enables all the N arrival voice processes.

In the simulation, two types of processes are defined for each voice source. One process is defined to generate cells, called cell arrival process, while the other serves the generated cells and is called cell service process. N voice cell arrival processes are defined where N is the number of sources (or load). Each consists of consecutive talkspurt (active) and silence periods, which in-turn are generated using the built-in exponentially distributed random variable generator. The arrival process generates cells only in the talkspurts periods where the interarrival time between the cells is the cell generation time. The mean silence period is the mean interarrival time between the talkspurts.

According to the access control scheme, each of the N voice sources generates cells with a coding rate that corresponds to 4 bits per sample as long as the buffer occupancy level is in the underload interval (i.e., the number of cells in the buffer is less than the first threshold K1). In this case, the interarrival time between cells, T1, is equal to the cell length in bits divided by the source peak bit-rate in bits per second. When the buffer occupancy level is bounded between the threshold levels K1 and K2 (transitional interval from the underload to the overload period), the coding rate drops to 3 bits/sample thus decreasing the source peak bit-rate correspondingly, therefore increasing the interarrival time between cells to T2. As the buffer occupancy reaches the overload interval (i.e.,the number of cells in the buffer increases beyond the second threshold K2) the coding rate drops further from 3 to 2 bits per sample hence the source peak bit-rate is reduced furthermore and the interarrival time between cells is increased to T3.

Each generated cell requests the service process for service, if the server is busy, the cell is queued in the buffer. An arriving cell will get blocked if the buffer is full. If the server is idle, the cell keeps the server busy for a time equal to the service time then it relinquishes the server. All the simulation events are scheduled by a timing mechanism or routine as shown in Figure 12. The timing routine scans the event set in order to pick up the next event to be executed. It removes the most imminent process notice from the future process set, updates the simulation time to the process time indicated, and passes control to the routine for this process. Upon completion of this process, the timing routine again turns to the future process set to determine the next process routine to be executed. This sequencing continues until all process notices in the future process set are exhausted. When this happens, control is returned to the statement directly after the start simulation statement. This process continues until the simulation time expires.

In Simscript II.5, the global variables were used to represent the number of generated cells and the number of blocked cells during the simulation times. The blocking probability was measured by evaluating the ratio of the number of blocked cells to total the number of transmitted cells during the simulation run time. We run several simulations with different seeds to evaluate the link performance with and without access control and under different burstiness sets. After the simulation time expires, the program terminates the simulation and calculates the final output results in the final report routine. By using Simgraphics package, all interesting performance can be plotted via a routine called Trace-plot which traces the values of the time dependent parameters. Some of the performance evaluation results are the histogram of the number of cells in the buffer with and without access control, the optimum values of the thresholds K1 and K2 that give us the desired value of the blocking probability, the mean bits/sample and the control triggering rate. The following section provides a detailed simulation analysis of the link performance with and without control.

2.4 Simulation Results and Discussions

In this section we discuss the simulation results obtained from this study. Assuming ADPCM coding technique at 4 bits/sample, then the source peak bit-rate is 32 kbps. As the first threshold level K1 is reached, the source peak bit-rate is reduced to 24 Kbps. As the second threshold K2 is reached, the source peak bit-rate drops to its final overload value of 16 kbps. The cell length is fixed and is chosen to be 53 bytes (with an overhead of 5 bytes) to suit the ATM application standards. The communication link capacity is set to be equal the T1 interface rate of 1.544 Mbps. Then the cell interarrival times, T1, T2 and T3 are equal to 12ms, 16ms and 24ms respectively and the cell service time is equal to 0.256 milliseconds. In order to limit the maximum delay of the voice cells to 5 milliseconds which is the maximum acceptable value for voice quality, the buffer length (M) is set to 20 cells. To study the efficacious of the access-control scheme with different bursty traffic, we use three possible burstiness sets. Burstiness set (1) corresponds to a mean talkspurt period of 352 milliseconds and a mean silence of 650 milliseconds, therefore the average burst length is 352ms and the activity factor is 0.35. Burstiness set (2) corresponds to a mean talkspurt period of 1200 milliseconds and a mean silence period of 2200 milliseconds with an activity factor of 0.35. Burstiness set (3) corresponds to a mean talkspurt period of 450 milliseconds and a mean silence period of 550 milliseconds with an activity factor of 0.45.. We evaluate the link performance under different offered loads ranging from 80 (load of 56%) to 150 (load of 100%) input voice sources. Each simulation is run for a period of 5 minutes which is sufficient to evaluate blocking probabilities in the order of 10^{-4} and our confidence interval is 95%.

Figure 2.5 represents the histogram of the number of cells in the buffer for burstiness set (1), set (2) and set (3) at heavy load. For burstiness set (1), with 150 sources, a link utilization of 100% is achieved, since the link utilization equals the number of voice sources multiplied by the ratio of the average bit-rate to the link capacity. Similarly the maximum loads supported are 130 and 120 for burstiness set (2) and set (3) respectively. The link is congested and the buffer is full 33.9%, 37.9% and 38.6% of the simulation time for burstiness set (1), set (2) and set (3) respectively. This leads to unacceptable blocking probabilities which causes a severe degradation in the delivered voice quality. Comparing burstiness set (1) and set (2) where the ON period has been increased to 1200 milliseconds for set (2), we observe that, using burstiness set (2), there has been a significant degradation in the performance of the system (the link is congested and the buffer is full 37.9% of the simulation time). Similarly comparing burstiness set (1) that has activity factor of 0.35 and set (3) that has activity factor of 0.45, we notice that, using burstiness set (3), the performance of the system has significantly degraded (the link is congested and the buffer is full 38.6% of the simulation time). It is easy to expect this results since the link has been overloaded with a higher average rate, in burstiness set (3), than the average rate of burstiness set (1).

Figure 2.6 shows the cell blocking probability versus the number of voice sources for the three burstiness sets without control. As shown in this figure, our objective value of cell blocking probability of 10^{-4} is satisfied for all burstiness sets when the number of voice sources is equal to or less than 100, 90 and 80 sources for burstiness set (1), set (2) and set (3) respectively. When the number of voice sources increases the cell blocking probability increases. The difference in cell blocking probability between burstiness set (1), set (2) and set (3) is that set (2) has more burstiness than set (1), and set (3) has more burstiness than set (1) and set (2).

In order to find the optimum sets of the control threshold levels (K1) and (K2) for each arrival characteristics (burstiness set), we postulate an optimization problem. We set an objective blocking probability of 10^{-4} as an acceptable rate for delivering voice traffic with high quality. We obtain the optimum values of all the sets (K1, K2) by running the simulation program for 5 minutes (i. e. The average duration of a voice call) using different seeds. Table 2.1 shows all optimum values of (K1, K2) for each burstiness set. The maximum load that can be supported by the control scheme is 150, 130 and 120 sources for burstiness set (1), set (2) and set (3) respectively. The notation (*) implies that the required QOS is achieved with no control threshold. For burstiness set (1) the optimum values for K1 and K2 are 11 and 17 to support 120 sources. For the same number of sources, K1 and K2 are 1 and 2 for burstiness set (3). Since the threshold levels K1, K2 have dropped to such small values for burstiness set (3) we conclude that it is not possible to support those 120 voice sources for burstiness set (3). The reason is that at such small threshold levels the average number of bits per sample over the buffer length would drop significantly leading to a degradation in the voice quality below acceptable levels.

It is easy to see from Figure 2.7, that as the access scheme is applied, a dramatic shift in the buffer length has happened. The peak of the graph has shifted towards the zero buffer occupancy level. The performance of the system has improved almost equally in all cases using the access control scheme. It is evident from the curve that the access control scheme has greatly reduced the mean number of cells in the buffer, hence reducing the blocking probability significantly and thus avoiding congestion. It then becomes possible to support certain number of voice sources at a less amount of allocated bandwidth which we call the bandwidth gain. Not only that, but the statistical multiplexing gain has also improved significantly since it has become possible to support more voice sources without increasing the allocated bandwidth. For example, without the access control the maximum number of supported sources to achieve the required QOS (blocking probability and mean bit/sample) are 100, 90 and 80 for burstiness set (1), set (2) and set (3) respectively as seen in Figure 2.7. While using control, the number of supported sources increases to 150, 130 and 120 respectively as seen in table 2.1.

Table 2.2 shows the optimum values of (K1, K2) for three chosen values of control window(W_c), namely 20ms, 100ms, and 300ms. As seen in Figure 2.8, for all values of W_c, the controller triggering rate increases as the load increases up to 140 sources, after which it saturates at heavy load situations. For all different cases of W_c, as the control window increases the controller triggering rate decreases. However, the quality of the voice (i. e., Mean Bits/Sample) has been reduced as shown in Figure 2.9. For example, at 130 sources, the mean bits/sample is 3.8 bits with no access control window, yet it drops to 3.3 bits when the control window is 300ms.

In Figure 2.10, it is clear that the individual blocking probability seen by each source varies significantly. In the No-Control case, although the average blocking probability, as calculated over the buffer length, was 6×10^{-3} for 120 sources, this value has been satisfied for only 80 sources. The remaining 40 sources had encountered blocking probability greater than 6×10^{-3}. This problem persisted even when the access control scheme was applied in Figure 2.11. In this case, our QOS objective was a blocking probability of 10^{-4}, which had been satisfied for only 60 sources out of the 120 sources. It is clear that the access control algorithm treats sources with the same characteristics unequally. To rectify the problem, we applied the access control scheme only to the sources that had encountered a blocking probability more than 10^{-4}. We call this control scheme "Control (B)" and the original control scheme "Control (A)". As we can see, in the case of Control (B), the performance improved significantly with only 16 sources not meeting the required QOS.

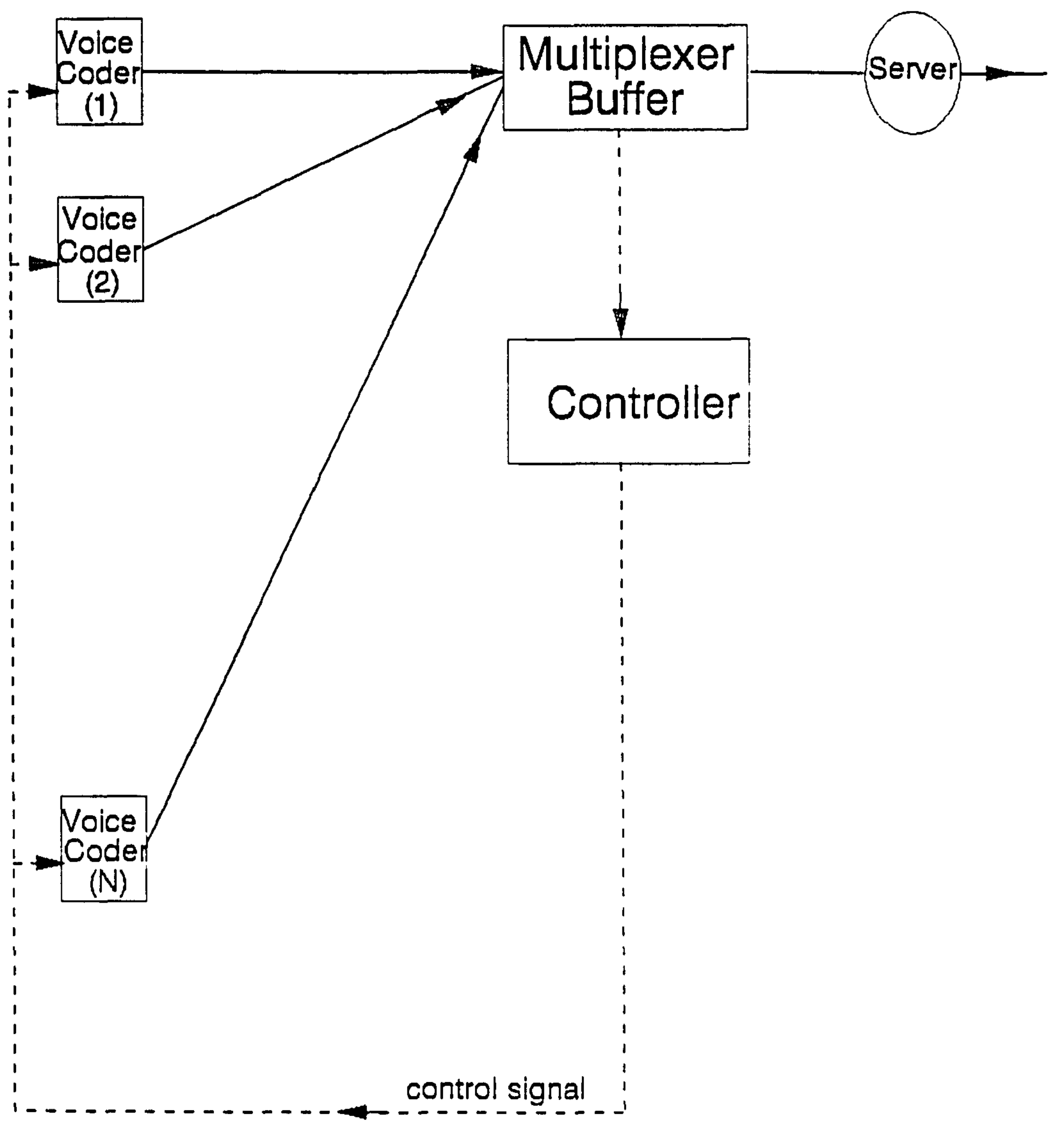


Figure 2.1 Voice multiplexer with feedback control

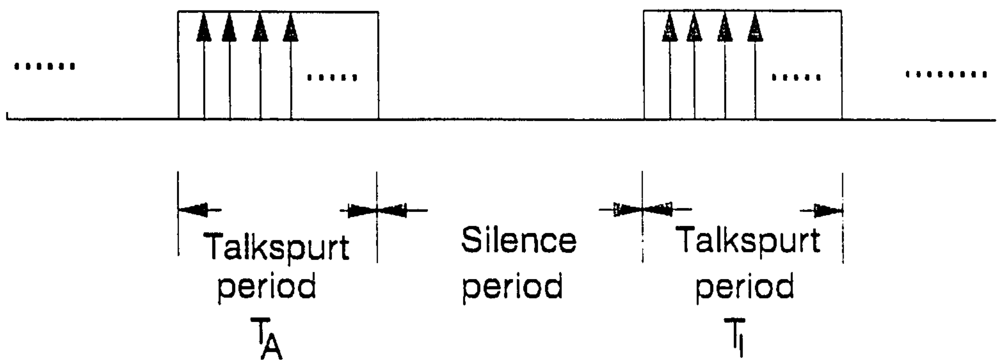


Figure 2.2 Single voice source model

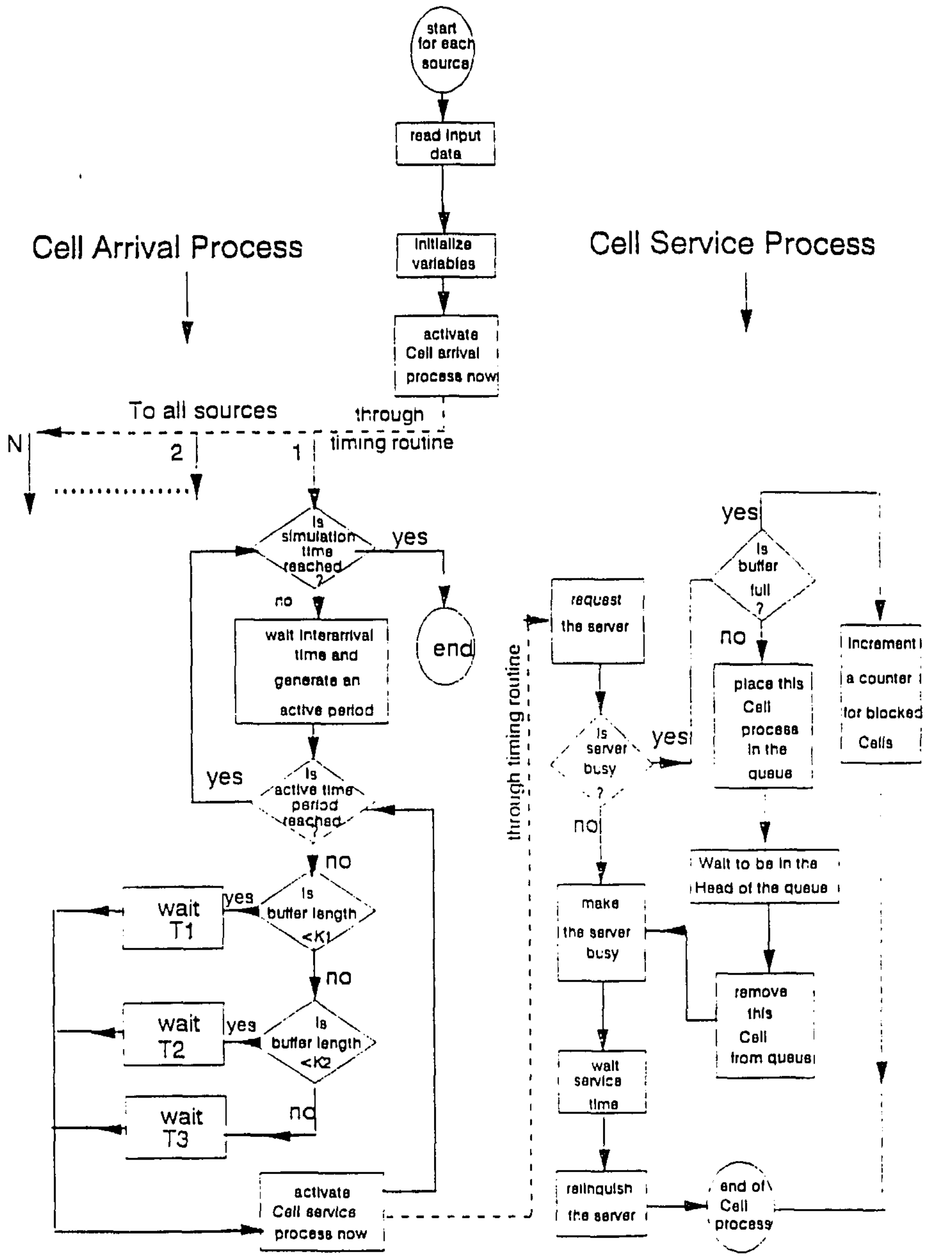


Figure 2.3 Flow Chart of the Simulation Model

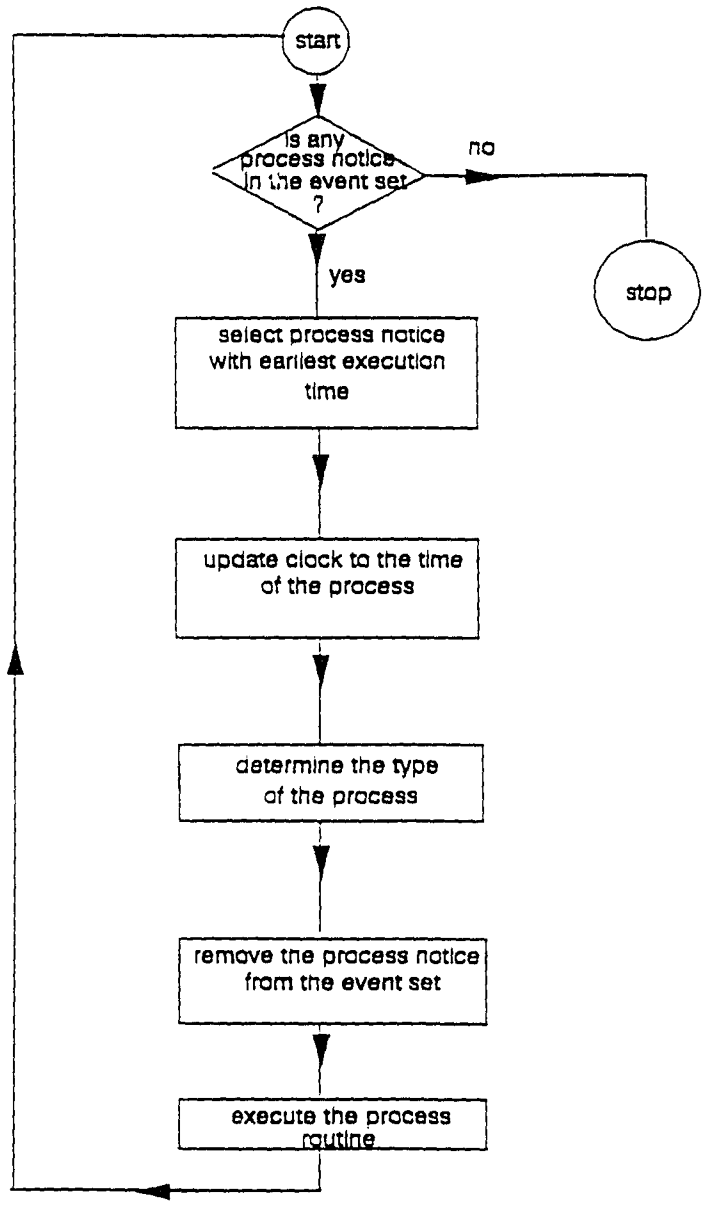


Figure 2.4 The Timing Routine

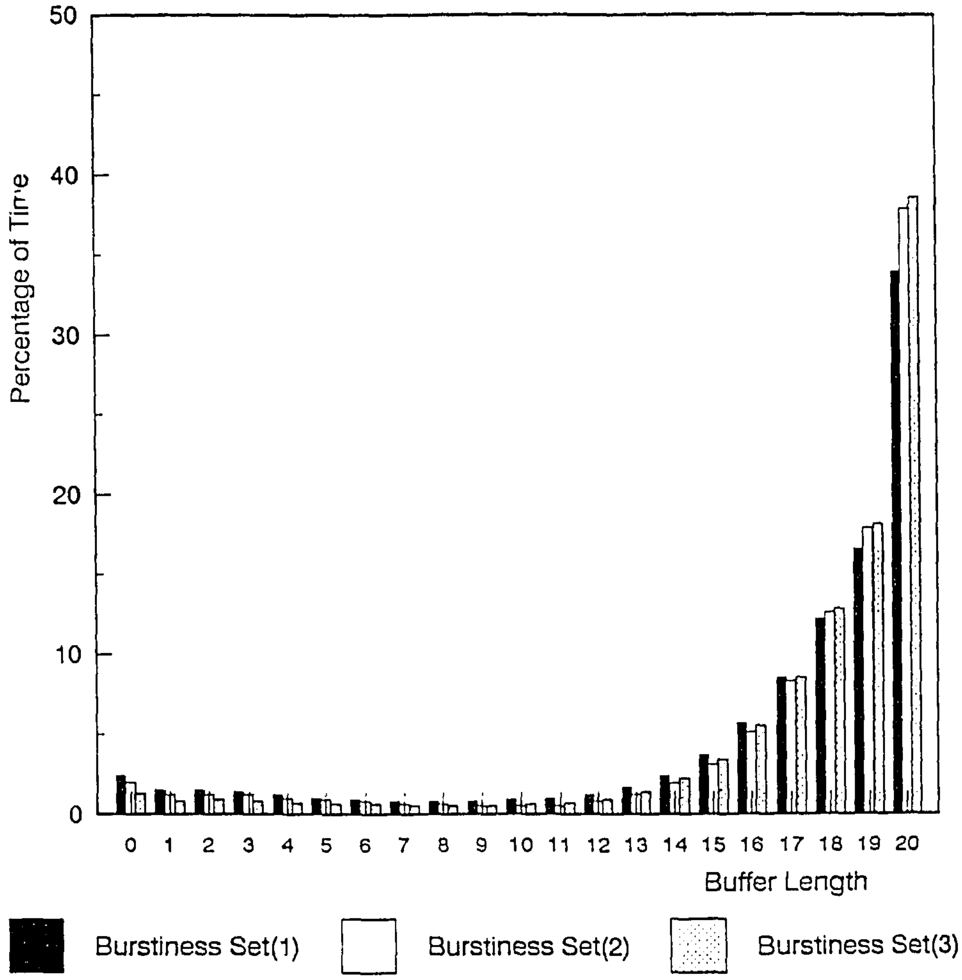


Figure 2.5 Histogram of the Buffer Length for Three Different Burstiness Sets at Heavy Load Utilization=100%, without Control

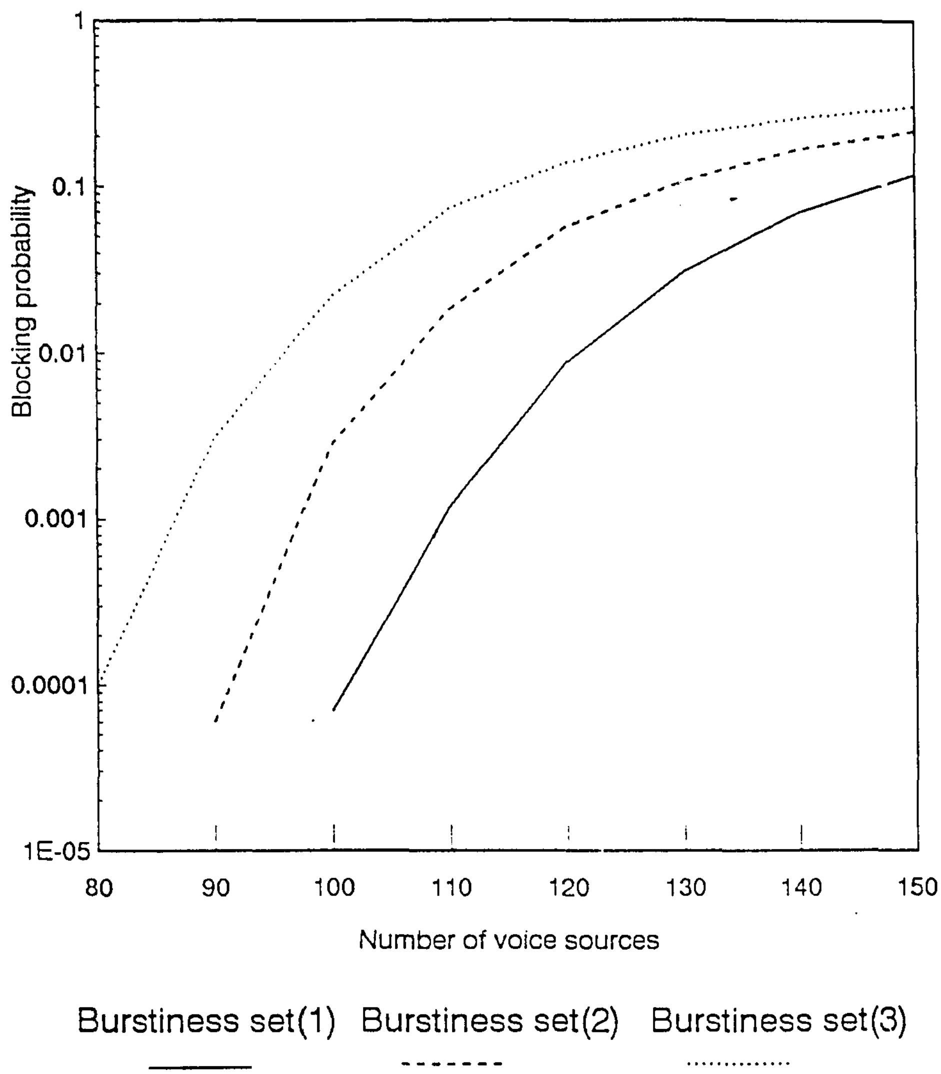


Figure 2.6 Blocking Probability Versus Traffic Load for Three Different Burstiness Sets

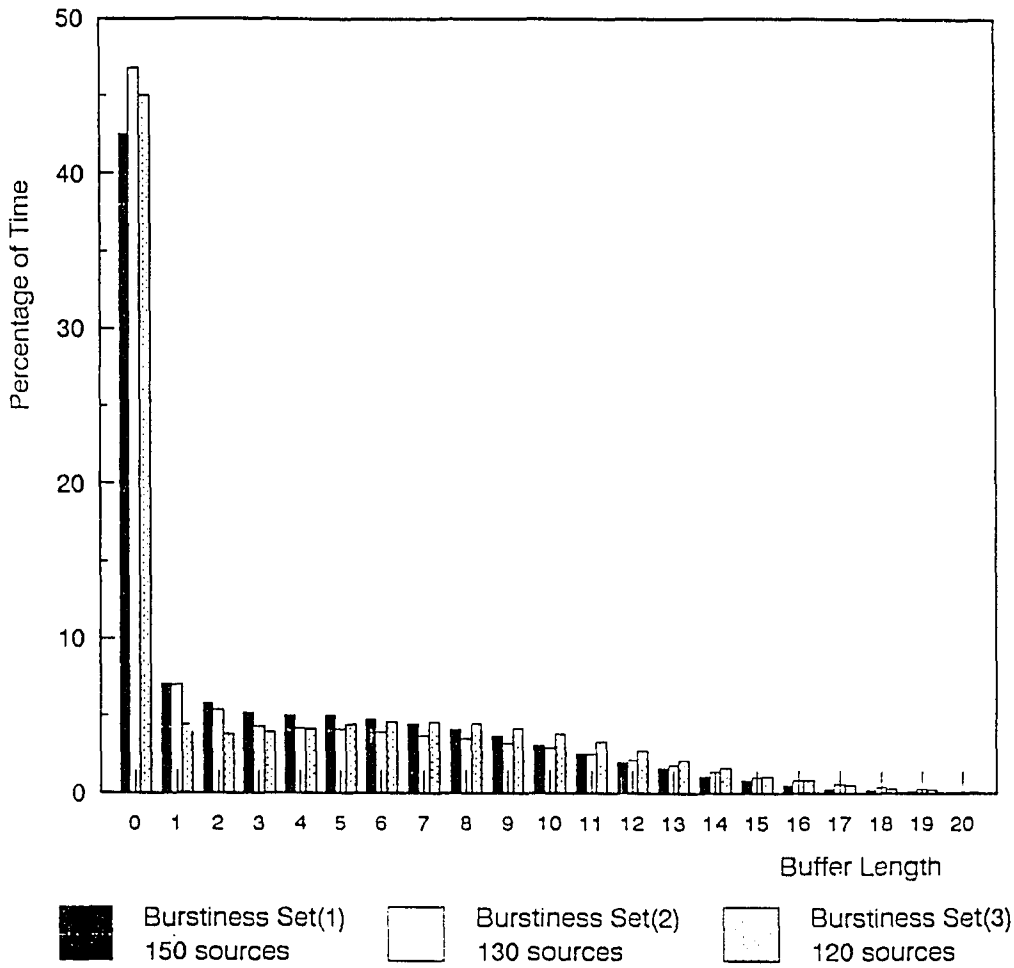


Figure 2.7 Histogram of the Buffer Length for Three Different Burstiness Sets at Heavy Load with Control

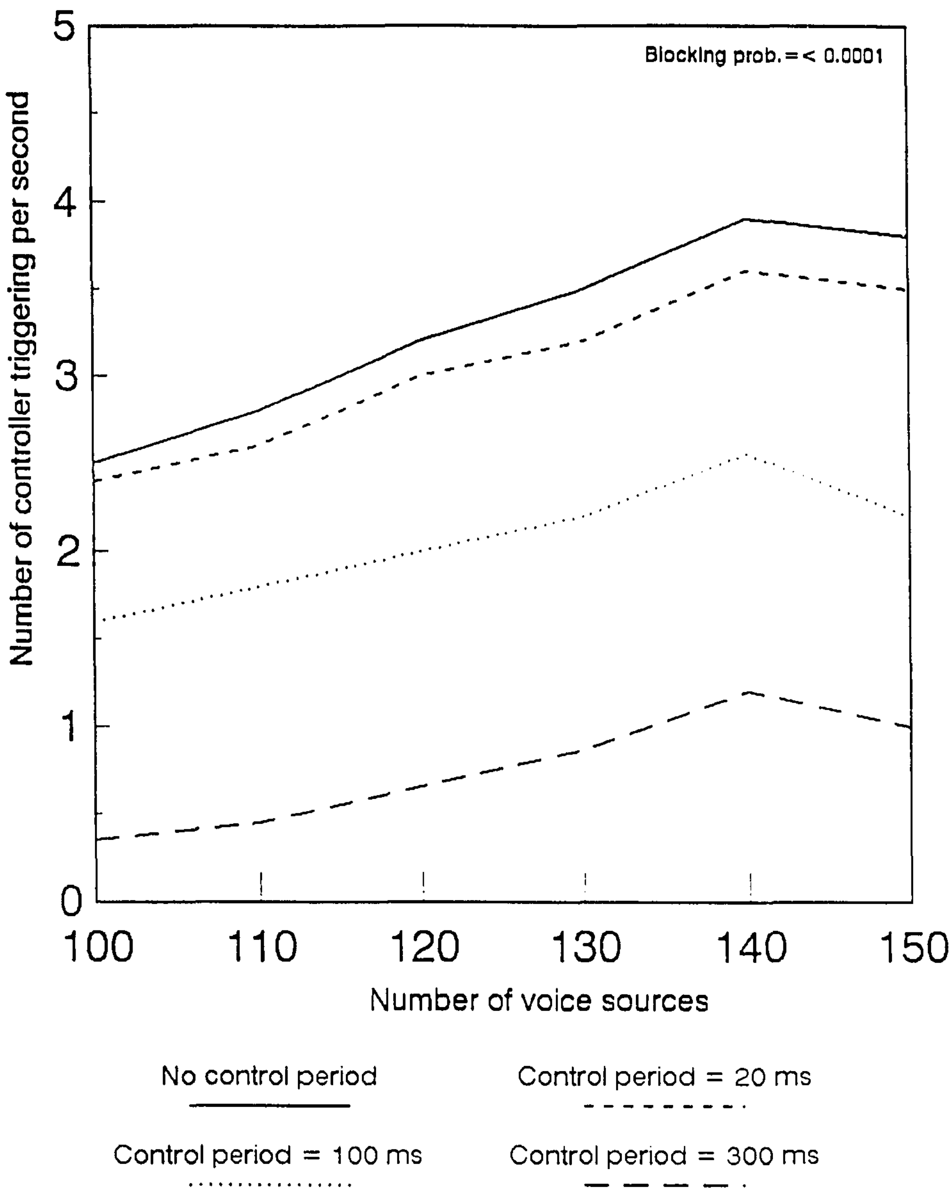
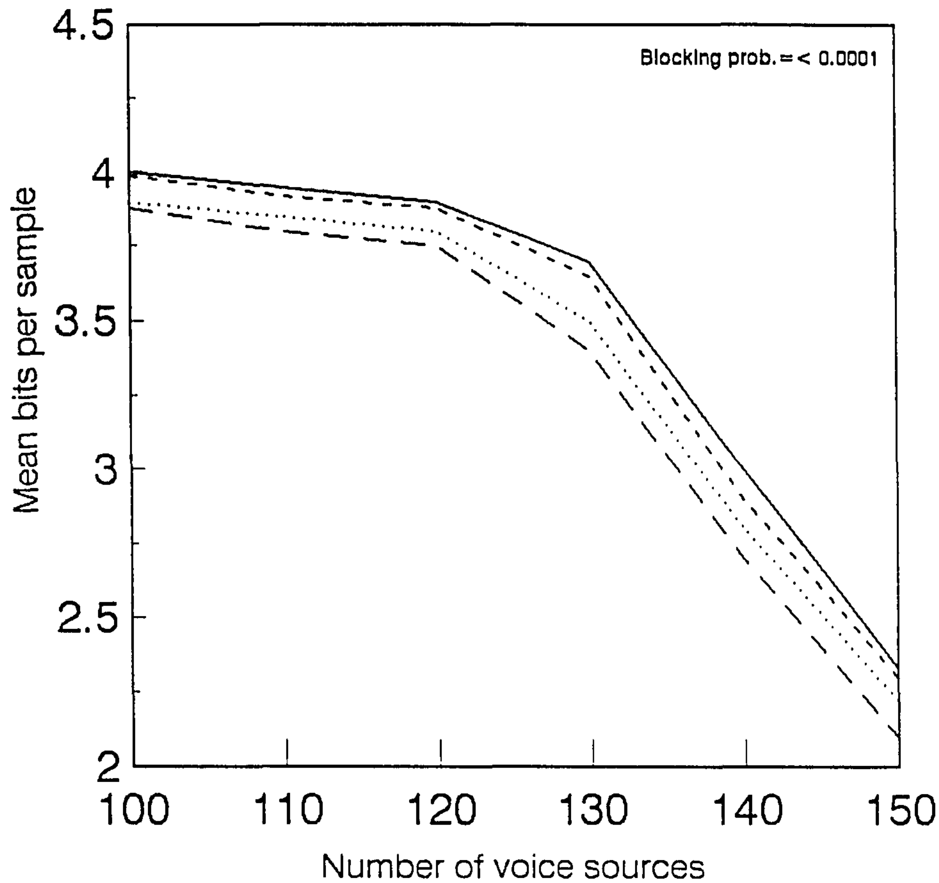


Figure 2.8 Controller Triggering Rate

Burstiness set (1)



No control period ——————

Control period = 20 ms - - - - - - - - - - -

Control period = 100 ms ....

Control period = 300 ms — — — — — .

Figure 2.9 The effect of the control window on the mean bits per sample
Using burstiness set(1)

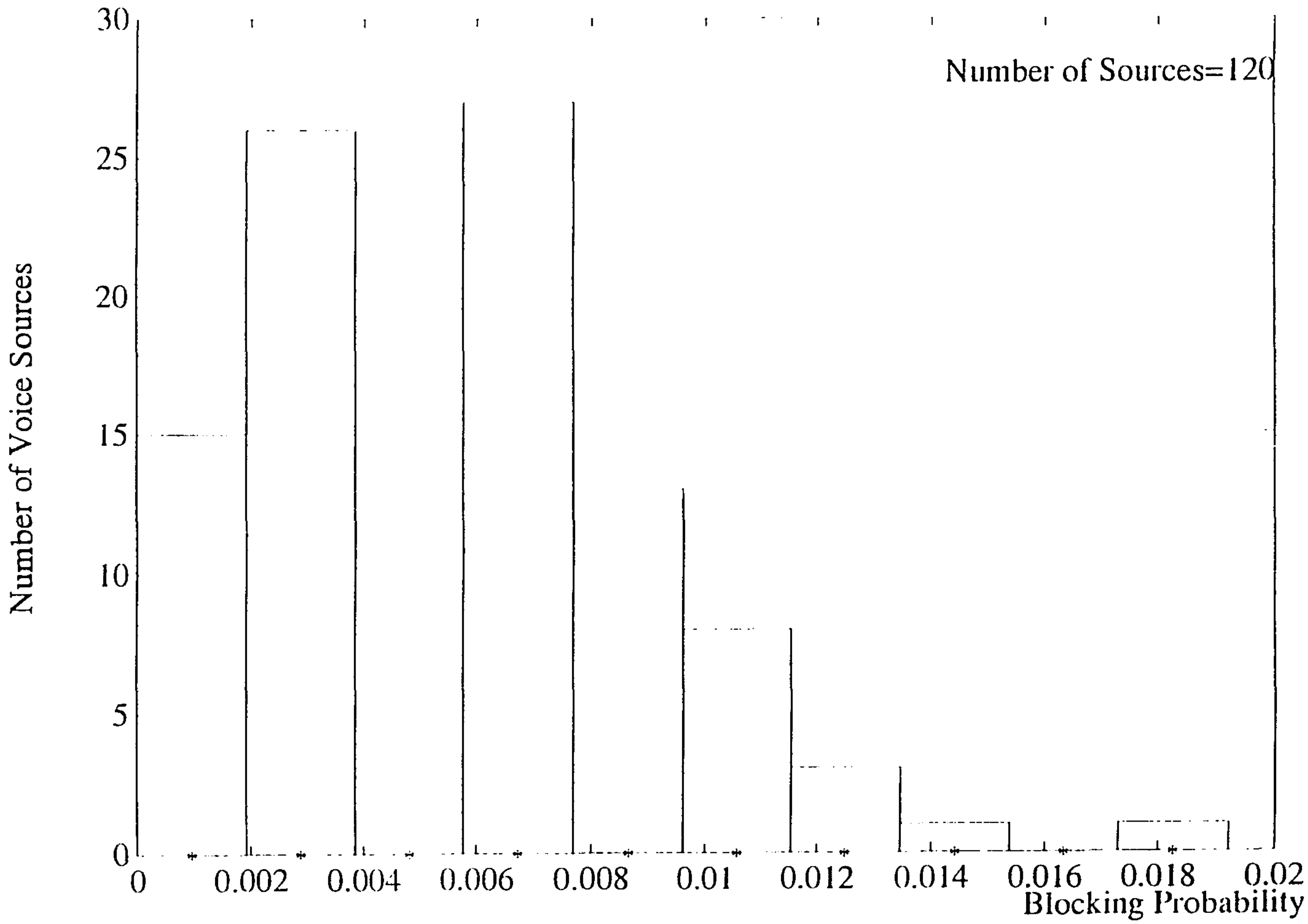


Figure 2.10 Individual Blocking Probability per each Source, No Control

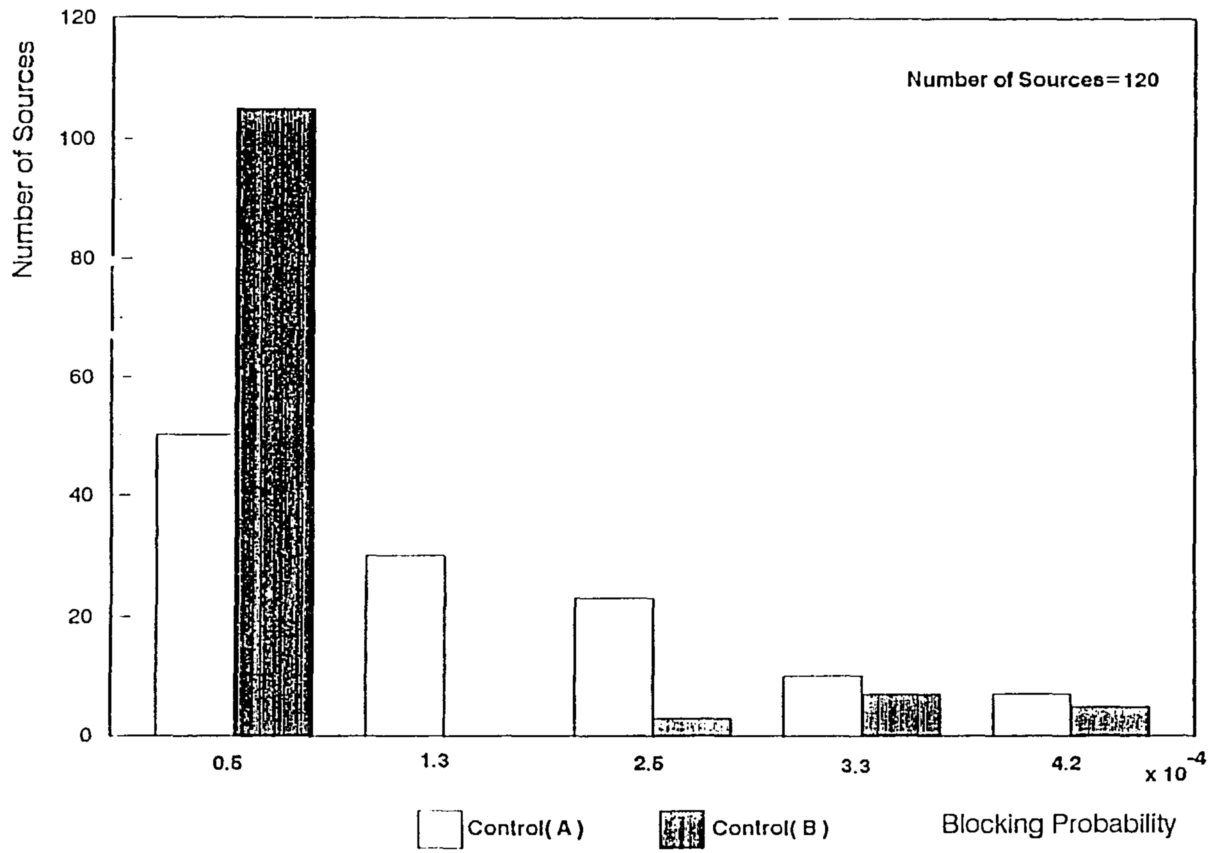


Figure 2.11 Individual Blocking Probability per each Source, with Control

Number of voice sources	K1	K2	K1	K2	K1	K2
	Burstiness set(1)		Burstiness set(2)		Burstiness set(3)	
80	*	*	*	*	*	*
90	*	*	*	*	14	*
100	*	*	15	*	8	14
110	15	*	10	15	3	4
120	11	17	4	5	1	2
130	8	15	2	3	-	-
140	6	12	-	-	-	-
150	5	10	-	-	-	-

Table 2.1: Optimum values of K1 and K2 for three different burstiness sets
Blocking prob.=0.0001

Number of voice sources	K1	K2	K1	K2	K1	K2
	Control period = 20 ms		Control period = 100 ms		Control period = 300 ms	
100	*	*	17	*	10	16
110	12	18	10	16	8	14
120	10	16	8	15	6	12
130	7	14	6	12	4	8
140	5	10	5	10	3	6
150	4	8	3	4	2	3

Table 2.2: Optimum values of K1 and K2 for different control periods
Burstiness set (1)
Blocking prob.=0.0001

3. Simulation Analysis of An ATM Access-Node with Statistically Multiplexed Homogeneous and Heterogeneous Traffic

3.1 Introduction

It is expected that video will be a major source on ATM networks, which will place a very large bit rate requirement on the network. Due to statistical variations in the bursty traffic, congestion may occur in the link with a number of sources sharing the same link. Congestion can seriously degrade the quality of service which is the cell loss and delay. Therefore traffic control is necessary to avoid congestion.

In this chapter, we study the performance of a statistical multiplexer whose inputs consist of a superposition of packetized video sources and a superposition of packetized voice and video sources. Also, we study the performance under an access flow control scheme and bandwidth assignment. The basic objective of a bandwidth management and traffic control scheme is to allow for high utilization of network resources, while sustaining an acceptable quality of service for all users.

3.2 Feedback Access Flow Control of Video Traffic

In ATM Networks, Congestion occurs when various sources compete for the network resources and these resources do not meet the demands. Due to the stochastic nature of the traffic, long bursts of cells can get formed even when the network resources are underutilized. The call acceptance rule in ATM networks is based upon the availability of bandwidth to support the call along its connection path. As the traffic gets more bursty, more bandwidth is required to support the call at the same QOS, else congestion occurs. It is then extremely important to enforce a congestion control scheme at the cell level that can dynamically interact with the call level control such that the traffic gets smoother, and the bandwidth allocated is utilized to the maximum extent.

In [69], the model represents the arrival rate $\lambda(t)$ by quantizing the bit rate into uniform discrete levels, and the rate variations over time are approximated by a continuous time process with discrete jumps at random Poisson times. Thus the state space (A) of the chain represents a quantization level of the original sampled process, measured in bits/pixel. The (M+1) states scan the range of the variations. The parameters α and β are the transitional rates of jumping from one quantization level to the other and these parameters were evaluated in [45] and the results are,

$$A = \frac{C_R(0)}{E(\lambda_R)} + \frac{E(\lambda_R)}{M} \qquad (III.1)$$

$$\beta = 3.9/\left(1 + \frac{E^2(\lambda_R)}{M C_R(0)}\right) \qquad (III.2)$$

$$\alpha = 3.9 - \beta \qquad (III.3)$$

Where $E(\lambda_R)$ and $C_R(0)$ are the average and the variance of the aggregate arrival process from R identical and independent sources. Each source transmits a random process with mean $E(\lambda)$ and autocovariance function $C(\tau) = C(0)e^{-3.9\tau}$. τ is the source frame number n divided by a frame rate of 30 frames/sec. The multiplexer buffer has a fixed buffer length N cells and is fed by the process described by the system of equations in (III.1-III.3). The system equations were written in the matrix form $PQ = 0$ where P is the steady state probabilities vector, Q is the transitional rates matrix. The matrix Q contains submatrices of dimension M. These matrices can be of extremely large sizes, yielding numerical difficulties in solving them. For example, if the number of sources are 10 video sources, and the buffer size is 100 cells, then the dimension of the matrix Q is 10,000 which is impossible to solve using direct matrix manipulations. Matrix-geometric techniques are used to solve the above system, the solution uses an iteration refinement technique which is needed to be slightly modified to suit the overload control in this case.

The approximations encountered with the above mathematical analysis, often limits the scope of the results rendering simulation tools more powerful and accurate to use. In our study, we use SimscriptII.5 simulating language to gain insight to the performance of the communication link. The simulation study covers a homogeneous case where N video sources are multiplexed and a heterogeneous case where a mix of voice and video sources are multiplexed.

Figures 3.1 and 3.2 show the schematic diagram of the access control scheme that applied to video multiplexer and voice/video multiplexer respectively. It controls the bursty arrival process at the input multiplexer via a feedback control signal that throttles the arrival process during over-load periods. The scheme uses the buffer-length of the multiplexer as an indication of possible congestion. The motivation behind this choice is that the arriving packets are blocked only when the buffer is full. Therefore, it is a good apriori indication of possible congestion. As the number of packets in the buffer reaches a first threshold level (K1), a feedback control signal switches the input source-coders to a higher coding rate (i.e. , decreasing the number of bits/sample). When another congestion threshold (K2) is reached a similar action is taken thus reducing the input bit-rate further more. For example, consider ADPCM with coding rate of 8 bits/sample. At threshold level K1, the coding rate is switched down to 6 bits/sample. While at threshold level K2, it is switched further down to 4 bits/sample only.

3.3 Bandwidth Allocation to Voice and Video Traffic

ATM networks are capable of supporting a wide range of connections with different bandwidth requirements and traffic characteristics. While this environment provides increased flexibility in supporting various services, its dynamic nature poses difficult traffic control problems when trying to achieve efficient use of network resources. One such problem is the issue of bandwidth management and allocation. Because all connections are statistically multiplexed at the physical layer and the bit rate of connections varies, a challenging problem is to characterize, as a function of the desired quality of service (QOS), the effective bandwidth requirement of both individual connections and the aggregate bandwidth usage of connections multiplexed on a given link. The basic objective of a bandwidth management is to allow for high utilization of network resources, while sustaining an acceptable quality of service for all connections.

In [70], three different types of traffic, video, voice and data are being multiplexed into the outgoing link. Cells belonging to each type of traffic are queued in a separate buffer and then multiplexed on a separate virtual path. The controller reads the input traffic characteristics and its required QOS and, based upon the bandwidth availability, either accepts the call or rejects it (CAC algorithm). As the number of queued cells, in each buffer, reaches a certain threshold, the access control algorithm throttles the input bit-rate by compressing the coding rate. At a certain desired bandwidth utilization (e.g. 0.8), the threshold levels required to support the voice calls at the voice buffer are quite different from those required at the video buffer.

The controller allocates the bandwidth to users sharing the same virtual path according to a predetermined rule which is based upon exploiting the statistical multiplexing gain. The bandwidth required, is less than the peak rate and greater than the average rate by some bandwidth allocation factor. This factor is function of the arrival statistics, the required QOS (i.e., cell loss rate, delay) and the number of multiplexed calls per each virtual path. This bandwidth is just enough to guarantee the call its QOS which is determined by an acceptable cell loss rate and maximum permissible delay. The delay is easily guaranteed by limiting the buffers' size. Whereas, the cell loss rate is guaranteed by the amount of allocated bandwidth.

The scheduler schedules transmission of cells, from each queue, such that the bandwidth allocated to each type of traffic, and hence the virtual path capacity, equals the value driven from the bandwidth allocation table.

The scheduling of cells from separate queues, is controlled via the Bandwidth Control Period (BCP) rule. The BCP, is a time-frame that is used, by the scheduler, to control the access to the link by scheduling both types of traffic (video and voice) such that they are always guaranteed their corresponding allocated bandwidth. The controller selects the required bandwidth to support each type of traffic according to its required QOS. This information is then used by the scheduler to initialize a value for the BCP, which specifies a clock frequency to control the switching speed of the scheduler. The controller monitors both the arrival statistics and the buffers' lengths to maintain the required QOS and avoid congestion. Whenever the controller detects a change in the arrival statistics (or a possible congestion) then it updates the BCP value such that the QOS is always maintained for each type of traffic. Hence, the assigned bandwidth to each virtual path will be also updated to accommodate any new changes in the traffic characteristics. For example, an increase in the number of accepted voice calls would imply a corresponding increase in the bandwidth allocated to the voice-traffic virtual path, at the expense of a decrease in the bandwidth allocated to the video-traffic virtual path. The BCP value, which is the sum of the two windows, could still be the same (in case there is no change in the queueing process performance) or it could be updated to accommodate any traffic variations that may lead to an increase in the number of queued video (or voice) cells. The BCP rule provides simple means to control each virtual path capacity in response to such traffic variations, by simply changing the width of the access-control time-frame. The performance of the queueing process (in terms of the cell loss rate) improves when the BCP value decreases (i.e., the switching time from one queue to the other is small).

The arrival process resulting from the superposition of M video sources depends upon the contents of the scene and the type of video coding algorithm implemented. The model that is used here assumes single motion activity scenes. The aggregate video arrival rate is also modeled by a Continuous-Time Discrete-State Markov Chain, where the state of the Markov chain represents a certain quantized bit-rate level of the bit-rate arrival process. Hence, the state of the Markov chain, (ia), $i \in (0, 1, .. N_v)$, scans all possible changes in the bit-rate process from zero to the maximum possible bit-rate N_v. The parameter N_v, is equal to the maximum number of quantization levels that represent the bit-rate process. The Markov chain transition-rate matrix, representing the video arrival process, has a dimension of (N_v+1). This model, yields an extremely high number of states, thus leading to numerical difficulties. For example, consider the multiplexing of 10 video sources, then the number of the Markov chain states will be 100 different states covering all possible bit-rates transmissions.

The infinitesimal generator for the queueing process for each of the voice and video queues were presented. The problem, here, is to solve two separate queues each of the type PH/PH/1/K, for the cell loss probability. The solution of the these matrices involve numerous numerical calculations using matrix geometric techniques.

In our study, we use simulation to analyze a bandwidth control scheme for integrating packetized voice and video traffic in an access node in ATM network. In this scheme, voice and video are queued separately to facilitate dynamic bandwidth sharing and mutual overload protection. It guarantees bandwidth to voice and video traffic by reserving N1 and N2 queue length limits for transmitting voice and video packets (cells) respectively. When one queue is exhausted, the transmission is immediately moved over to the other queue if it is not empty. Our simulation results show that the scheme enables it to meet the disparate performance requirements for voice and video traffic and to increase the efficiency of transmission bandwidth usage.

In Figure 3.3, three bandwidth allocation algorithms to voice and video traffic are shown. Figure 3.3a shows a statistical bandwidth allocation to voice and video traffic. In this algorithm, a transmission bandwidth is shared between the voice and video traffic where the voice and video cells are treated equally on a first come first serve (FCFS) basis. Figure 3.3b shows a deterministic bandwidth allocation to voice and video traffic. In this algorithm, a transmission bandwidth is deterministically subdivided for video and multiplexed voice traffic. Peak bit rate (PBR) bandwidth is deterministically allocated to video traffic, and the rest of the bandwidth is allocated to multiplexed voice traffic. Figure 3.3c shows a bandwidth control window allocation to voice and video traffic. In this scheme, voice and video are queued separately to facilitate dynamic bandwidth sharing. The scheduling of cells from separate queues, is controlled via the optimum values of the bandwidth control windows. When one queue is exhausted, the transmission is immediately moved over to the other queue if it is not empty. The bandwidth control window, is set to be the scheduler maximum switching period of time required to support both types of traffic (voice and video), such that their respective cell loss rates and delays are delivered.

3.4 Simulation Model

The simulation program used in this study is written using Simscript II.5 and implemented on Sun workstation. The flow chart and the timing routine are the same as given in Figure 2.3 and Figure 2.4 respectively that were shown and explained in the previous chapter.

The communication link is represented by a server with service time equal to L/C where L is the packet length in bits and C is the communication link capacity in bits per seconds. The bursty sources access the link via asynchronous statistical multiplexing on a first come first serve (FCFS) basis. The server serves one packet per transmission time which is the unit time. Packets arriving during busy server intervals are queued in the buffer of length M packets. The length of the buffer (M) is chosen to limit the maximum permissible delay per packet. An arriving packet will get blocked if the buffer is full. If the server is idle, the packet keeps the server busy for a time equal to the service time then it relinquishes the server.

Two types of processes are defined for each source. One process is defined to generate packets, called packet arrival process, while the other serves the generated packets and is called packet service process.

3.4.1 Statistical Multiplexing of Video Sources

In the case of homogeneous sources, we defined N packet arrival processes where

N is the number of video sources. We assume video sources generating 30 frames/sec. We model the bit rate of a single video source during the nth frame by a first order autoregresive Markov process. A first-order autoregresive Markov process R(n) is generated by the regressive relation:

$$\lambda(n) = a\lambda(n-1) + bw(n) \qquad (III.4)$$

Where w(n) is a sequence of independent Gaussian random variables and a & b are constants and their values are determined in [70].

3.4.2 Statistical Multiplexing of Voice and Video Sources

In the case of heterogeneous sources, we defined N1 packet arrival processes for voice sources where N1 is the number of voice sources and N2 packet arrival processes for video sources where N2 is the number of video sources. Each video source generates bit rate as given in the previous section. Each voice source alternates between active and silence periods which are generated using the built-in exponentially distributed random variable generator. The voice arrival process generates packets only in the active periods where the interarrival time between the packets is the packet generation time. The mean silence period is the mean interarrival time between the active periods.

According to the access control scheme, each video/voice source generates packets with a coding rate that corresponds to 8/4 bits per sample as long as the buffer occupancy level is in the underload interval (i.e.,the number of packets in the buffer is less than the first threshold K1). In this case, the interarrival time between packets, T1 is equal to the packet length in bits divided by the source peak bit-rate in bits per second. When the buffer occupancy level is bounded between the threshold levels K1 and K2 (transitional interval from the underload to the overload period), the coding rate for video/voice drops to 6/3 bits/sample thus decreasing the source peak bit-rate correspondingly, therefore increasing the interarrival time between packets to T2. As the buffer occupancy reaches the overload interval (i.e., the number of packets in the buffer increases beyond the second threshold K2) the coding rate for video/voice drops further from 6/3 to 4/2 bits/sample hence the source peak bit-rate is reduced furthermore and the interarrival time between packets is increased to T3. The blocking probability is measured by evaluating the ratio of the number of blocked packets to the total number of transmitted packets during the simulation run time. We run several simulations with different seeds to evaluate the link performance with and without access control and the confidence interval was 95%. Some of the performance evaluation results are the blocking probability with and without access control, the optimum values of the thresholds K1 and K2 and the mean bits/sample. A comparison between statistical, deterministic and bandwidth control window allocations to voice and video traffic are presented.

3.5 Simulation Results and Discussions

In this section, we discuss the simulation results obtained from this study. For each voice source, we set the mean of the talkspurt period equals to 352 microseconds and the mean silence period equals to 650 microseconds. Assuming ADPCM coding technique for the voice sources at 4 bits/sample, then the source peak bit-rate is 32 kbps. As the first threshold level K1 is reached, the source peak bit-rate is reduced to 24 kbps. As the second threshold K2 is reached, the source peak bit-rate drops to its final overload value of 16 kbps. The video source model used in our study, generates 30 frames per second, each frame contains 512 x 512 pixels. Using first order autoregressive continuous-state, discrete-time Markov process which is given in the previous section, we generate each video source rate. The values of the constants a and b are 0.878 and 0.1108 respectively, leading to an average bit rate of 3.9 Mbits/sec and a peak bit rate of 10.6 Mbits/sec for each video source. These values are adopted from the work done [70]. The communication link capacity is assumed to be 150 Mbits/sec , and the cell length is the ATM standard of 53 bytes. The buffer length is set at 20 cells for both voice and video queues, which limits the maximum acceptable delay to 50 μ secs.

Figure 3.4 shows the statistical multiplexing gain of independent video sources without and with the access flow control. It shows the blocking probability for multiplexing one to five video sources at the same utilization value of 80%. This result indicates that the blocking probability of video sources drops dramatically as the number of multiplexed sources increases. This demonstrates that statistical multiplexing can efficiently absorb temporal variations of the bit rate of individual sources. In the case of video traffic, the access flow control is more effective than the voice traffic, (see our work in [71], therefore only one threshold for the video buffer is used. To see the effect of the threshold on the blocking probability, we set this threshold value to 10 cells. When the access flow control is applied, more dramatic reduction in blocking probability is obtained as the number of multiplexing sources increases.

Figure 3.5 shows the blocking probability for multiplexing 32 (load of 80%) to 40 (load of 100%) video sources. At 32 video sources, the blocking probability is below the required value for QOS which is 10^{-7}. As the number of video sources increases, the blocking probability increases to reach a value of 0.01 at 40 video sources. From the figure, it is clear that the access flow control is applied from 33 to 40 video sources which represent 100% utilization of the link. The access flow control activates only one threshold level for the video buffer to satisfy the required QOS which is 10^{-7} of blocking probability. The optimum values for the threshold level K are shown in Table 3.1. It is also shown the percentage of time for activation of the threshold level for each number of video sources. At 33 video sources, this percentage of time has very small value which is equal to 0.01%. As the number of video sources increases, this percentage of time increases and reaches a value of 2% at 40 video sources. It is clear from Figure 3.5 that as the access flow control is applied, all the blocking probabilities have been dropped to the required value of 10^{-7}. The price to be paid is a slight degradation in the service quality measured by a decrease in the mean bits/sample as shown in Figure 3.6. This degradation is very graceful versus the input load.

Figures 3.7 and 3.8 show the statistical multiplexing of different number of voice sources with one and three video sources respectively. As shown in these figures when the voice and video cells are treated equally, the blocking probability of video is worse than that of voice. We applied the access flow control using two threshold levels (K1, K2) for the common multiplexer buffer and we set these values of the threshold levels to 6 and 12. It is clear from Figures 3.7 and 3.8 that as the access flow control is applied, a dramatic reduction in blocking probability is obtained at each number of multiplexed voice sources. In the case of one video source, if the video blocking probability is 10^{-7} and the voice blocking probability is 10^{-4}, the number of multiplexed voice sources is 1000 sources, but after applying the access flow control this number is increased to 11300 sources. In the case of three video sources, if the video blocking probability is 10^{-7} and the voice blocking probability is 10^{-4}, the number of multiplexed voice sources is 8000 sources and becomes 8800 after applying the access flow control.

Comparisons between deterministic and statistical bandwidth allocation to voice and video traffic are shown in Figure 3.9 and 3.10 for one and three video sources respectively. For statistical bandwidth allocation in both cases, if the video blocking probability is 10^{-7} and the voice blocking probability is 10^{-4}, the number of multiplexed voice sources decreases about 10% in comparison to deterministic bandwidth allocation. Comparing the two figures, it is easy to see that if the peak bit-rate of the multiplexed video sources is the same, the number of multiplexed voice sources is insensitive to the number of multiplexed video sources.

Figure 3.11 shows the values of the total bandwidth control windows assigned to voice and video traffic for multiplexing voice sources with one and three video sources. The bandwidth control window is equal to the maximum total time spend in both buffers to support the given video and voice calls under the required QOS and is given in units of cell transmission time. As the number of voice sources increases, the total bandwidth control window decreases. In the case of three video sources, the values of the bandwidth control windows decrease in comparison to the values assigned in the case of one video source. Table 3.2shows the optimum values of the bandwidth control windows that assigned to voice and video buffers at different load values to guarantee the traffic required QOS in terms of its cell blocking probability of 10^{-7} for video and 10^{-4} for voice and attainable quality.

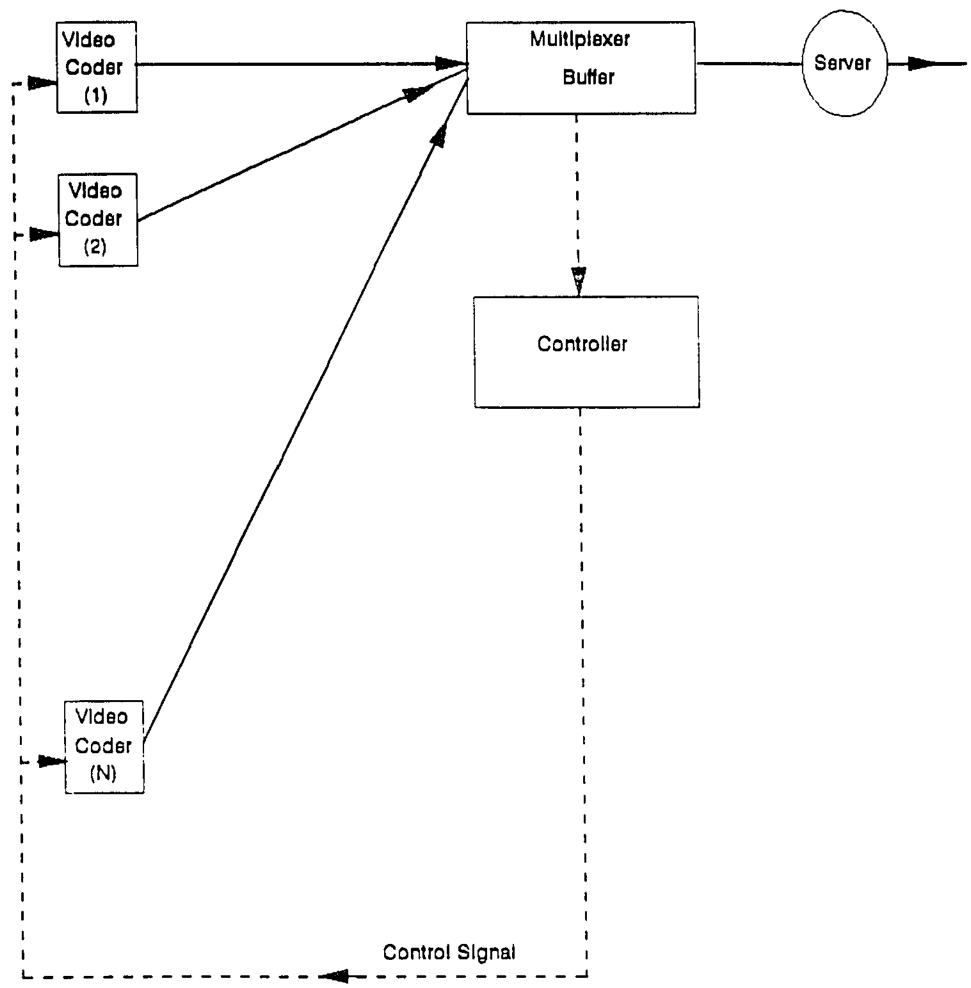


Figure 3.1 Video Multiplexer with Feedback Control

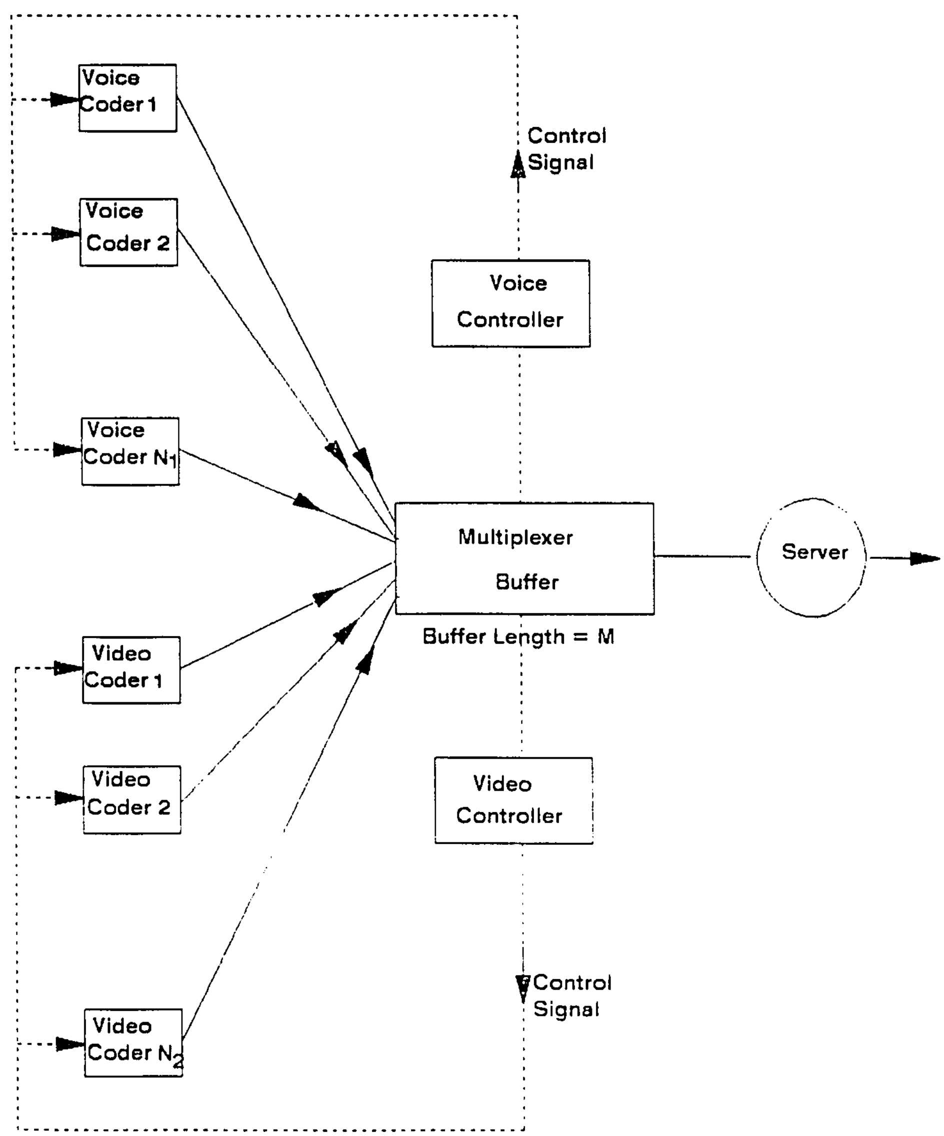
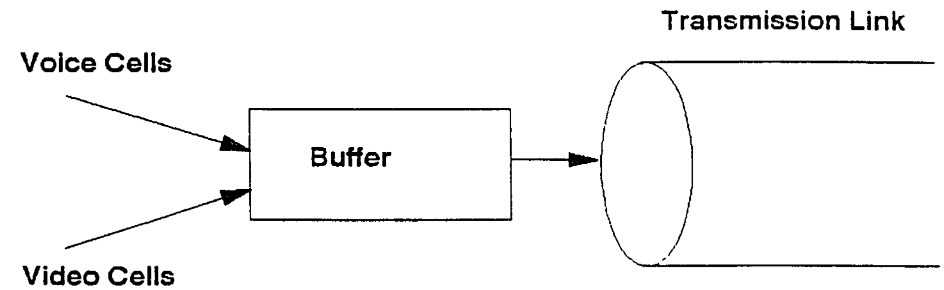
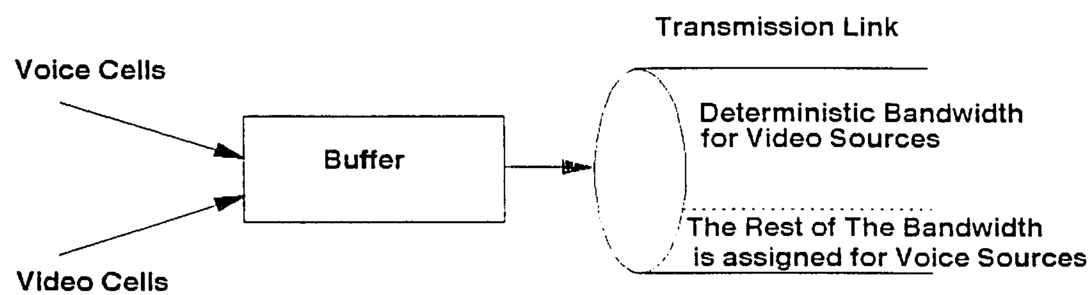


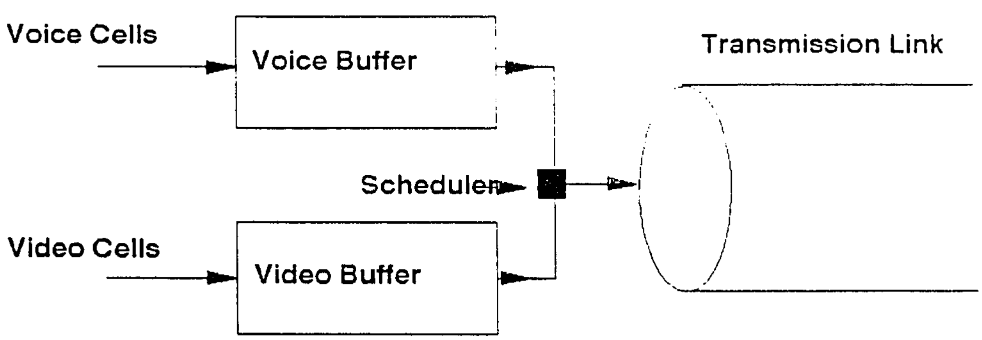
Figure 3.2 Voice and Video Multiplexer with Feedback Control



a: Statistical Bandwidth Allocation to Voice and Video Traffic



b: Deterministic Bandwidth Allocation to Voice and Video Traffic



c: Bandwidth Control Period for Voice and Video Traffic

Figure 3.3 Bandwidth Allocation to Voice and Video Traffic

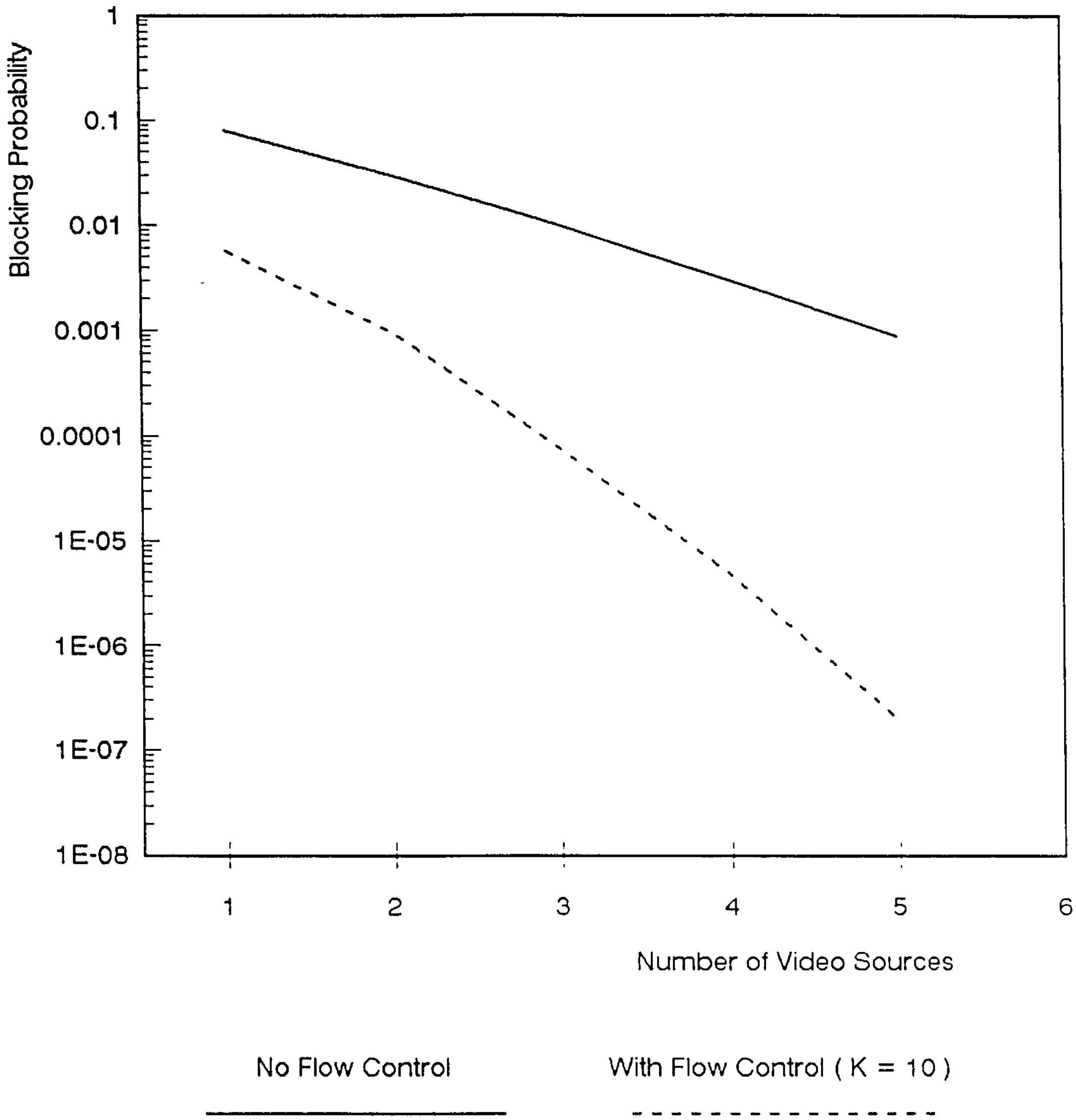


Figure 3.4 Statistical Multiplexing Gain

With and Without Flow Control

Utilization = 80 %

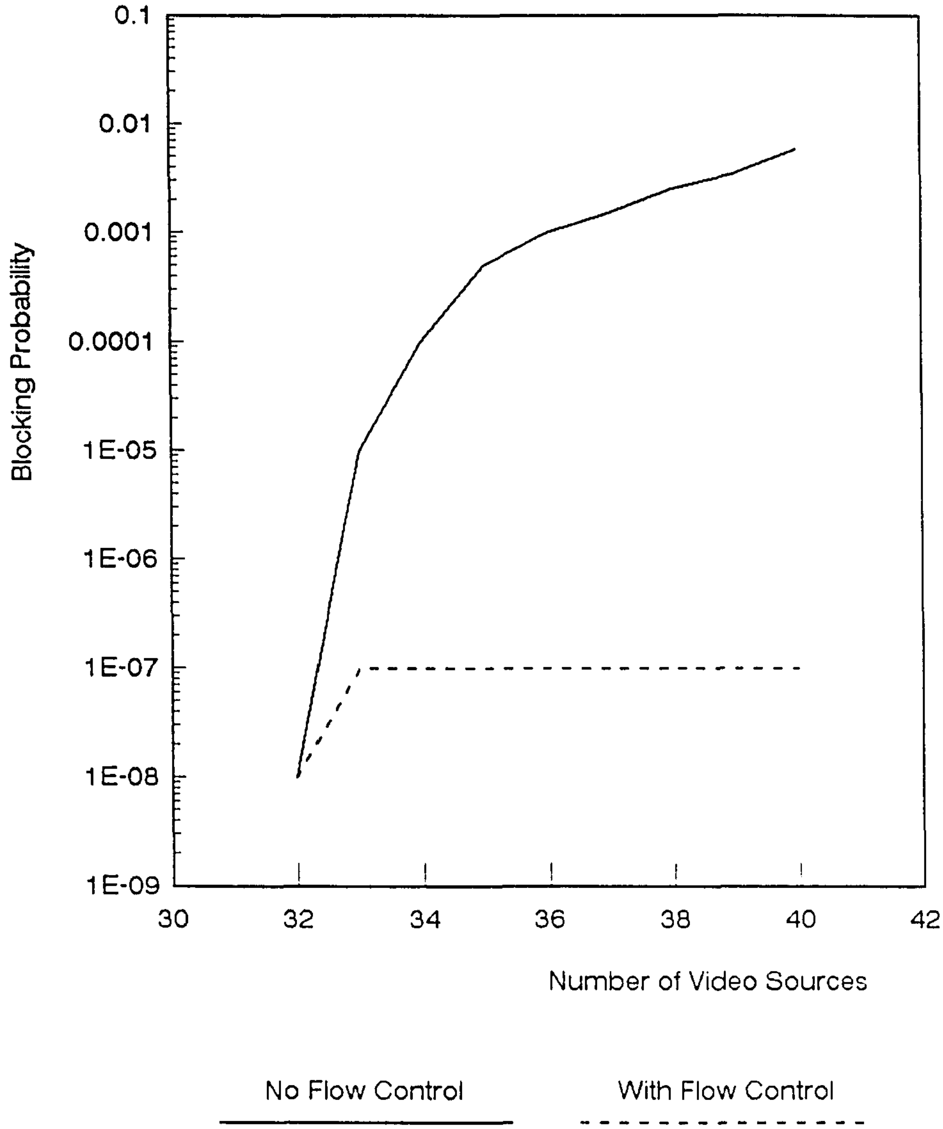


Figure 3.5 Blocking Probability Versus Number of Video Sources With and Without Flow Control

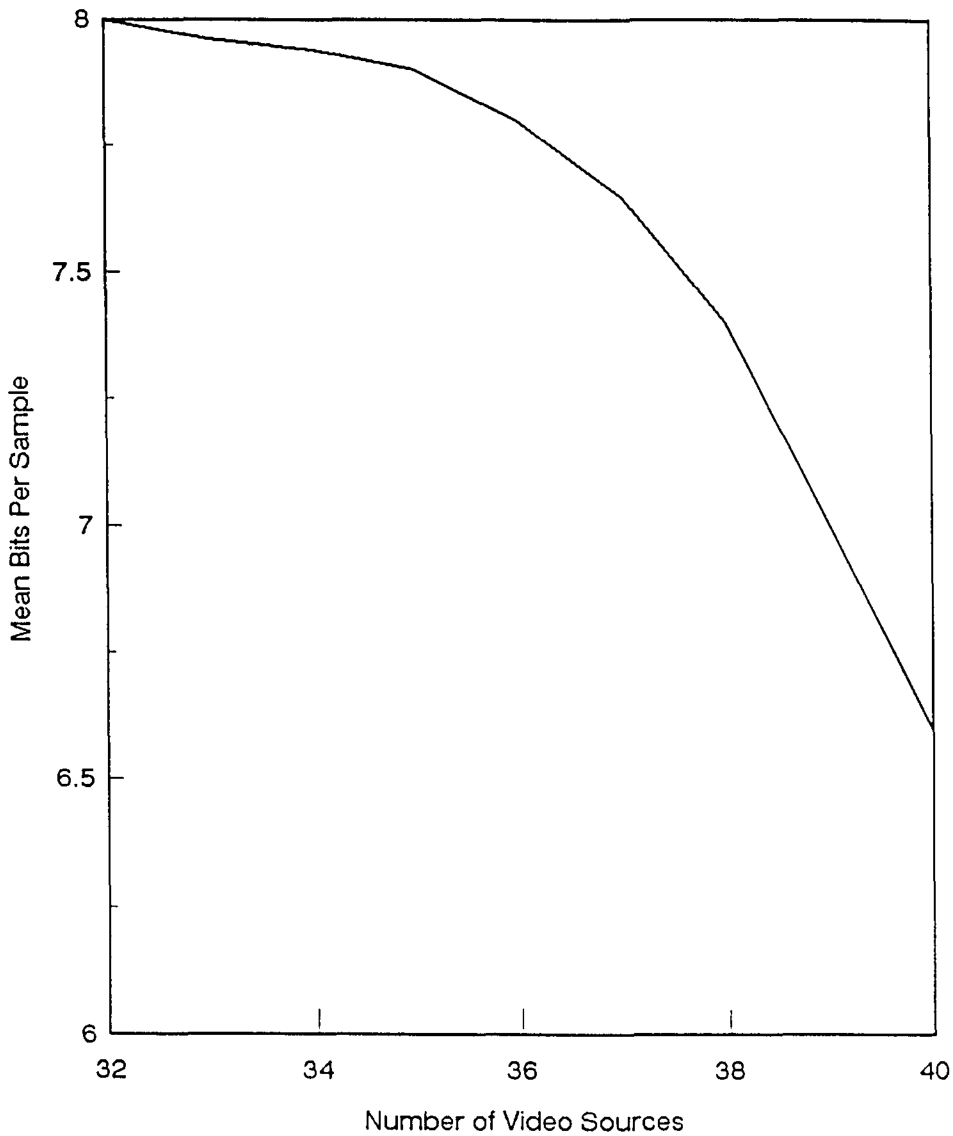


Figure 3.6 Effect of Flow Control on The Mean Bits

Per Sample

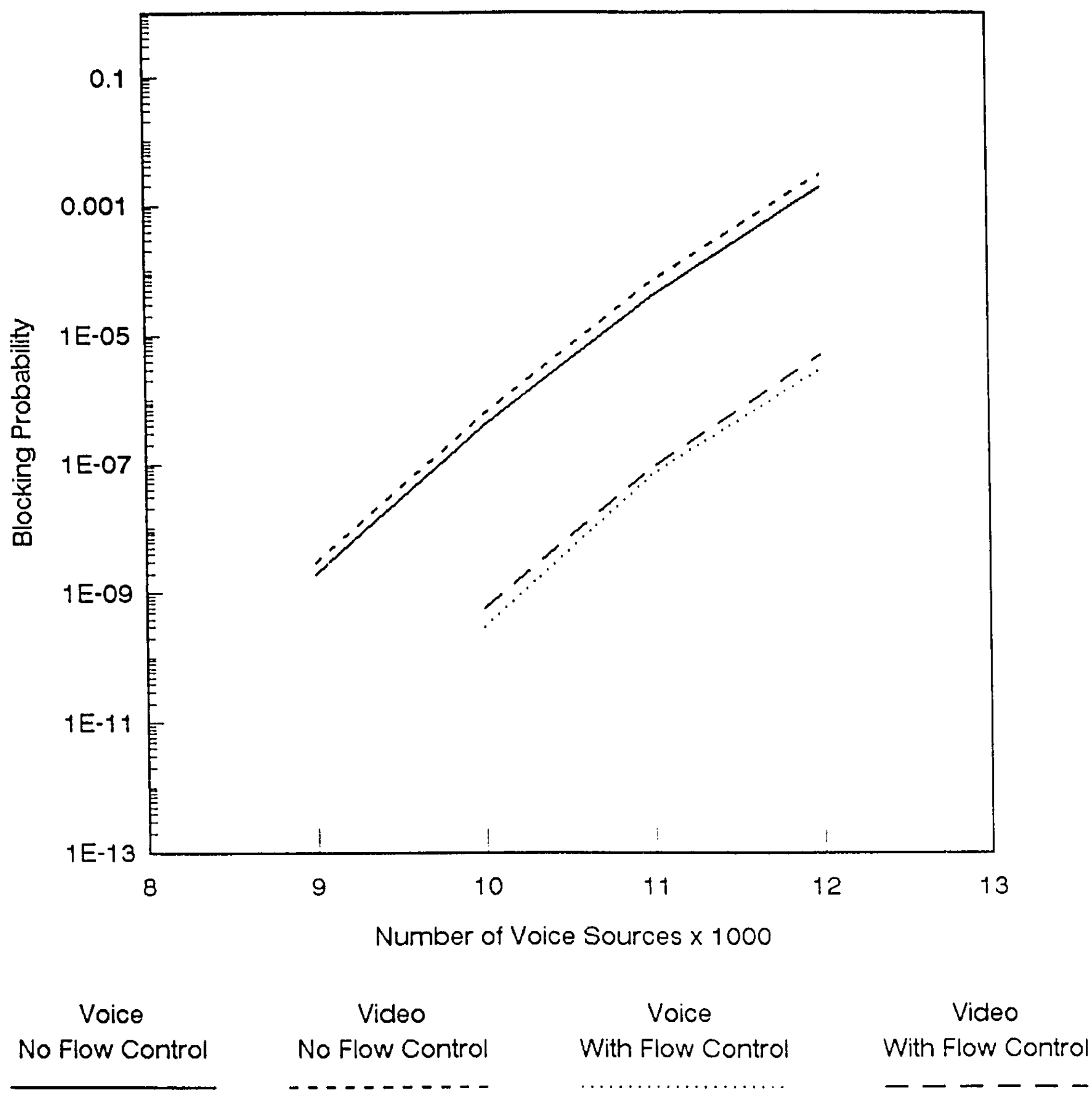


Figure 3.7 Blocking Probability Versus Number of

Voice Sources (1 Video Source)

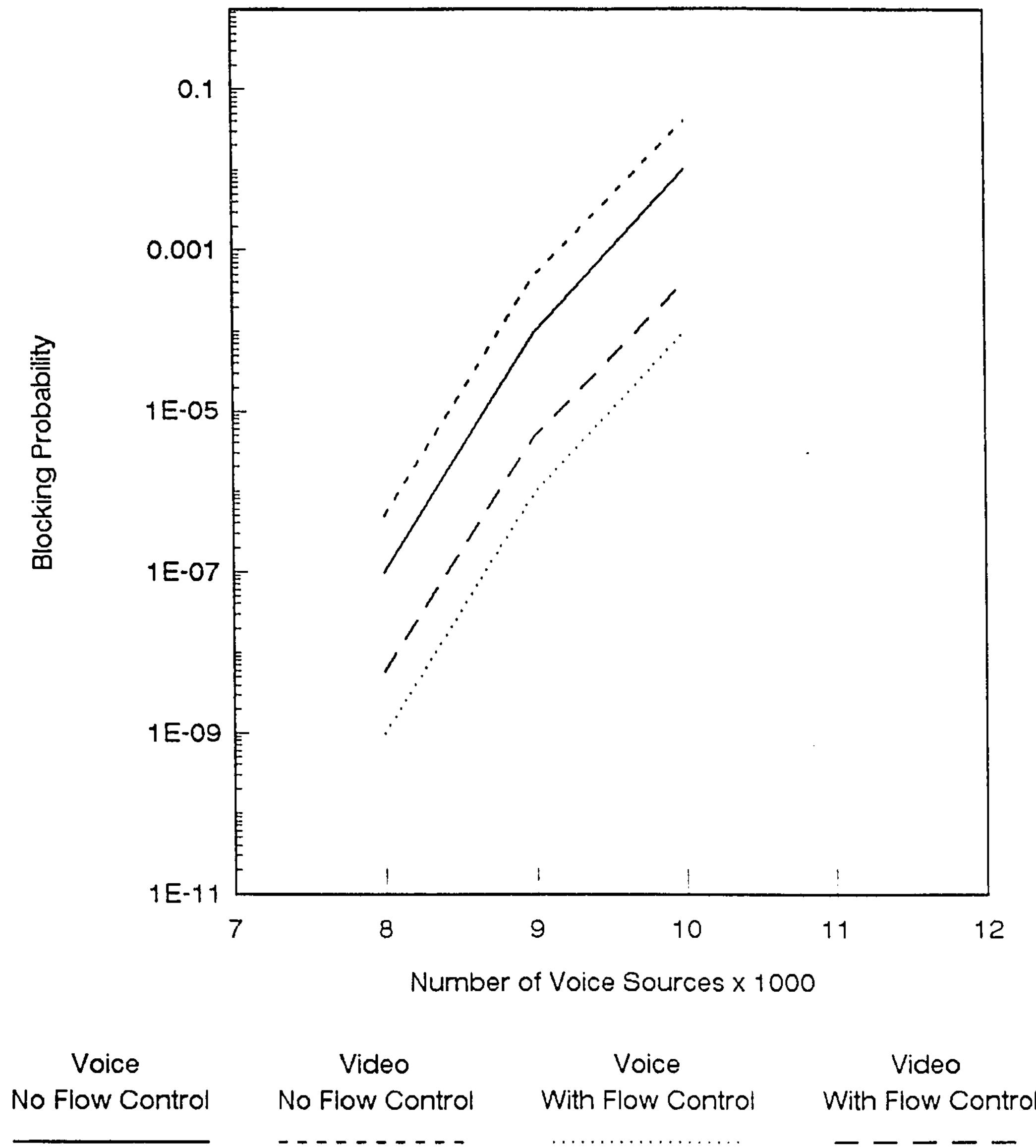


Figure 3.8 Blocking Probability Versus Number of

Voice Sources (3 Video Sources)

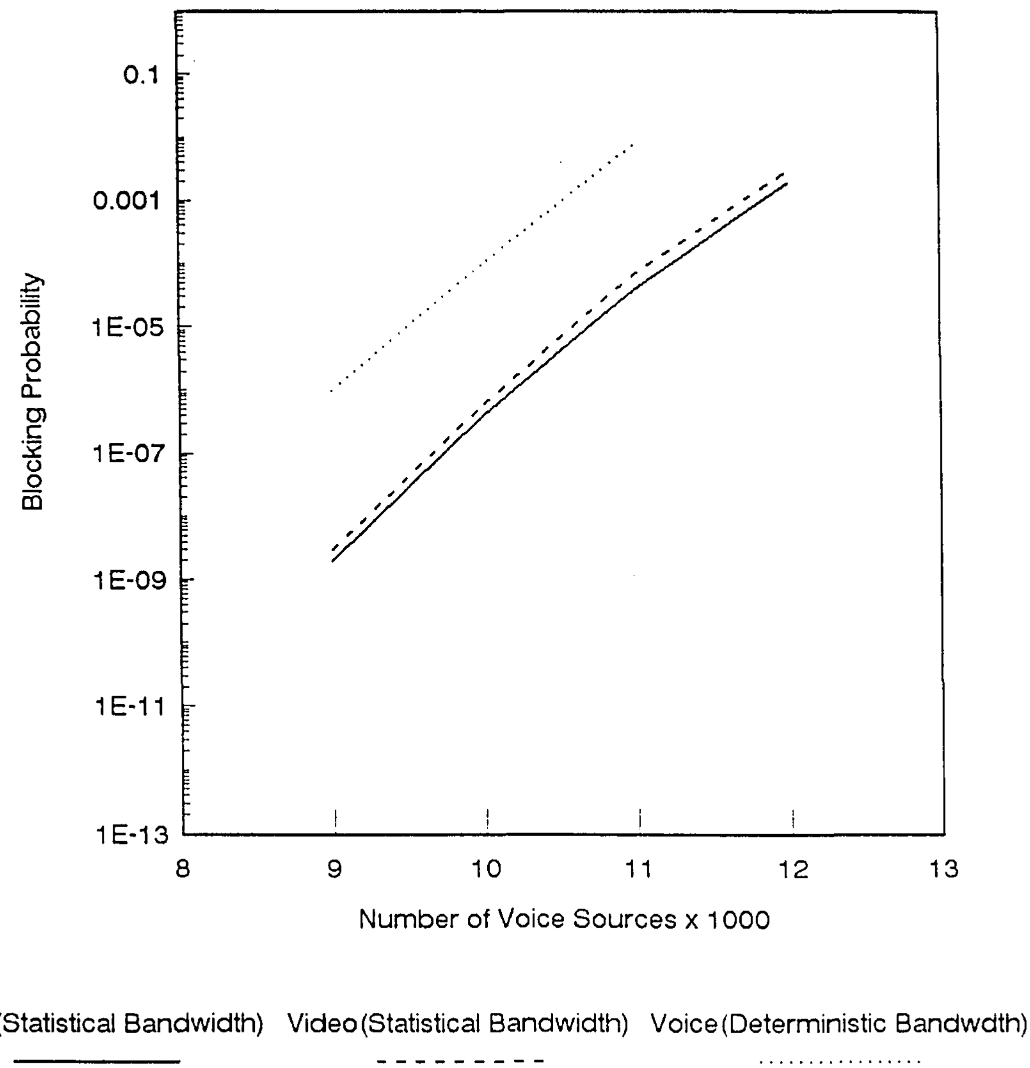


Figure 3.9 Comparison Between Deterministic and

Statistical Bandwidth Allocation

(1 Video Source)

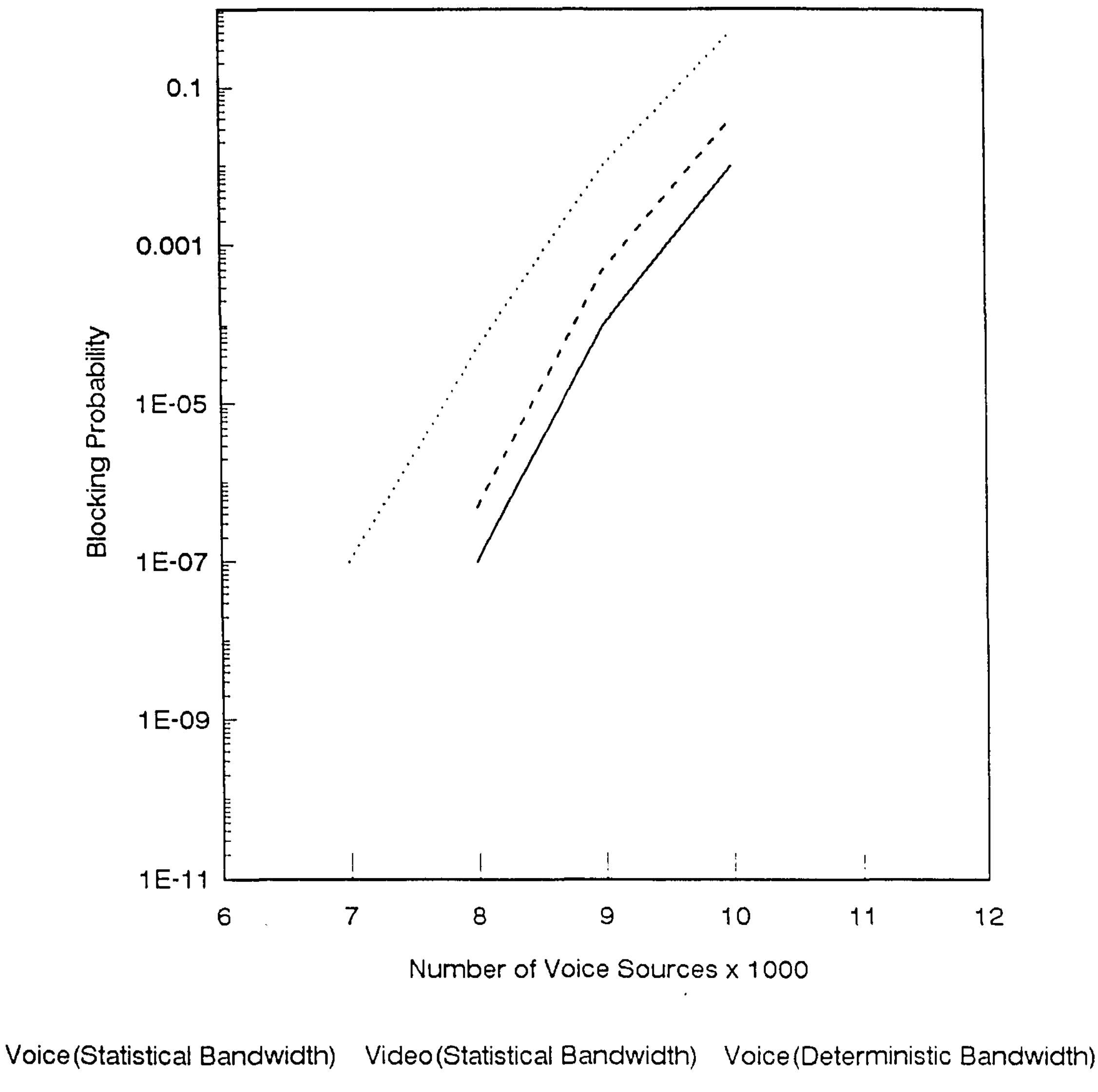


Figure 3.10 Comparison Between Deterministic and

Statistical Bandwidth Allocation

(3 Video Sources)

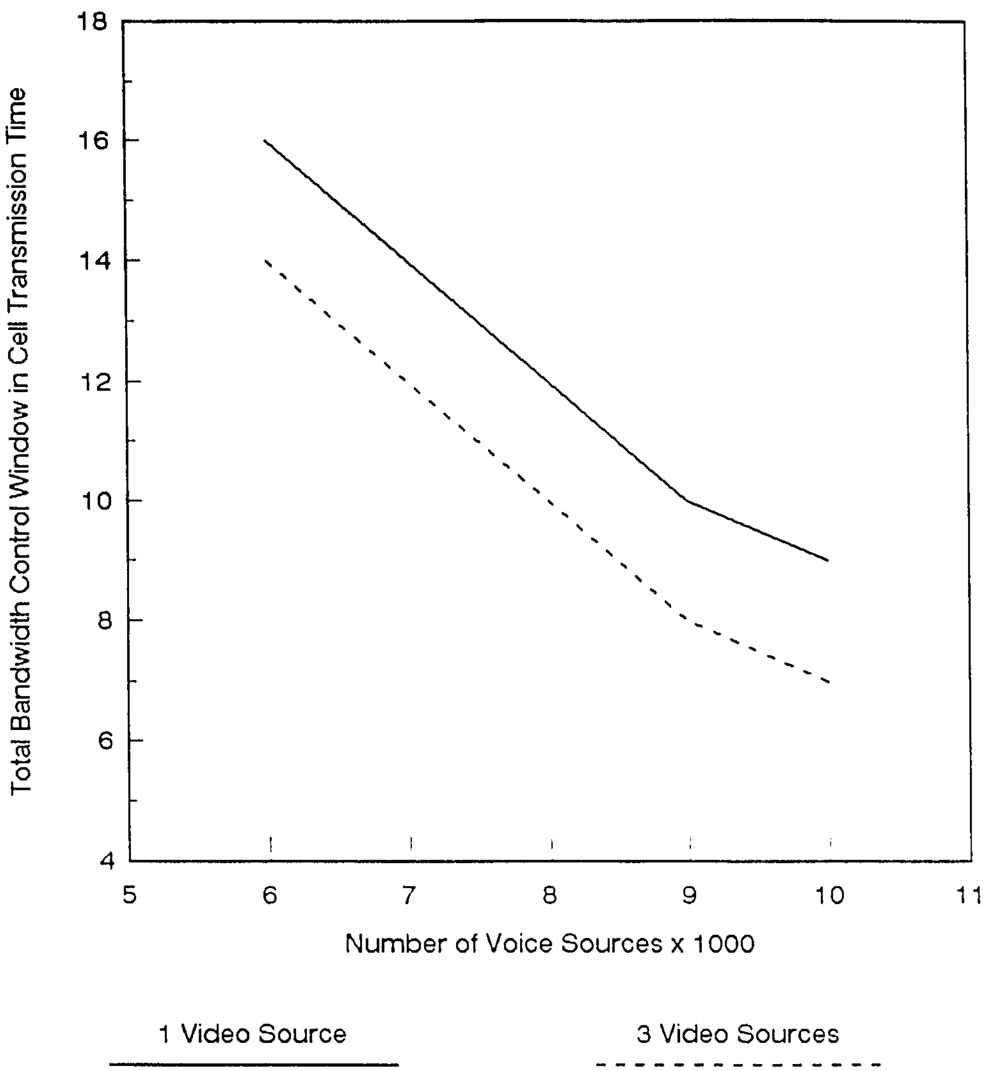


Figure 3.11 Total Bandwidth Control Window Versus Number of Voice Sources

Number of Video Sources	Value of the Threshold Level " K "	Percentage of Time for Activation of Threshold Level
33	16	0.01%
34	15	0.0123%
35	14	0.0315 %
36	13	0.26 %
37	12	0.53%
38	11	0.82%
39	10	1.12%
40	9	2%

Table 3.1 Optimum Values of The Threshold Levels to satisfy The Required Blocking Probability of 10 For Different Number of Video Sources

Number of Voice Sources	1 Video Source		3 Video Sources	
	Assigned Window To Video Buffer	Assigned Window To Voice Buffer	Assigned Window To Video Buffer	Assigned Window To Voice Buffer
6000	6	10	6	8
7000	5	9	5	7
8000	4	8	4	6
9000	3	7	3	5
10000	2	6	2	4

Table 3.2 Optimum Values of Bandwidth Control Windows for Voice and Video Buffers at Different Loads

4. Conclusions and Future Work

We have applied simulation methods to evaluate the performance of voice, video and voice/video multiplexers with an access flow control algorithm and a bandwidth control window assignment. The basic principle of the simulation is to operate simultaneously N independent voice or video packet generation and service processes. The accuracy of the results from a simulation run are a function of the length of the simulation and the correctness of the source and multiplexer models. We adopted an empirical approach to establish the accuracy of the simulation by running the simulation programs for incrementally longer periods of time until the differences in the simulation outputs of interest converged to appropriately small values. Because the aggregate process resulting from multiplexing homogeneous and heterogeneous traffic is very complex, we used simulation analysis to gain some sight of the multiplexer behavior. Also we used simulation to obtain the distributions (histograms) of the multiplexer buffer length at different load which are difficult to obtain analytically and we obtained the all optimum values of the threshold levels of the multiplexer buffer and the optimum values of the control windows for the access control scheme and bandwidth allocation algorithm.

We summarize our contributions in the following points:

1) The optimum sets of the thresholds values (K1, K2) were obtained to satisfy the QOS for voice and video traffic in terms of its blocking probabilities and mean bit/sample at given buffer size.

2) The access control scheme with different access control windows are presented and the optimum sets of (K1, K2) for different access control windows were obtained. Also, we show the effect of the access control window on the triggering rate of the controller.

3) Using the bandwidth control window assignment, a bandwidth can be allocated to voice and video traffic such that to guarantee the traffic required QOS in terms of its cell blocking probability and attainable quality. The optimum values for the bandwidth control windows, which were assigned to voice and video buffers at different loads, were obtained.

The problem of flow control in ATM networks needs more work to achieve a novel scheme. Therefore, our future work will include further studies in flow control algorithms and new models for real time traffic. Also, we will study dynamic routing and bandwidth allocation of real time traffic which are issues need to be resolved.

5. References

[1]. R. Handel, "Evolution of ISDN towards broadband ISDN," IEEE Net., pp. 7-13, 1989.

[2]. M.Kawarazaki, T. Okada, and M. Sasakawa, "Perspective of ATM network," J. IEICE, vol. 71, no. 8, pp. 806-814, 1988.

[3]. J. S. Turner, "New directions in communications (or which way to the information age?)," IEEE Commun. Mag., vol. 24, no. 10, pp. 8-15, 1986.

[4]. H. Saito, and K. Shiomoto, "Dynamic Call Admission Control in ATM Networks" IEEE Journal on Selected Areas in Comm., vol. 9, no. 7, Sept 91.

[5]. S-Qi. Li, J. Mark "Traffic Characterization for Integrated Services Networks" IEEE Trans. on Commun., vol. com-28, pp.915-925, Aug. 1990.

[6]. D. Minoli, "Issues in packet voice communication," Proc. IEEE, vol. 126, pp. 729-740, 1979.

[7]. "General geometric arrival-constant server queue with application to packetized voice," in Proc. Conf. Rec. Int. Conf. Com., vol. 3, pp. 36.6.1-5, 1978.

[8]. G. I. Coviello, O. l. Lake, and G. R. Redinbo, "System design implications of packetized voice," in Proc. Conf. Rec. Int. Comput., 1977, pp. 38.3.49-53.

[9]. G. J. Coviello, "Comparative discussion of circuit versus packet-switched voice," IEEE Trns. Commun., vol. COM-27, pp. 1153-1160, 1979.

[10]. J. Seguel, Y. Tanaka, and M. Akiyama, "Simulation analysis of the waiting time distribution of a packetized voice concentrator," Trans. IEEE Japan, vol. E65, pp. 115-122, 1982.

[11]. Y. C. Jeng, "Approximation for packetized voice traffic in statistical multiplexer," in Proc. IEEE INFOCOM, Apr. 1984, pp. 256-259.

[12]. K. Sriram and W. whitt, "Characterizing superposition arrival processes in packet multiplexers for voice and data," IEEE J. Select. Areas Commun. vol. SAC-4, pp. 833-846, 1986.

[13]. H. Heffes and D. M. Lucantoni, "A Markov modulated characterization of packetized voice and data traffic and related statistical multiplexer performance," IEEE J. Select. Areas. Commun. vol. SAC-4, pp. 856-868, 1986.

[14]. N. Janakiraman, B. Pagurek, and J. E. Neilson, "Multiplexing low-speed buffered data terminals," IEEE Trans. Commun., vol. COM-28, pp. 1838-1843, 1980.

[15]. Y. Tanaka, J. Seguel, and M. Akjyama,"Performance analysis of a voice packet multiplexer," Trans. IEEE Japan, vol. J65-B, pp. 539-546, 1982.

[16]. T. E. Stern, "A queueing analysis of packet voice,"in Proc. GLOBECOM'83, vol. 1, pp. 71-76, 1983.

[17]. J. N. Daigle and J. S. Langford,"Models for analysis of packet voice communcations systems,"IEEE J. Select. Areas Commun., vol. SAC-4, pp. 847-855, 1986.

[18].D. Anick, D. Mitra, and M. M. Sondhi,"Stochastic theory of a data-handling system with multiple sources," Bell Syst. Tech. J., vol. 61, pp. 1871-1894, 1982.

[19]. C. J. Weinstein, "Fractional speech loss and talker activity model for TASI and for packet-switched speech," IEEE TRANS. Commun., vol. COM-26, pp. 1253-1257, 1978.

[20]. D. Sparrell, "Wideband packet technology," in Proc. IEEE GLOBECOM'88, Hollywood, FL, Nov. 1988, pp. 1612-1615.

[21]. W. J. Giguere. "New applications of wideband technology," in Proc. IEEE Supercomm/ICC'90, Atlanta, GA, Apr. 1990, vol.3, pp. 997-999.

[22]. N. Yin, S. Q. Li, and T. E. Stern, "Congestion control for packet voice by selective packet discarding," in Proc. GLOBECOM'87, Tokyo, Japan, Nov. 1987, vol. 3, pp. 1782-1786.

[23]. D. W. Petr, L. A. DaSilva, Jr., and V. S. Frost, "Priority discarding of speech in integrated packet networks," IEEE J. Select. Areas Commun., pp. 644-656, June 1989.

[24]. R. Guerin, H. Ahmadi, and M. Naghshineh,"Equivalent capacity and its application to bandwidth allocation in high-speed networks," IEEE J. Select. Areas Commun., vol. 9, No. 7, Sept. 1991.

[25]. K. Sriram, "Dynamic bandwidth allocation and congestion control schemes for voice and data multiplexing in wideband packet technology," Proc. ICC'90., vol.2, pp.201-209, 1990.

[26]. G. Gallassi, G. Rigolio, L. Fratta, "ATM: Bandwidth assignment and bandwidth enforcement policies," Proc. GLOBECOM'89, Dallas, vol.1, pp. 61-68, 1989.

[27]. M. Decina, T. Toniatti, P. Vaccari, L. Verri "Bandwidth Assignment and Virtual Call Blocking in ATM Networks" Proc. INFOCOM 90, vol.3, pp.210-216, 1990

[28]. W. Wang, T. Saadawi, K. Aihara "Bandwidth Variation and Control for ATM Networks" Proc. ICC 90, vol.2, pp.119-126, 1990.

[29]. S. Ohta, K. Sato and I. Tokizawa, "A Dynamically Controllable ATM Transport Network Based on the virtual Path Concept",Proc.INFOCOM, 88,vol.2,pp.185-195,december 1988.

[30]. I. Tokizawa, T. Kanada and K. Sato,"A New Transport Network Architecture Based on Asynchronous Transfer Mode Technique", ISSLS,88,vol.1,pp.51-60, 1988.

[31]. B. Amin-Salehi, D. Spears "Support of Transport Services in BISDN" Proc. GLOBECOM 89, vol.2,pp.134-141, 1989.

[32]. H. Ishii, M. Kawarasaki "BISDN Signaling Protocol Capabilities" Proc. GLOBECOM 89, vol.2,pp.160-168, 1989.

[33]. J. P. Coudreuse "ATM Status of Definition and Discussion of Some Open Issues" Proc. Multimedia 89, vol.1 pp.31-39, 1989.

[34]. S. Yoneda "Broadband ISDN ATM Layer Management: Operations, Administration, and Maintenance Consideration" IEEE Net. Mag.,vol.14,no.10pp.10-16,1990.

[35]. J. Kulzer and W. Montgomery, "Statistical Switching Architecture for Future Services," Proc. ISS 84, pp. 43A.1.1-43A.1.6,1984.

[36]. T. Y. Choi, "Statistical Multiplexing of Bursty Sources in an ATM Network," Multimedia 89, vol.1,pp. 52-60, 1989.

[37]. C. Anthony Cooper and Kun I. Park, "Toward a Broadband Congestion Control Strategy" IEEE Net. Mag., vol. 14, no.10, pp.28-36, 1990.

[38]. T. Kamitake and T. Suda, "Evaluation of an Admission Control Scheme for an ATM Network Considering Fluctuations in Cell Loss Rate," Proc. IEEE GLOBCOM 89, pp.49.4.1-49.4.7,1989.

[39]. Murata, Y. Oie, T. Suda and H. Miyahara, "Analysis of a Discrete-Time Single-Server Queue with Bursty Inputs for Traffic Control in ATM Networks," Proc. IEEE GLOBECOM 89, pp.49.5.1-49.5.7,1989.

[40]. S-Q. Li, "Study of Information Loss in Packet Voice Systems," IEEE Tras. Commun.,vol.37,No.11, pp.1192-1202, November 89.

[41]. I. Ide, "Superposition of Interrupted Poisson Processes and Its Application to Packetized Voice Multiplexers," Proc. 12th Intl. Teletraffic Congress, Torino, Italy, vol.1,pp.61-69, 1988.

[42]. K. Q. Liao and L. G. Mason, "A Discrete-Time Single Server Queue with a Two-level Modulated Input and Its Applications," Proc. IEEE GLOBECOM 89, pp.26.1.1-26.1.6,1989.

[43]. P. Sen, B. Maglaris, N. E. Rikli and D. Anastassiou, "Models for Packet Switching of Variable-Bit-Rate Video Sources," IEEE J. Select. Areas Commun., vol.7, No.5, pp.865-869, June 1989.

[44]. M. Nomura, T. Fujii and N. Ohta, "Basic Characteristics of Variable Rate Video Coding in ATM Environment," IEEE J. Select. Areas Commun., Vol.7, No.5, pp.752-760, June 1989.

[45]. B. Maglaris, D. Anastassiou, P. Sen, G. Karlsson and J. D. Robbins, "Performance Models of Statistical Multiplexing in Packet Video Communications," IEEE Trans. Commun., Vol.36, No.7, pp.834-844, July 1988.

[46]. A. Papoulis, Probability, Random Variables, and Stochastic Processes, New York: Mc-Graw Hill, pp. 62-128, 1984.

[47]. Y. Yasuda, H. yasuda, N. Ohta and F. Kishino, "Packet Video Transmission Through ATM Networks," Proc. IEEE GLOBECOM 89, pp.25.1.1-25.1.5,1989.

[48]. H. Yamada, K. Miyake, F. Kishino and K. Manabe, "Modeling of Arrival Process of Packetized Video and Related Statistical Multiplexer Performance," Proc. IECEJ National Conference, vol.1, pp.21-29, 1989.

[49]. L. Dittmann and S. B. Jacobsen, "Statistical Multiplexing of Identical Bursty Sources in an ATM Network," Proc. IEEE GLOBECOM 88, pp.39.6.1-39.6.5,1988.

[50]. M. Hirano and N. Watanabe, "Characteristics of a Cell Multiplexer for Bursty ATM Traffic," Proc. IEEE ICC 89, pp.13.2.1-13.2.5,1989.

[51]. S. Akhtar, "Congestion Control in a Fast Packet Switching Network," Master's Thesis, Washington University, December 1987.

[52]. S. B. Jacobsen, K. Moth, L. Dittmann and K. Sallberg, "Load Control in ATM Networks," Proc.ISS 90, vol.2, pp.141-149, 1990.

[53]. G. Ramamurthy and R. S. Dighe, "Distributed Source Control: A Network Access Control for Integrated Broadband Packet Networks," IEEE J. Select. Areas Commun. Vol.9, No.7, September 1991.

[54]. A. Gersht and K. J. Lee, "A Congestion Control Framework for ATM Networks," Proc. IEEE INFOCOM 89, pp.701-710, 1989.

[55]. G. M. Woodruff, R. G. Rogers and P. S. Richards, "A Congestion Control Framework for High-speed Integrated Packetized Transport," Proc. IEEE GLOBECOM 88, pp.7.1.1-7.1.5,1988.

[56]. J. M. HOLTzman,"Coping with Broadband Traffic Uncertainties: Statistical Uncertainty, Fuzziness, Neural Networks,"presented at IEEE Workshop on Computer Commun., Data Point, CA, October 1989.

[57]. P. T. Brady, "A model for Generating ON-OFF Speech Patterns in Two-Way Conversation." Bell Syst.Tech.Journal Vol 48,pp.314-322, Sept.1969.

[58]. M. Nomuror, T. Fuju, N. Ohta, "Basic Characteristics of VBR Video Coding in ATM Envoirment" IEEE J.Sec. Area Comm, vol.11,no.2, June 1989.

[59]. J. P. Coudreuse, et al, "ATD Techniques: An experimental Packet Network Integrating Videocommunication" Proc. ISS'84, vol.1, pp.50-59, 1984.

[60]. J. S. Turner, L. F. Wyatt "Apacket Network Architecture for Integrated Services" Proc.GLOBCOMM 83, vol.2, pp.112-120, 1983.

[61]. M. De Prycker "Definition of Network Options for Belgian ATM Broadband Experiment" IEEE J. Sec. Areas. Comm., vol.1, no.3, Dec.1988.

[62]. J. Hui, "Network, Transport and Switching Integration for Broadband Commu-nication "IEEE Network, vol.14, no.10, pp.21-30, March 1989.

[63]. R. Ballart, Y. Ching, "SONET. The Standard Optical Network," IEEE Communication Magazine, vol.24, no.10, pp.18-27, 1989.

[64]. J. Gruber, et al, "Improvement in Availability and Error Performance of SONET compared to Asynchronous Transport Systems,"Proc., ICC 90, vol.2, pp.115-122, 1990.

[65]. J. Y. Hui, "Resource Allocation for Broadband Networks," IEEE J.Sec.Area Comm., vol. SAC-4, pp.240-251, 1988.

[66]. K. Sririam, D. Lucantoni, "Traffic Smoothing Effects of Bit Dropping in Packet Voice Multiplexers," IEEE Trans.Comm., vol. 28, pp.312-321, 1989.

[67]. I. Habib, T. Saadawi, "Access Flow Control Algorithms in Broadband Networks" Journal of Computer Communication, vol.4, pp.310-321, 1992.

[68]. I. Habib, T. Saadawi, "Flow Control Techniques in Packet Voice Multiplexers" Proc. ISMM International Conference on Parallel and Distributed Computing, vol.1, pp.52-59, 1988.

[69]. I. Habib, T. Saadawi "Congestion Control in Video Packet Multiplexers" Proc.3rd IFIP Conference on High Speed Networks, Berlin, March 1991, Also appeared in "High Speed Networking," edited by O. Spaniol, North Holland 1991.

[70]. I. Habib, T. Saadawi, "Dynamic Bandwidth Allocation and Congestion Control

of Virtual Paths in ATM Networks" Proc., INFOCOMM'92, vol.2, pp.305-315, 1992.

[71]. M. Saleh, I. Habib, and T. Saadawi, "Simulation analysis of a communication link with statistically multiplexed bursty voice sources," IEEE J. Select. Areas in Commun., vol. 11, No. 3, Apr. 1993.