Order Number 9207078

Performance analysis of flow control algorithms in asynchronous transport mode broadband networks

Habib, Ibrahim Wahby, Ph.D.

City University of New York, 1991

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

A

PERFORMANCE ANALYSIS OF FLOW CONTROL ALGORITHMS IN ASYNCHRONOUS TRANSPORT MODE BROADBAND NETWORKS

BY

IBRAHIM WAHBY HABIB

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

1991

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

8/29/91
Date

Chair of Examining Committee

8/29/91
Date

Executive Officer

Prof. S. Ahmed

Prof. J. Barbra

Prof. M. Lee

Dr. K. Sohraby

Supervisory Committee

The City University of New York

ABSTRACT

PERFORMANCE ANALYSIS OF FLOW CONTROL ALGORITHMS IN ASYNCHRONOUS TRANSPORT MODE BROADBAND NETWORKS

by

Ibrahim Wahby Habib

Adviser: Professor Tarek N. Saadawi

Asynchronous Transport Mode (ATM) Broadband Networks are designed to carry a very diverse mixture of traffic with different burstiness and correlation parameters. In this dissertation, we propose and analyze an access flow control scheme which throttles peak bit rate of the arrival process at the User Network Interface (UNI). It uses a two threshold control levels (K1,K2) to minimize the buffer occupancy level. We, also, propose and analyze a dynamic bandwidth allocation scheme based upon the virtual path principle. The bandwidth is statistically assigned to each virtual path, according to the declared traffic characteristics and required class of service. To dynamically control the allocated bandwidth, a Bandwidth Control Period (BCP) rule is proposed to control the scheduling of different classes of traffic. The multiplexer queueing performance has significantly improved, when the proposed flow control schemes are applied.

ACKNOWLEDGEMENTS

I must begin my acknowledgements by a dedication. It goes to those whom without their relentless support, persistant encouragement and perpetual motivation, I could not have concluded this dissertation. To my parents, I dedicate my dissertation, for there are no words that can express my foremost gratitude and indebtedness to them.

To my Ph.D. advisor, Professor Tarek Saadawi, go my sincere gratefulness for his continuous help and guidance throughout the duration of my studies at the City College.

My deepest thanks to the faculty members of my supervisory committee, for their remarkable efforts in bringing my dissertation to its final form.

Many friends and colleagues have been very helpful and cooperative. My deepest thanks to all of them. Finally, I thank all my colleagues at the Computer Networks Laboratory for making my stay there, pleasant and joyful.

Table of Contents

List of Figures

I.INTRODUCTION

I.1. Background

Recent advances in fiber optics communications, switching and buffering technologies, voice and video sources coding have made integrated networks, also known as Broadband ISDN, a possible reality. In such networks, we transmit multimedia information using the same transmission links and switching fabrics. Present day technology, uses separate circuit and packet switching networks to accommodate heterogeneous traffic sources. Video and voice information are transmitted via separate circuit switching networks, whereas data file transmission is accommodated over the packet switching network.

Synchronous time division multiplexing, currently employed in circuit switching telephone networks, are not suitable to support the diverse mixture of multimedia services with widely different bandwidth requirements. Packet switching networks, which allow high delay and use complicated protocol structures, are not useful with this new environment. Thus, we have to design new transport techniques that are simple, flexible and

efficient in order to support the different traffic mixture over the same integrated network. One straight forward solution is to combine the advantages of both circuit and packet switching into one single technique. But before we start addressing the details of such possible new transport technique, we will look into the objectives of integrated networks [1]-[5].

The objectives of integrated network can be summarized in the following points:

1. The network must be able to provide the users with reliable and successful means of communications. It must ensure each user that his required quality of service is met without affecting other users.
2. It must be flexible enough, to support new services and demands. The protocols functionality must be transparent to the introduction of new traffic sources which are unforeseen at the present time. This leads to an important principle of separating network transmission functions from users to network interactive detailed protocols.
3. It must efficiently utilize its resources while providing the different users with their required grade of service. It must be able to dynamically allocate its resources, upon demand, while being fair to all users. It must protect its resources

from being monopolized by a single or group of users. Thus the network must be able to monitor, on a real time basis, the traffic status on its links and switching fabrics and accordingly allocate resources, accept new calls or re-route traffic.

4. The network uses integrated access, transport and switching, to support the heterogeneous traffic. Thus it must employ a simple transport protocol in order to support real time traffic which has very stringent delay requirements. The transport protocol must be free from excess processing overhead per packet and eliminate the needs for transit nodes processing functions such as link by link flow control.

5. Full connectivity is required, in order to support various modes of operations. Multipoints communication, broadcasting, should be supported without the need of creating replicates of the transmitted information that may cause overload and introduce congestion problems.

The multimedia information, expected to be carried by these networks, covers a wide spectrum of traffic characteristics, ranging from low bit rates data to broadband bit rates suitable for video transmission. Not only does the multimedia information differ in their transmitted bit rates, but also in their

traffic shape and characteristics. Traffic sources differ, also, in their required services and demands from the network. Traffic sources, expected to be supported, cover a wide spectrum from low speed bulk data transfer to interactive high speed data, from low speed telephone transmission to high speed high quality interactive audio distribution, from image transfer and video teleconference to video broadcasting and high definition T.V. We can categorize these traffic sources into separate classes according to their declared statistics and expected quality of service from the network. Thus, a certain class of service is a set of performance measures, provided by the network, to the traffic sources sharing that specific class of service. Data transfer traffic, for example, are loss sensitive and can tolerate delay whereas real time traffic, such as, video and voice applications, are less prone to packet losses but have stringent delay and delay variability requirements [6]-[8].

I.2. ATM Networks: Concepts and Principles

After this brief discussion, we can now realize the need for a new transport technique. Two similar alternatives have

emerged, both provide the same functions yet they differ in some aspects, one possible solution is Asynchronous Transfer Mode (ATM), the other is fast packet switching (FPS). Both terms are widely used as synonyms, since they are very similar. ATM stems from circuit switching, it provides us with the advantages of two different types of digital multiplexing, one is the packet multiplexing and the other is synchronous time division multiplexing (STDM) (fig.I.1). ATM is similar to the STDM in the sense that it uses slotted time format, however, it does not allocate time slots on a fixed per call basis, but rather on a dynamic basis. It does not identify calls by their position but by a label, similar to packet multiplexing, which identifies a logical connection and is called virtual channel identifier (VCI). It does not use circuits as transport units but fixed short labelled packets called cells. Cells belonging to different calls are, thus, statistically multiplexed and identified according to their VCI's.

ATM combines the advantageous of packet switching and the Time Division Multiplexing (TDM) techniques. Each of them has been modified to accommodate all types of traffic (e.g., data, fixed bit rate voice, variable bit rate video, etc..). For what concerns packet switching, the following modifications

apply:

1. No Error or Flow Control on the links inside the ATM network.
2. Connection oriented at the lowest level. All information is transferred in a virtual circuit assigned for the total duration of the connection.
3. Packets are fixed size length (called cells) and are short (53 bytes). This choice of small size cells allow for the use of very high speed switching nodes and puts no constraints to services, since large information entities will be chopped into cells.
4. Limited functionality in the header of the cells. The sole functionality supported by the header of the ATM cells is the identification and characterization of the virtual circuits. In addition some error detection and correction on the virtual circuit identifier is provided.

For what concerns TDM, the time slotted operation is kept but the time transparency is shifted to the edges of the network. This means that no time relation is maintained inside the network: time slots are no longer characterized by their position in a frame as it was the case in TDM (this is the meaning of "A" in ATM). That explains the need to identify a certain time slot by having an additional field called header

containing a virtual circuit identifier. Finally, ATM uses out of band signalling for transport of control information used to set-up and release connections. This is contrary to the concept of in band signalling adopted in the OSI protocols, or the X.25 protocol where control and data packets are mixed together.

Network resources, such as bandwidth, are allocated to each call at the call set up time and controlled over the logical connection during the call duration. Resource control in ATM is, therefore, based upon end to end connection oriented type rather than connectionless one (although ATM can also support connectionless services through an overlaying adaptation layer). In short, ATM is the combination of asynchronous time division multiplexing (ATDM) and call connection control [1],[4],[5]. The principle of connection control, has motivated the introduction of virtual path identifier (VPI), where bundles of logical channels are aggregated and treated as a single entity by higher layers, for crossconnection functions. These functions include dynamic routing and bandwidth allocation. The VPI concept has made the network reconfiguration much easier to implement. Also, the VPI has made it possible to direct multiplex traffic with different bit rates, without the

need for complex multiplexing stages which are needed in STM networks. The VPI concept will be explored, in detail, in chapter II.

I.3. ATM Protocol Reference Model

Fast packet switching [2], evolves from packet switching and is based upon reducing the packet overhead and the transit nodes processing functions. It does that by eliminating the link by link flow control and simplifying the protocol stack. It relies on the fast switching of packets in the transit nodes and processing protocol functions in hardwire. Most of the packet switching protocol functionality are done on an end to end basis, leaving only simple functions to be performed within the network. This concept leads to the important design principle of decoupling the connection control oriented functions into two parts, off-line functions and real time functions [5],[9]-[11]. Off line functions can be implemented using universal signaling protocols (out-band signaling), whereas real time functions are done by in-band signaling in the form of a header attached to the information cell. To conclude, we can summarize the differences between the two

approaches as follows:

1. In ATM, the label functionality is reduced to an absolute minimum, while FPS contains extra functionality in its label.
2. The cell size in ATM is 53 bytes, whereas in FPS it is in the range of 100 bytes.

An important issue in ATM is the simple, fast and efficient transport of information without complicated protocol functions. The cell header functions (fig.I.2) reveal this principle. 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, [7],[8],[12],[13]:

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 NNI 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. As explained before, 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. As mentioned before, it greatly enhances the concept of dynamic routing, and resource management according to the traffic required class of service.

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 now 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.

In fig.(I.3), we show a possible protocol architecture of the Broadband ISDN network [6]-[8]. The figure shows the ATM transport layer and its position above the physical layer and below the adaptation layer. To summarize, we can identify that the physical layer can be further subdivided into two sublayers: the physical medium (PM) sublayer and the transmission convergence (TC) sublayer. The PM sublayer is responsible for the correct transmission and reception of its bits on the physical medium and is medium dependant (optical, wireless, electrical). The transmission convergence sublayer's main function after bit reconstruction is the mapping of the

ATM cells to the transmission system used (synchronous or cell based hierarchies). The ATM layer has four main functions:

1. Multiplexing and Demultiplexing of cells of different VCIs onto a single cell stream
2. Before (after) the cell is delivered to (from) the adaptation layer, the cell header is extracted (added).
3. In addition a translation of the VCI might be required at the switching nodes.
4. An access flow control is required at the User Network Interface (UNI).

The ATM adaptation layer (AAL) functionality is mainly to provide segmentation and adaptability of the information units into the cell type format of the ATM layer. It is also responsible to enhance the services provided by the ATM layer according to the requirements of specific services. These services can be user services as well as control and management functions. Four main service classes are identified:

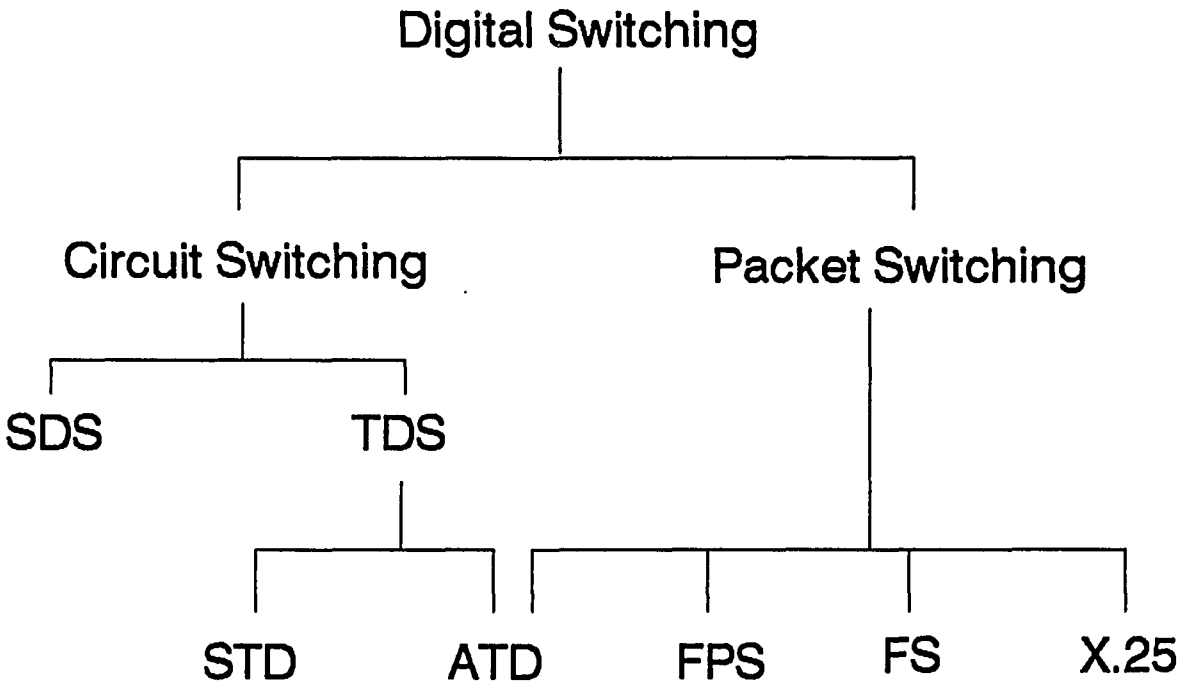
Class A, time relation exists between the source and the destination. The bit rate is fixed and the service is connection oriented.

Class B, time relation also exists for a connection oriented service, however, the bit rate can be variable (e.g., variable

bit rate video or voice).

Class C, no time relation exists, and the bit rate is variable, with a connection oriented service (e.g., data transfer in the user plane, signalling in the control plane).

Class D, similar to class C but the service is connectionless (e.g., LAN interconnection traffic).



SDS: Space Division Switching

TDS: Time Division Switching

STD: Synchronous Time Divsion

FPS: Fast Packet Switching

FS: Frame Switching

ATD: Asynchronous Time Division

Fig.(I.1) Digital Switching Technology, Ref.[4]

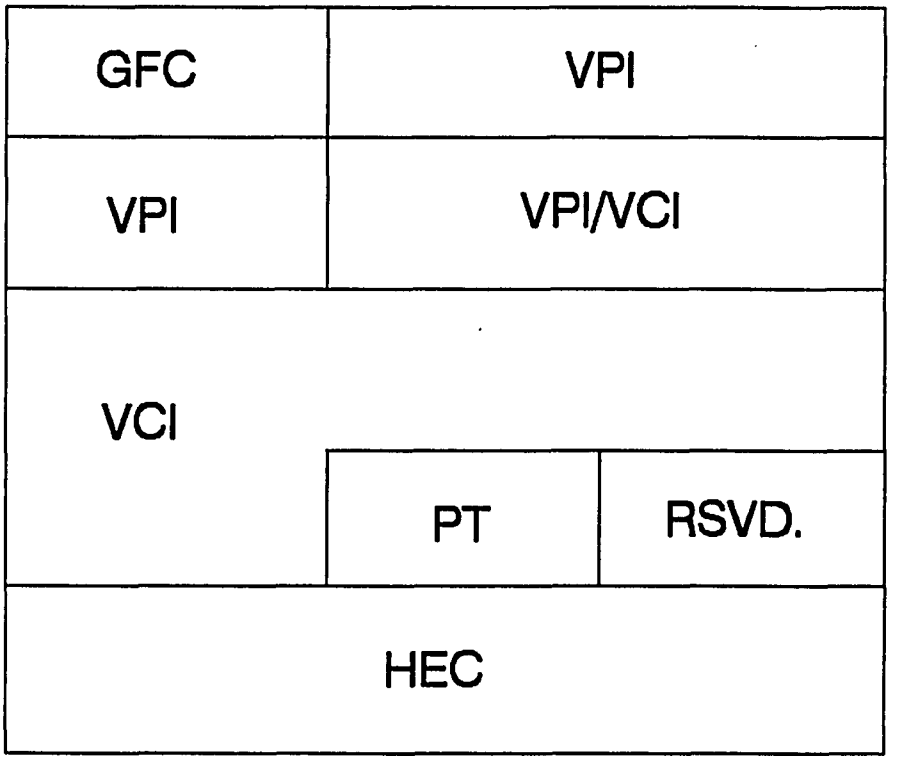


Fig.(I.2) ATM Cell Header at UNI, Ref.[8]

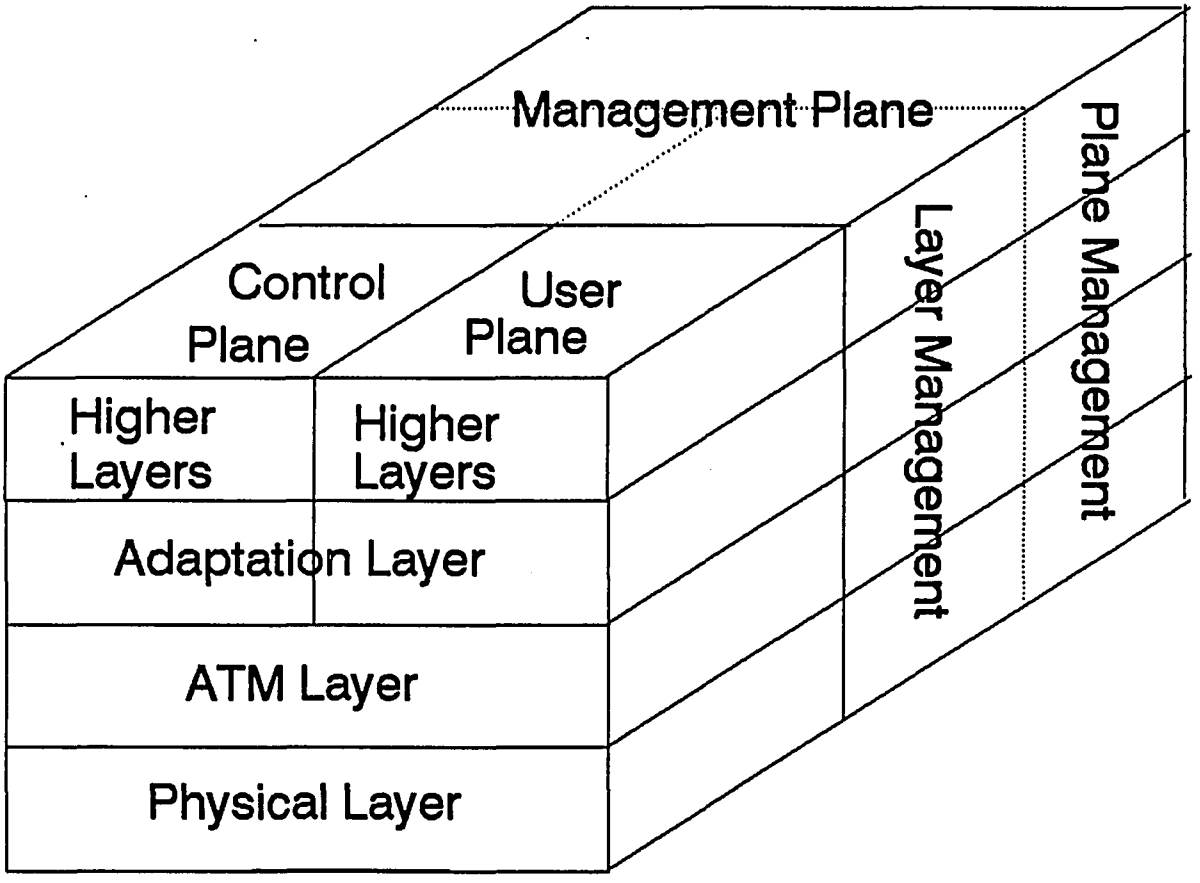


Fig.(I.3) ATM Protocol Reference Model, Ref.[13]

II Flow Control And Congestion Avoidance

II.1. Flow Control in High Speed Networks

Flow control is required to avoid congestion and allocate the network resources, efficiently, to each call according to a predefined class of service that defines performance measures. Congestion is a major problem to be faced in Broadband Networks. Before tackling flow control actions at different network levels, we present a global view of the congestion problem.

Broadband networks operate under a very peculiar environment. It supports a diverse mix of traffic sources with different characteristics, different classes of services, and at very high speeds. A fundamental characteristic of traffic propagating in BISDN environment is that the ratio of the propagation time to the cell transmission time is very high. Congestion occurs when various sources compete for the network resources and these resources do not meet the demands. The problem can happen at the network level, i.e. over links capacity assignments or over a particular area of the network (switches, multiplexers, interface plane,..etc.). It can happen at the call level, when logical channels request

bandwidth that can not be supported, or when a certain call exceeds its reserved bandwidth limits, or when the network admits more calls than the link can handle.

At the cell level, congestion can happen at any node due to poor buffer management, or as a result of upper layers congestion leading to the direct impact of congestion at the cell level, or when long bursts of cells are formed at some buffer leading to buffer overflow. Note that due to the stochastic nature of the traffic, long bursts of cells can get formed even when the network links are underutilized. Flow control and resource management policies at one level have a direct impact on the other levels and the congestion problem can be only solved within a global solution that relates different actions at different levels. This layered approach to the problem has been considered by a number of authors, for example see [24]-[28].

II.2. Flow Control Requirements

We can set a requirement framework for the solution of the congestion problem and apply it to possible implementation methods at the network, call and cell levels. The basic requirements can be summarized in the following points:

1. The flow control algorithm should be flexible, in the sense

that it must be efficient to handle traffic characteristics which are yet to come.

2. The algorithm must be effective in handling real time traffic at high speeds, also it must be able to handle traffic with different burstiness characteristics.

3. It must be simple with little overhead. Simplicity is required at broadband speeds in order to minimize the overhead processing of complicated algorithms.

4. It must be robust, in the sense that it must be able to solve the congestion problem without relying on specific detailed information about the type of traffic generated by the end users. It must not be sensitive to short term high fluctuations in traffic conditions, since any actions taken during those periods can be misleading and can lead to false flow control decisions.

5. It must be able to preserve the network resources, without degrading the users' requested performance measures. However in protecting the network resources, the control actions may cause degradation in the performance and in some severe cases disrupt the service.

6. It must be fair to all users. Fairness is not a clear defined term, allocating resources to some users while denying others can be unfair. Uneven distribution of network resources can be unfair. However in Broadband networks, we are faced with different traffic characteristics with widely different demands that define different classes of services. Therefore

it is impossible to allocate even resources to different classes of services. We will define fairness as the even distribution of resources to users belonging to the same class of service, as long as none of the users has tried to exhaust those resources.

II.3. Flow Control Design Guidelines

To provide an effective flow control architecture based upon the above framework, is a challenging task. In the following, we present the basic design principles of flow control in ATM networks.

A. Connection control principle

As described above, ATM based networks are designed upon a connection oriented transport principle. Connection oriented mechanism implies a set up of a virtual connection between the virtual connection end points, and then the call is progressed over that connection. We can distinguish two phases of operation:

1. Call set-up phase, during which network resources are checked and reserved along the connection route before the call starts. Once network resources are available, and can satisfy the required class of service, then the call is accepted.
2. Call progress phase, during which the actual call progresses

and traffic is transferred. Throughout this phase, the traffic management functions monitor the traffic status and maintain the network resources .

Connection control implies that an important flow control decision is taken during call set up. The decision is whether or not the network will accept the call (admission control). The importance of the decision stems from the fact that based upon it, the network can be either in a congested or congestion-free status. Once the decision is taken to accept the call, the network must allocate the necessary resources to accommodate that call and maintain the call's class of service. If the network can not support the required class of service, it can either reject the call or notifies the user that the required class of service can not be honored and a less stringent requirement class of service will be delivered. When the call is in progress the network must control its resources by monitoring the input traffic and restricting the call traffic characteristics to its declared parameters (traffic enforcement or policing function) [29]. It is clear to see that the flow control philosophy will be to avoid congestion rather than reacting to it (i.e. preventive control rather than reactive control).

B. Buffer Management and Scheduling

ATM networks provide different users with different services

according to their respective class of service. The class of service principle has a direct impact, not only on the traffic management functions, but also on the buffers management and scheduling policies. In order to implement the class of service principle, each class of service traffic will be accommodated in a separate queue. Buffer size and bandwidth requirements vary greatly from one traffic to the other, according to the traffic characteristics. Real time traffic, such as video and voice, needs small buffer sizes in order to limit the maximum delay. Data traffic, is more sensitive to cell losses and hence require large buffer sizes. Assigning different types of traffic to separate queues will make it possible to enforce traffic enforcement functions. For example, if congestion occurs, we can easily drop low priority cells with the minimum effect on the quality of the transmitted voice or video. Control signals must be assigned separate buffers, as they require highest priority. Efficient buffer management schemes are required to manage different buffers with different traffics, in order to keep buffers occupancy levels under control and avoid excessive cell loss due to buffer overflow. Scheduling policies reflects not only, service priority schemes, but also bandwidth assignment schemes. For example, video traffic will be assigned higher bandwidth than voice or data traffic. The virtual path concept will simplify the implementation of dynamic bandwidth assignment and the use of priorities [30].

C. Reactive Control

As stated in (A) above, flow control in ATM networks is based upon the philosophy of preventing congestion rather than reacting to it. In general, flow control techniques are functionally categorized into two categories, preventive type or reactive type. In point (A), we realized the need for admission control and traffic enforcement actions at the input access nodes to the network, these actions are preventive type control. Reactive control tries to alleviate congestion after occurring, and in principal it does that by sending a feedback signal from the destination point to the source informing the source of the congestion and throttling the input traffic. There are numerous number of schemes to achieve this goal and they differ mainly in the manner by which the feedback message is conveyed (for example see [31],[32] and the references therein). Window based control or rate based control are examples of reactive type flow control. In window based control, the destination sends a feedback signal, accordingly the sender limits the window size and hence the number of transmitted packets. In rate based control, the destination specifies the number of packets per seconds that the sender can transmit. These schemes are widely used in conventional data packet switching networks.

It is clear that in ATM networks, reactive control schemes are not the answer to the congestion problem due to several

reasons.

1. The ratio of the propagation delay to the cell transmission time is very high, hence any end to end feedback message will be clearly outdated and is not useful in relieving congestion.
2. These schemes are sensitive to the transient traffic fluctuations and this effect may lead to instability in the feedback loop dynamics.
3. These schemes use excessive overhead functions and complicated protocols are needed to implement them, which is totally contradicting to our requirements in BISDN networks.
4. Packet acknowledgement messages may lead, under the very high speed links, to increase congestion and catastrophic build up of queues.

In ATM networks, we still need to have some form of simple and effective reactive control scheme. Although the main defense line against congestion is preventive control, congestion may occur and we must use reactive control to relieve it. One solution, is dynamic routing, another solution is to employ end to end window type mechanism which is modified to suit the high speed environment (e.g., to acknowledge blocks of information units instead of single message acknowledgment. Fig.(II.1) shows a classification of the flow control functions according to their location within the network. In the following, we shall provide a detailed discussion of these functions.

II.4. Network Level Flow Control

II.4.1. Dynamic Bandwidth Control

Flow control at the network level is concerned with traffic management functions over different links and nodes in the network. The management functions exercise flow control over traffic travelling through the network. It maintains an acceptable level of traffic utilization, over different links, such that to avoid congestion and relieve it, in case it happens. We can realize, immediately, the necessity to perform the following functions, Link capacity assignment, dynamic routing and call congestion control [26],[28].

It is crucial that flow control, at the network level, be able to perform efficient capacity assignment per link, which will lead to the maximum utilization of the link capacity and provide control over the allocated bandwidth to the virtual connections. The result, of course, will be to minimize the probability of denying a call acceptance request and call control call congestion. Statistical bandwidth allocation is best suited for the variable bit rate traffic, expected in ATM networks. If fixed capacity virtual paths (no bandwidth control) is used, then this capacity must be large enough to support the maximum number of calls over the path. Clearly, this method will waste the unused portion of the bandwidth

when the total number of calls are not supported. This unused bandwidth can be used to avoid congestion on another highly utilized path. Even if the number of supported calls are fixed, the stochastic nature of the incoming traffic can greatly change the bandwidth required, hence we must have efficient control over the bandwidth. Therefore dynamic path bandwidth control must be the answer. If the number of supported calls is small, the assigned bandwidth is decreased, as the number of calls increases, the assigned bandwidth is increased accordingly, see chapter V for our proposed scheme.

Two different methods for bandwidth assignment are possible (fig.II.2). In the first one, traffic with similar characteristics are combined over the same virtual path, whereas in the second different types of traffic are supported over the same path [34]. The former method has several advantages, firstly it is easier to implement the class of service control according to each type of traffic. Secondly, it is also easier to implement dynamic bandwidth control per path since we know that the traffic per path is of similar characteristics and therefore a unique bandwidth allocation scheme is used per path. Thirdly, multiplexing several classes of traffic, with different characteristics, decreases the multiplexing gain compared to the multiplexing of traffic with same multiplexing [35],[36]. In [33], [34] and [37], a dynamic path bandwidth control scheme was analyzed. The scheme works as follows

1. An arriving call requests bandwidth, if available the call is accepted, otherwise the scheme requests an increase in the path bandwidth.
2. If the increase is granted then the call is accepted otherwise it is rejected.
3. The scheme monitors the utilization of the path bandwidth, if it is not used then the bandwidth assigned is decreased.

Both schemes provide a significant increase in the transmission efficiency achieved by the dynamic path bandwidth control (fig.II.3). Of course, as the transmission efficiency increases the processing load increases (fig.II.4). The processing load is defined as the ratio of the frequency of the virtual path bandwidth change request and that of the call setup attempts.

The call decision acceptance and the dynamic path bandwidth control at the network level are highly correlated . At the network level, the network controller takes a decision to accept or deny an increase in the bandwidth allocated to each path. As with the call level acceptance rule, the decision to increase the path bandwidth is based upon the call rejection rate. The call rejection rate is a measure of congestion at the call level. If this rate increases above a maximum limit then the corresponding path bandwidth must be increased. The availability of link capacity to increase the path bandwidth

is function of the number of the bandwidth required to support other paths on the same link at the same instant. If the bandwidth is not available, then dynamic routing is the answer.

II.4.2. Dynamic Routing

Dynamic routing, in essence, provides the capability to perform network dynamic reconfiguration. There are several benefits to this important function

1. Network adaptability to acquire short or long term traffic demand variations in a simplified transport structure.
2. Providing back-up alternate routing in case of link failure.
3. An efficient means of alleviating congestion, per link, in case it happens. Note that we have classified dynamic routing under the reactive type flow control.

As mentioned before, the concept of the virtual path provides us with the tool to implement an efficient and flexible traffic control functions at the network level, which in turn has a direct impact on the efficiency of the flow control functions at the underlying call and cell levels [33]. For sake of completeness, we repeat the definition of the virtual path. The virtual path is a logical connection, between the virtual path terminators, that is composed of a bundle of virtual channels (also known as virtual circuits). Virtual paths are crossconnected and controlled, by the network upper

management functions as a single entity (fig.II.5). Thus virtual paths define the crossconnection functions across the network, while virtual channels (or circuits) are concerned with switching and connection establishment functions.

The virtual path concept has the following characteristics

1. A predefined route is associated with each virtual path in the physical network.
2. A virtual path is identified using a label attached to the cell (called virtual path identifier). Each VPI has a local significance over the link, it does not provide a global significance over the network. The reasons behind this assignment method are to keep the VPI length small because there is an upper limit to the number of multiplexed paths per link, and to avoid management of virtual paths in a centralized manner, and to provide flexibility in assigning virtual paths per physical link.
3. Each virtual path is assigned a certain bandwidth, that determines the number of virtual channels it can support. The bandwidth allocated can be either deterministic or statistical, in the former case the virtual path is called a labeled deterministic path (LDP) whereas in the latter case it is called a labeled statistical path (LSP). LSP has several advantages over LDP, such as the ability to exercise path bandwidth control, and improvement of the transmission efficiency by exploiting the statistical multiplexing gain, and

optimum allocation of bandwidth according to the traffic characteristics of each call.
4. Virtual paths are statistically multiplexed on the physical link on a cell multiplexing basis.

Although there are several similarities between the virtual path concept and the digital path concept in STM digital transmission networks, there are significant differences that has enhanced the virtual path solution. Firstly, virtual paths are labeled digital paths that are multiplexed on a cell basis whereas STM paths are positioned paths over the physical transmission link within a certain transmission frame. Secondly, VPs' can be allocated statistical bandwidth assignment while STM can be only allocated deterministic bandwidth. Note that positioned paths can accommodate both labeled and positioned paths but labeled paths can accommodate only statistical or deterministic labeled paths (fig.II.6).

The virtual path concept has provided several advantages that have contributed in the realization of flow control functions at the network level. These advantages are:
1. Elimination of the transit node processing per call set up, thus simplifying the node functions and providing fast switching per cell. Also, no processing is required at the transit nodes when the path capacity is allocated or changed. Also, all switching fabrics need only simplified functions.

2. Separation of the logical transport network from the physical transmission network, thus providing flexibility in performing traffic management. For example, we can change the virtual path capacity without affecting the physical interface structure. This feature will greatly simplify the network architecture.

3. Direct multiplexing of virtual paths with different capacities while using simple hardware and software. No need to employ the hierarchical multistages multiplexers as with the STM case. Dynamic path routing and dynamic path bandwidth allocation are now simple to implement.

4. Statistical bandwidth allocation for the calls per virtual path provides efficient utilization of the total link capacity.

II.5. Call Level Flow Control

II.5.1. Admission Control and Call Acceptance Rule

We have discussed the necessity for dynamic bandwidth control over each path. The bandwidth allocated per path depends upon the number of calls supported per path which in turn depends upon the required bandwidth to support each call. It should be clear, that there is a strong dependency between flow control at the network level and at the call level. If it is possible to design an efficient algorithm for allocating

bandwidth to each call, this will reflect upon the bandwidth allocated per virtual path. We add that the decision to accept or reject a call (admission control) relies upon another decision taken at the network level, which is to increase, or decrease, the virtual path capacity. If the increase in the virtual path capacity is granted, then the call will be admitted. The following parameters that influence the decision of call acceptance:

1. Type of traffic characteristics, which specifies the bandwidth required to accept the call.
2. Required class of service, for example a certain amount of bandwidth can be sufficient to support the call with a certain class of service, while the same bandwidth may not be sufficient to support the same call if the class of service is more stringent.

The acceptance decision is simply based upon the availability of bandwidth to support the call along its connection path. There are several approaches to the indication of the bandwidth availability. One is to use the long term probability of cell loss at the multiplexer buffer of the input access node to the network. Another approach is to use the instantaneous cell loss rate [38], as a criteria to the call acceptance. The instantaneous cell loss rate is a measure of the cell loss rate but taking into consideration the dynamic change in traffic conditions due to the bursty sources. Although

the second approach captures the effects of traffic variations, it makes the flow control management functions very susceptible to short term fluctuations in the traffic. Since it is possible that these variations be very abrupt and steep, the network access control may take decisions that are changing very frequently over short periods of time. This behavior is not recommendable since it contrary to the robustness requirement of flow control. So the cell level provides the upper call level with an important performance parameter which is the cell loss rate. Notice the similarity, in the interaction between the cell and call control levels on one hand and the interaction between the call and network control levels on the other hand (fig. II.7).

The decision to accept or deny a call request at the call level depends primarily upon the availability of the requested bandwidth. To perform this function efficiently, we must provide the call level control with an efficient bandwidth allocation and management schemes. Several schemes have been studied in the literature, for example see [39]-[48] and the references therein. To start with, the deterministic bandwidth allocation scheme, which is based upon the peak rate assignment, is not useful in ATM networks. This scheme has been widely used in circuit switching networks, and is best suited for STM transport mode. In Integrated traffic networks, this scheme does not provide the flexibility required to integrate different

types of traffic with variable bit rates and bandwidth requirements. Also, this scheme does not exploit the statistical variations of the input traffic and consequently wastes the bandwidth. Hence the network resources are wasted and the achieved efficiency is very low.

In [42] a call acceptance decision criteria, is based upon a parameter Rs which is equal to the call/line bit rate ratio. This ratio determines the maximum call throughput in a short duration of time divided by the link bit rate. The decision criteria is dynamic according to the traffic characteristics which defines the ratio Rs. In cases of small Rs, the call acceptance decision is based only upon the trunk line utilization, whereas for large Rs situations the decision is based upon two parameters one is Rs and the other is the call throughput over a medium period of time. For extremely bursty traffic, a short hold mode ATM service is proposed. In this mode as soon as the user terminates a certain active period, the network temporary releases the network resources while keeping the logical call. Before the user starts to transmit again, he must ask the network to re-allocate the resources to the call. The motivation behind this scheme is that as the traffic gets more bursty the line utilization decreases, and more bandwidth is require to support the call. So to achieve high bandwidth efficiency, it is better to have more than one criteria of call acceptance according to the traffic status.

II.5.2. Dynamic Bandwidth Allocation

Statistical bandwidth assignment is required to fully utilize the network resources and exploit the statistical variations of the traffic. Since we are adopting a transmission mode with class of services where statistical multiplexing is performed per class of service, then the allocated bandwidth will depend upon the class of service to whom the traffic belongs. We can distinguish three main types of traffic, data, voice and video. Each type has its own characteristics, class of service and is defined in terms of the burstiness parameters. A bursty traffic source is a one that transmits over a certain period of time (active period) then enters a period of no transmission (idle period). A simple description to characterize the source is with three main parameters:

1. Peak rate (bits/sec)
2. Average rate (bits/sec)
3. Average burst length (secs.) which is the average duration of an active period of transmission at the peak rate.

The burstiness is measured by the ratio of the peak to the average bit rates (also, referred to as the burstiness index). There are other parameters to characterize the burstiness of the source. For example, variable bit rate video sources can be represented by time domain characteristics such as, the coefficient of variation, the autocorrelation function and its distribution. Burstiness can also be represented by the squared

coefficient of variation of the interarrival times which is the ratio of the variance of the call interarrival times to the square of the average value of the cell interarrival times [48] Sources exhibit different traffic patterns according to their type (i.e., video, voice, data, interactive image,..etc.). They also produce different patterns according to, the coding technique used, and the contents of scenes (in case of video sources). Some variable bit rate video sources have burstiness index ranging from 1.9 to 3.5 [43],[44].

Therefore an ideal bandwidth allocation scheme must consider the bursty nature of the traffic. In [47], a scheme was proposed that allocates an amount of bandwidth which is between the peak and the average rates. This amount is called the effective bit rate which is a constant multiplied by the peak rate. This constant reflects the variance of the traffic characteristics. In [43] and [45], a similar principle was applied to allocate bandwidth to a variable bit rate video source, where the allocated bandwidth is the sum of the average bit rate and constant term representing the standard deviation. The scheme approximates the arrival statistics from N independent video sources by a Gaussian distribution and limits the maximum achievable load to be 0.8 of the total link capacity. The bandwidth allocated is determined according to an estimation of the mean and standard deviation of the gaussian distribution.

In [35] a similar allocation scheme is proposed. The bandwidth allocated is equal to x% value of an estimation of the distribution of the arrival traffic.

The ideal scheme, will be to allocate some value between the mean and peak rates of the traffic. The question is, what is this optimum value?. In [40]-[42], the characteristics of the traffic were studied to investigate the effects of burstiness parameters on the allocated bandwidth. A scheme called the class related rule was proposed. It allocates bandwidth to each type of traffic according to a rule which is function of the average, peak and burstiness of the source, subject to a required cell loss rate. It was proven that as the burstiness of the source decreases, the bandwidth required to satisfy a certain class of service, decreases. A very important parameter is the ratio of the source peak rate to the link capacity. As the input source peak rate approaches the link capacity, the bandwidth assigned to that source reaches the source peak rate. The effect becomes more appreciable when the multiplexer buffer length is less than the average burst length. The reason is that as the peak rate increases, and with a limited buffer size, the statistical multiplexing gain decreases and we have to increase the assigned bandwidth to avoid excessive cell loss. In [48], it was proven that we can avoid the required increase in the bandwidth by a proper feedback control of the arrival rate such that to limit increase

in the peak rate. The ratio of the source peak rate to the link capacity must not increase above 0.1 in order to have an effective statistical multiplexing gain and avoid congestion. Also, the burst length is quite important when it is comparable to the size of the multiplexer buffer. However, in most real time applications, the buffer length will be in the order of several microseconds and the burst length will be greater than the size of the buffer, so the effect of the burst length on the allocated bandwidth is not appreciable. In chapter V, we provide a detailed analysis of our proposed scheme called Bandwidth Control Period (BCP) rule.

II.5.3. Dynamic Bandwidth Management and Scheduling Policies

From the above discussion, allocating the bandwidth to each class of traffic, is not enough to prevent congestion. In an ATM environment, the link capacity is statistically shared among the users to provide maximum efficiency and flexibility. Buffer sizes are limited, and although segregating the traffic into different classes of services provide a minimum bandwidth allocation to each class, we need to have an efficient management technique of the bandwidth. Since the early introduction of ISDN, there has been a considerable work in the area of bandwidth management to voice and data traffics. The problem was formulated so as to find a scheduling mechanism that can accommodate circuit switching traffic (voice) and

packet switching traffic (data). The transport media was basically synchronous TDM frame of a certain duration representing the bandwidth and the question was basically how to divide this time frame (bandwidth) among a heterogeneous traffic mix (voice, data). Notice that voice was supported on a separate circuit switched resources and data was supported on the packet switched resources.

One of the early bandwidth management techniques is the movable boundary scheme. In this scheme, the bandwidth is divided into two parts, one is reserved to the voice and the rest is used by data traffic. The boundary is movable according to the utilization of the data traffic. Voice traffic is blocked if there is no enough bandwidth to support it, whereas data traffic is queued. Several service scheduling policies were investigated [50], such as First input first output (FIFO), Preemptive priority (PP) and Sorting. FIFO policy does not really provide any service scheduling control but was found to attain fairness at the expense of bandwidth utilization. PP policy provides preemptive priority to voice traffic, but the data traffic is allowed to use the any available bandwidth at the risk of being preempted. Sorting policy maintains a list of the waiting customers sorted in a descending order of their required bandwidth.

It was found that a dynamic scheme that would switch from one policy to the other according to the traffic load, provided the best results in terms of maximum throughput (for voice traffic calls) and minimum delay (for data traffic) [50]. Another version of the movable boundary scheme is possible, when voice traffic is packetized. In this version, data packets is allocated a predefined number of time slots, in order to avoid bursts of voice packets from causing excessive delay to the data packets. The rest of the available bandwidth is shared among both voice and data packets. If voice packets arriving per frame exceed the available capacity, then the extra packets are discarded. The problem with movable boundary scheme is that voice is not allowed to use idle data time slots, thus the efficiency decreases. Burst switching is another alternative scheme to the movable boundary one. A burst can be either voice or data packets that are generated when a source is in an active period of transmission. Each burst has its own header and switching is done on the burst level. Voice bursts are given non preemptive priority over data bursts. Data bursts can be queued, if the channel is not available to support it, while voice bursts are queued for up to a maximum of two milliseconds only. Burst switching does not provide the flexibility in bandwidth assignment per user demand. The reason for that, is that the transmission link

is divided into narrow time slots of equal bandwidth (channel), and each channel is allocated to the burst. Intolerable delays may happen to either types of traffic.

In [49], an alternative scheme to the famous movable boundary strategy was proposed. The scheme divides the available bandwidth among two classes of traffic according to a fixed time ratio called (T1,T2), where T1,T2 is a fraction of the total bandwidth. This ratio is the minimum bandwidth guaranteed to each class of traffic. Each traffic is queued in a separate queue, which is consistent with the ATM multiplexing and bandwidth assignment according to the class of service required. Each queue is visited, alternately, and the scheme is dynamic in the sense when a queue is exhausted transmission is immediately switched to the other queue. The performance measures of this scheme when compared to the FIFO scheme, is more superior in terms of the voice quality and bandwidth utilization. The choice of T1 and T2 were motivated by the overload protection of one queue from the other. In other words, T1 is set to be the maximum size of queue (2) (in secs.) such that we can not stay in queue (1) time more than the time required to evacuate queue (1) at its maximum capacity and vice versa.

II.6. Cell Level Flow Control

II.6.1. Traffic Regulation and Multiplexing Efficiency

An important function of flow control, is that during the progress of the call, it must monitor the traffic conditions along the path of the call. If the call does not behave according to its declared characteristics, the controller must then arbitrate the call and limit the cell arrival rate to the original declared value at the call set-up time. In an overload situation the same action must be taken to avoid severe congestion problems. The problem that must be considered in any bandwidth enforcement technique is that high traffic bursts can be due to either overload or due to natural statistical variations in the arrival process. If these bursts are measures over short time periods, then it is more likely that they are due to normal statistical variations. However, if they are measured over relatively long time periods it is quite difficult to tell whether they are due to congestion overload or not.

In [38], it was proven that if congestion occurs it tends to stay for long periods of time. This problem has motivated the use of simple deterministic parameters to describe the arrival traffic process, such as the average value, peak value and burstiness length. So a design principle in a traffic enforcement scheme (policing function), is to perform traffic

shaping over the input bursty traffic. We have discussed before, as the burstiness increases, more bandwidth is required to support the traffic. The smoothing effects of the traffic through buffering, is not very effective in ATM networks because of the limited buffer sizes. Also as the peak rate of the input source traffic increases, the problem gets worse. As this ratio increases above 0.1 [25], [41] and [42], the number of supported sources drops. Due to that effect, the statistical multiplexing gain decreases and the transmission efficiency drops. This problem is more obvious in bursty traffic with high peak rates such as video retrievals. It becomes essential then, to apply traffic control at the input access node of the network. This scheme must not only regulates the traffic to its declared values in order to avoid congestion, but it must be able to smooth down the characteristics, specifically it must decrease the input peak rate which is a main parameter in an efficient statistical multiplexing [48]. It is to be implemented at the input access multiplexer and at other multiplexers and switches along the cells path.

At the cell level, it is quite clear that there is a strong correlation and dependance between the shape of the traffic arrival process and the flow control actions at the call level. In other words, there is a strong dependance between flow control at the call level and flow control at the cell level. Indeed, a successful functionality of bandwidth assignment and

avoiding congestion at the call level, depends upon how the flow control at the cell level is successful in controlling and shaping the statistics of the arrival process. The main objective, here, is to control the input flux of cells at heavy congestion states, or when a certain source transmits at a rate higher than its declared characteristics. Another important objective is, as mentioned before, 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.

The above discussion has motivated the study of the performance of the statistical multiplexer under various types of traffic (i.e., voice, video, data arrivals). Several authors have addressed this important issue in order to gain some sight of the multiplexer behavior (for example see [51]-[57]). The main analytical problem was to characterize the bursty traffic source, whether video or voice source, and describe then describe the overall superposition process by a simple analytical tool. Several models were introduced to model the voice source [58], [59]. The basic model is to represent each source by a periodic process alternating between a talkspurt and a silent modes. This periodic process is then represented by a simple two state Markov chain, where the time that the process spends in each state is approximated by an exponential distribution. The aggregate arrival process from N input sources is a complex nonrenewal process, hence we render to

approximations. One approximation is to represent the aggregate arrival process by a continuous time phase process. Another possible representation of the aggregate arrival process is by a two state Modulated Markov Poisson Process (MMPP) [55]. 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.

The modeling of the video source is more complicated than the voice source model. The output bit rate stream of a video source depends upon the specific scene contents and the coding technique used. In ATM networks, the variable bit rate coding techniques will be implemented to transmit the video information [60]-[65], due to the flexibility and data compression capabilities of these types of codes. Video-phone or single activity motion scenes encountered in teleconferencing produce different characteristics than those produced by broadcast T.V., under the same coding technique. Hierarchical layers coding [65], is best suited for high activity motion scenes, thus different analytical models are needed to represent different kinds of video signals. In general the bit rate of a video source is represented by a continuous state auto-regressive process [53], [60], [61], unfortunately it yields significant difficulties in any analytical analysis. In [53], a single activity scene was represented by the

auto-regressive model, and then approximated by a simpler one dimension discrete state Markov chain, very similar to the phase process of the voice aggregate stream. We must stress, that there is a strong dependency between the transport principle of the ATM network and the coding method employed. Specifically, sub-band coding has made congestion control and efficient statistical multiplexing much easier. It is extremely difficult to apply efficient flow control techniques with high activity video signals with high peak rates and burstiness. However with sub-band coding, the highly fluctuating bit rate of such kind of signals, can be decomposed into separate bands of smoother characteristics [65]. Each band can then be accommodated on a separate multiplexer buffer according to the class of service [48]. During periods of high overloads, the less important information can be dropped, or preemptive priority can be assigned to the bands conveying the main video information.

II.6.2. Traffic Enforcement and Congestion Control

In [48] and [68], an access flow control algorithm was proposed and analyzed. It controls the input arrival process upstream the network (i.e., at the input access node) based upon the feedback throttling of the arrival process to the input statistical multiplexer. A feedback control signal, proportional to the congestion level of the multiplexer, is

applied to the input source coder which controls the source rate via decreasing the coding rate (number of bits/sample). When another congestion threshold is achieved a similar action is taken, thus reducing the input rate further more . The threshold levels are activated, by the controller, according to the specific traffic characteristics and required class of service. In chapters III and IV, we shall explore this method in details.

The advantages of this scheme, are several. First, it prevents congestion from happening and greatly reduces the potential of congestion down stream the network. Second, it is applicable regardless of the type of coder used. It is good for variable bit rate coders, as well as fixed bit rate ones. Third, it provides the means for the maximum possible shaping of the input arrival process. Therefore, we can decrease the bandwidth allocated to the input call and achieve the same required class of service. Also, the statistical multiplexing gain is enhanced and we can achieve a significant improvement in the throughput. The price to be paid, is that we may perceive a slight degradation in the quality of the voice or video delivered. Interestingly, the degradation is imperceptible over most of the multiplexer utilization levels. This scheme deals with the problem of controlling bursty traffic via a two steps procedure, first it decreases the connection peak rate, second as a result of decreasing the peak arrival

rate we can multiplex more connections and in doing so enhance the smoothing effect of the multiplexer (recall that as the multiplexer traffic intensity increases the multiplexing gain increases too). The scheme avoids the need to drop excessive cells, which is not quite difficult to accomplish as explained in the leaky bucket scheme in the next paragraph.

The leaky bucket scheme is another congestion control scheme. The scheme is based upon an estimation of the source average bandwidth. When this average exceeds a certain threshold, new arriving cells are dropped until the average rate drops back to the original estimated value. The scheme highly depends upon the estimate of the input average bandwidth, thus it does not provide efficient policing function with statistically variable traffic. If the control parameters are set close to the declared mean rate, only a small infrequent variations are admitted and a high cell loss rate will occur. Thus while trying to avoid congestion, the traffic characteristics have been greatly altered due to the high dropping rate of the incoming cells. On the other hand, if the control parameters are set with large tolerance to guarantee the admission of the declared mean value, then the scheme will be insensitive to large variations above the declared parameters and congestion will occur.

In [70], another improved version of the leaky bucket scheme was proposed and analyzed. The scheme provides a limited size buffer to avoid excessive cell drop. To enforce the bandwidth limitation, tokens are generated into a token pool according to a predefined rate corresponding to the declared average rate of the traffic. Cells are not allowed into the network unless they obtain a token. To allow for the declared degree of burstiness, there is a maximum limit on the number of tokens in the token pool. If the token pool is filled, the token generation process is shut off. There is a tradeoff between the pool size and the generation rate of the tokens. As the token generation rate increases, the waiting time decreases and the variance of the inter-departure time increases to reach that of the input cell arrival process. This effect is clearly intuitive, for as the token generation rate increases, almost all the arriving cells are allowed into the network and the bandwidth enforcement becomes ineffective. Similarly as the pool size increases, the waiting time decreases, and the variance of inter-departure time increases to reach that of the input cell arrival process.

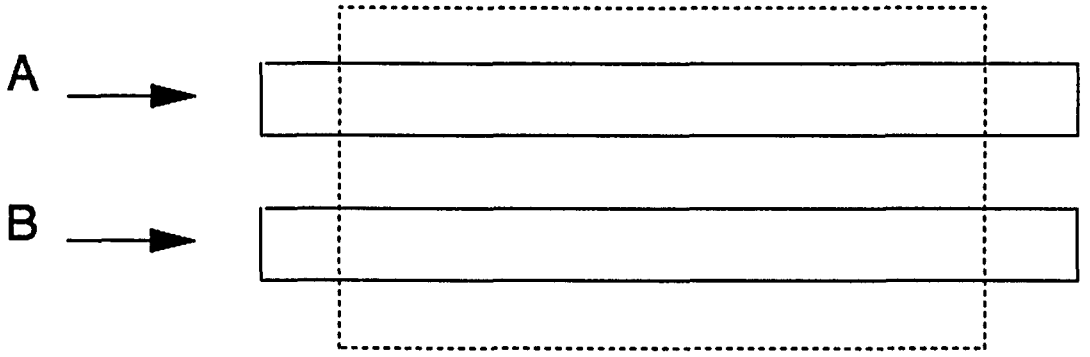
In [40], [56] and [71], the virtual leaky bucket scheme was analyzed. It is based upon the principle of improving the throughput via marking the cells that have violated the declared traffic parameters, and let them enter the network. As congestion is detected in one point of the network, these cells

are then dropped. The motivation behind this scheme is to minimize the cell dropping at the input access node which is inevitable in the leaky bucket scheme. We recall that in the leaky bucket scheme, the cell loss can be improved via increasing the buffer size, however in doing so the waiting time increases to unacceptable values for real time traffic and congestion will be literally introduced instead of being prevented. The virtual leaky bucket scheme, can be easily implemented using a threshold level on the buffer size, after which cells are marked. The problem with the virtual leaky bucket is how to identify the violating cells and mark them without marking the necessary nonviolating cells. In ATM networks, due to the bursty nature of the traffic, a burst of cells may arrive in a relatively short time followed by a silent period. In this case the declared average value has not been violated, yet some of the cells, which are generated during the burst period, will be marked and then discarded at any congested node. A solution would be to use a large threshold value, at the buffer, before marking the violating cells. However, and similarly to the leaky bucket case, choosing a large threshold will cause the bandwidth enforcement to be really ineffective. We realize now that due to the bursty nature of the traffic it is difficult to choose an optimum control parameters that would enforce the bandwidth allocated and prevent congestion without discarding cells carrying essential information for the call connection. In our view, any effective congestion control scheme must be

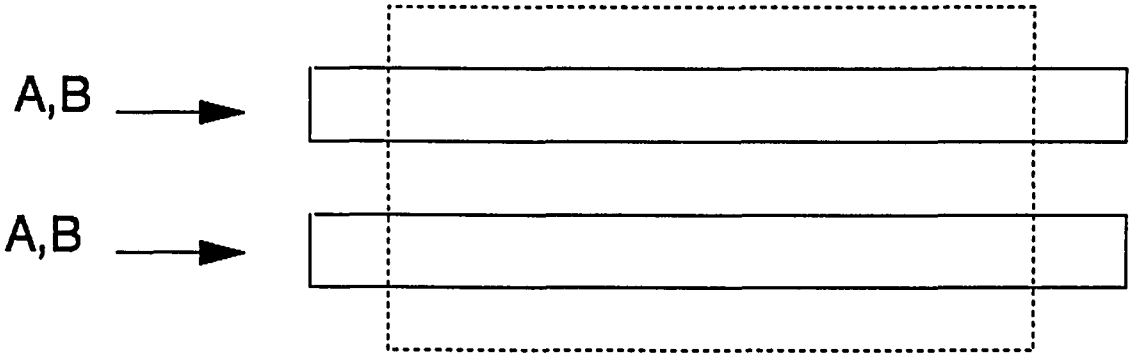
capable of performing traffic smoothing functions. The access flow control scheme, proposed and analyzed in chapters III and IV, is more suitable for the input access node to the network whereas the virtual leaky bucket is more suitable for the transit nodes.

WHO	WHAT	WHERE	WHEN
Cell Level Control	Traffic Policing Scheduling	Access Nodes Transit Nodes	Call Setup Call-Progress
Call Level Control	B.W. Allocation Call Acceptance Route Selection	Access Nodes Access Nodes Transit Nodes	Call Setup Call Setup Call Setup
Network Level Control	B.W. Control Path Routing	Access Nodes Transit Nodes	Call Setup Call-Progress

Fig.(II.1) ATM Control Layers Functionality



(a) Scheme 1



(b) Scheme 2

A: Traffic Class A

B: Traffic Class B

Fig.(II.2) Path Bandwidth Allocation Schemes, Ref.[34]

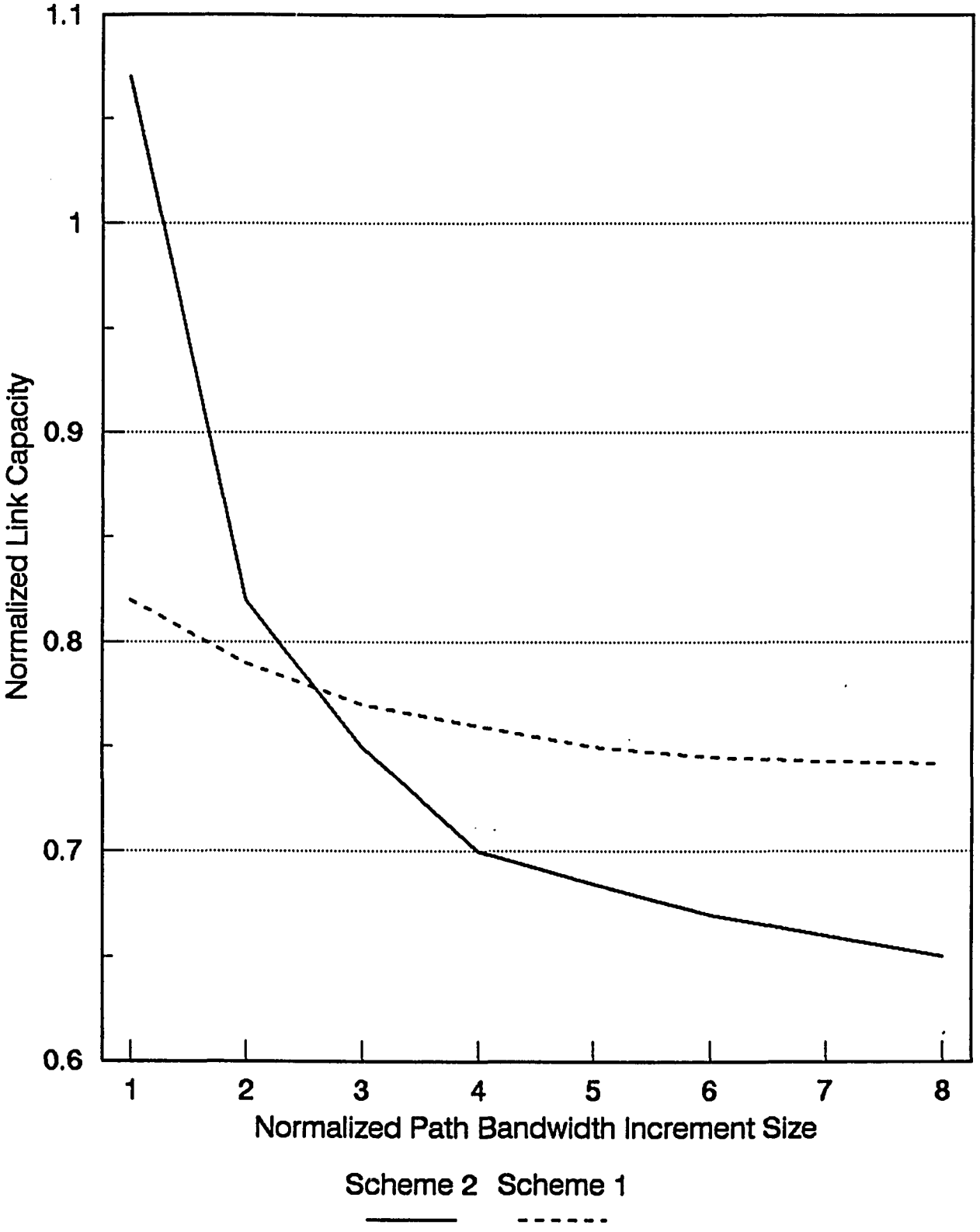


Fig.(II.3) Path Bandwidth Vs. Link Capacity

Ref.[34]

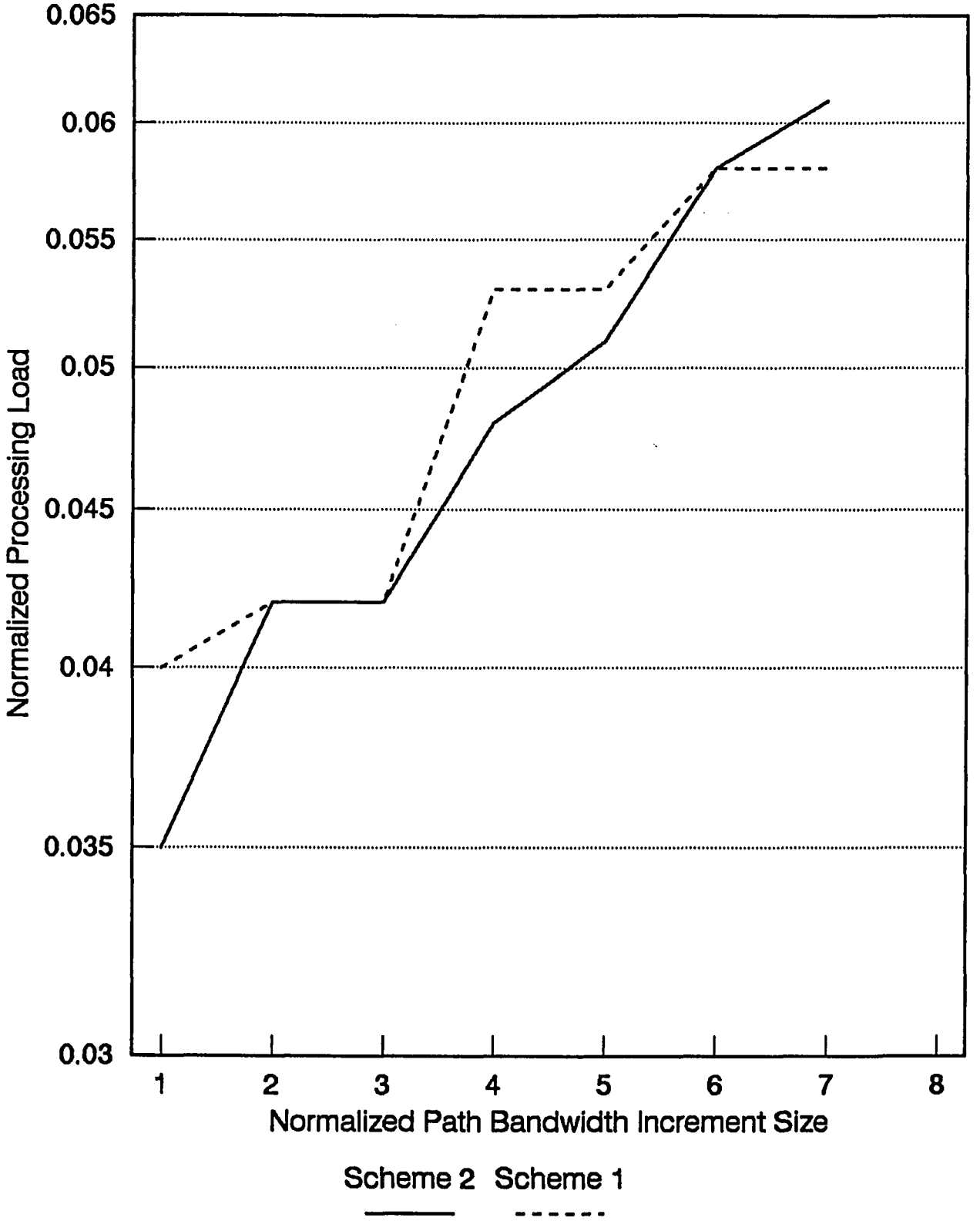


Fig.(II.4) Path Bandwidth Vs. Processing Load

Ref.[34]

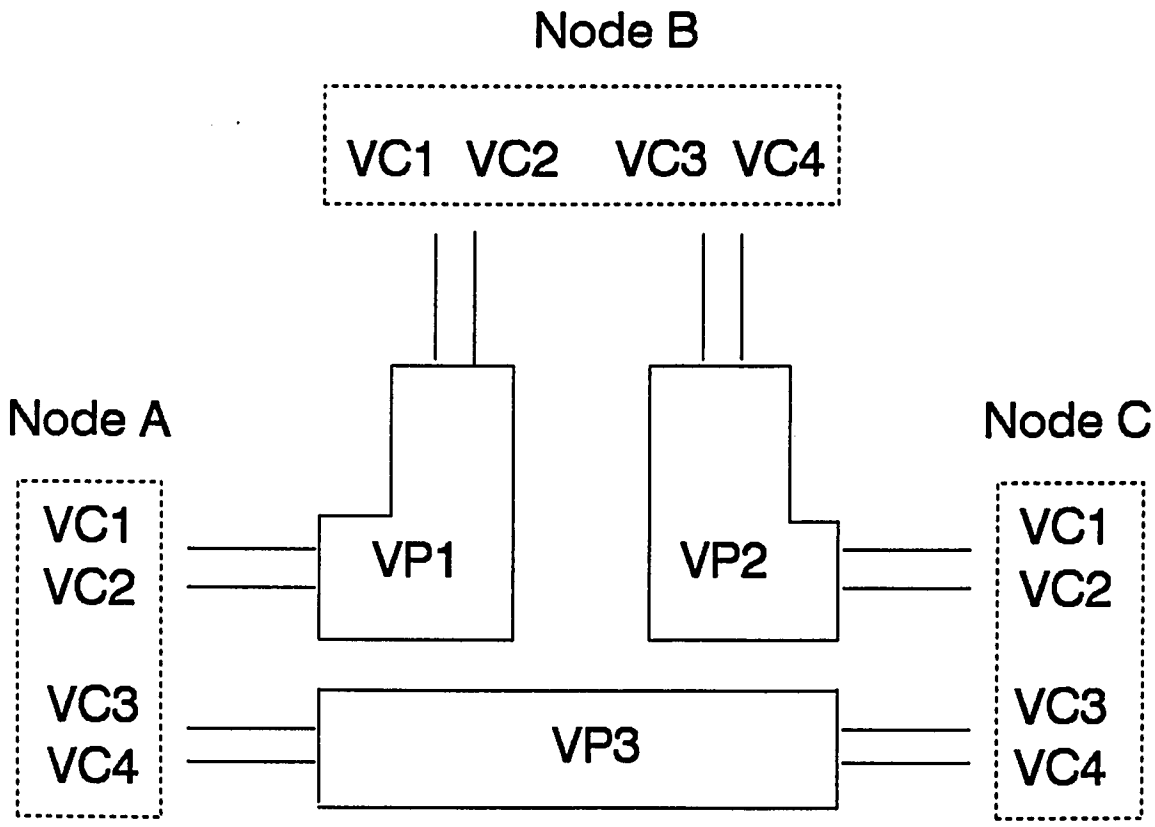
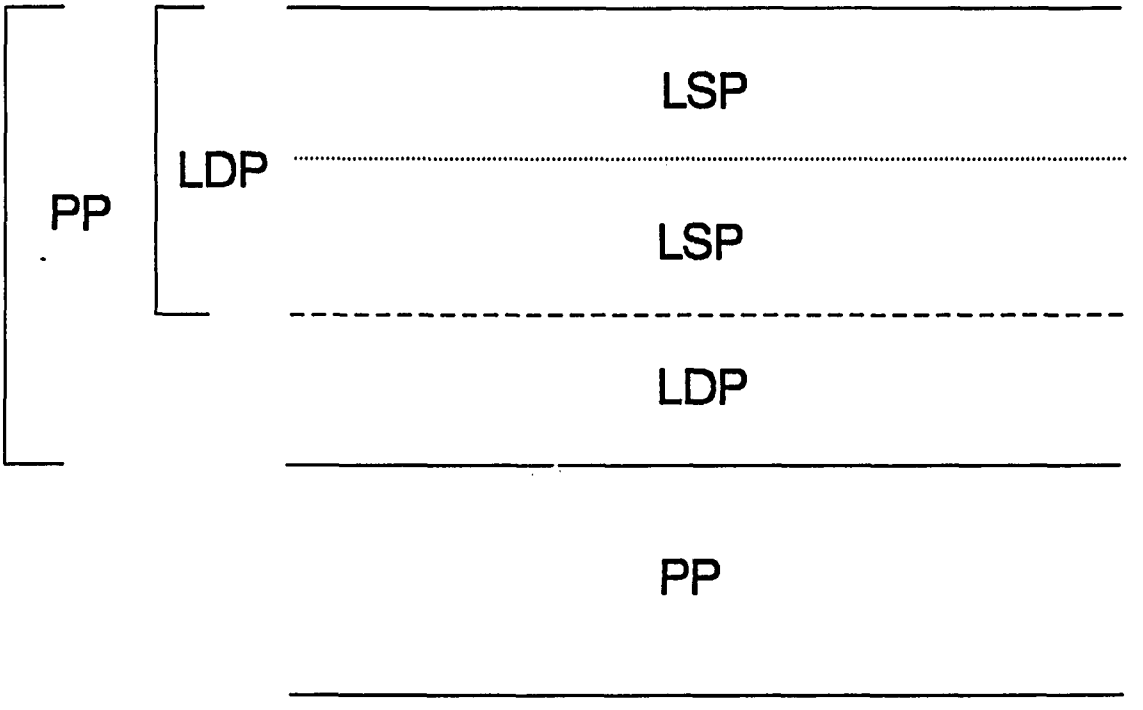


Fig.(II.5) Virtual Path Concept, Ref.[33]



LDP: Labelled Deterministic Path

LSP: Labelled Statistical Path

PP: Positioned Path

Fig.(II.6) Multilevel Path Structure, Ref.[33]

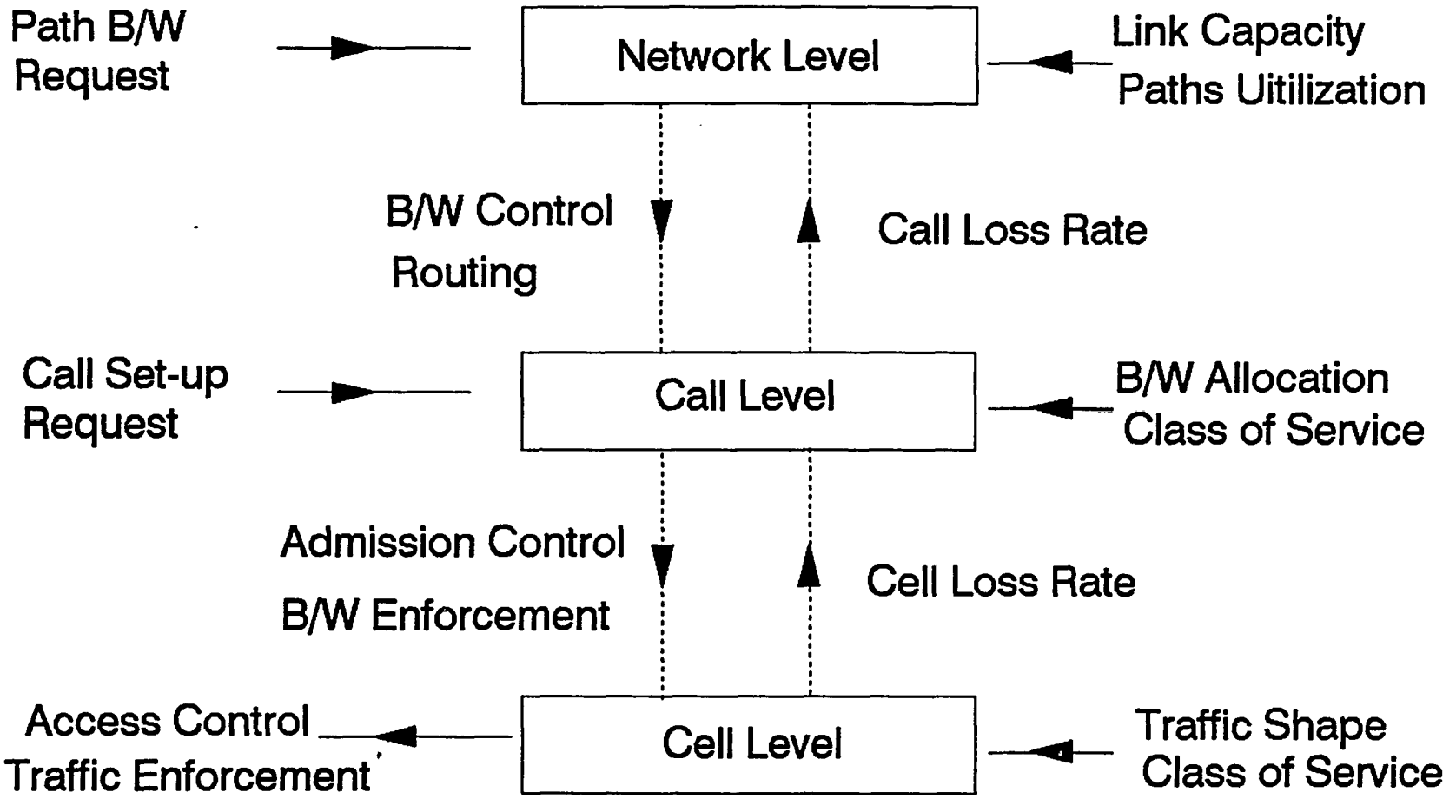


Fig.(II.7) ATM Multilevel Flow Control Model

III ACCESS FLOW CONTROL OF VOICE TRAFFIC

III.1. Background

Network resources, such as bandwidth, are allocated to each call at the call set up time and controlled over the logical connection during the call duration. Resources control in ATM, is therefore based upon end to end connection oriented type rather than connectionless one (although ATM can also support connectionless services through an overlaying adaptation layer). An important issue in ATM is the simple, fast and efficient transport of information units (called cells in ATM environment) without complicated flow control protocol functions, thus being flexible to support heterogeneous traffic. A fundamental characteristic of traffic propagating in BISDN environment is that the ratio of the propagation time to the cell transmission time is very high, thus end to end window type flow control is not useful. 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 links are underutilized.

The call acceptance decision is simply based upon the availability of bandwidth to support the call along its con-

nection path. As the traffic gets more bursty, more bandwidth is required to support the call at the same COS, 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

III.2. Multiplexer with Feedback Rate Control

Fig.(III.1) shows a schematic of the proposed scheme. It controls the input arrival process upstream the network (i.e., at the input multiplexer) based upon the feedback throttling of the arrival process. As the number of cells in the buffer reaches a first threshold level (K1), a feedback control signal, proportional to the congestion level of the multiplexer, is applied to the input source coder which controls the source rate via decreasing the coding rate (number of bits/sample). When another congestion threshold (K2) is achieved a similar action is taken, thus reducing the input rate further more. Since, at the access node, the delay between the source and the multiplexer is limited, any control signal will arrive at the source in time to throttle the traffic and avoid congestion. The control signals, in ATM networks, are transmitted as an out-of-band signalling and hence are given highest priority,

thus the control signal can be used to control sources connected directly to the ATM network, or to control sources connected via high speed LANs.

The threshold levels (K1,K2) are highly dependant upon the traffic characteristics and the required COS. The controller reads the traffic declared characteristics, such as the peak, average and coefficient of variation. Based upon the required COS, and the current link traffic situation (revealed via the cell loss rate) the controller then selects a certain set of threshold levels (K1,K2) to be used during the connection. The selection can be made using a Look-up table, or, better yet, the controller can be implemented using neural networks. The latter approach is more favorable, since there are so many input/output parameters involved in the decision and the application here, which is basically a pattern recognition one, is an ideal situation for swift control actions provided by neural networks.

The advantages of this scheme, are several. First, it prevents congestion from happening and greatly reduces the potential of congestion down stream the network, 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 higher perceived performance over FBR codes. Third, it provides the means for the maximum possible shaping of the input arrival process, without the need to design complicated window algorithms to regulate the traffic burstiness. Therefore, we can decrease the bandwidth allocated to the input call and achieve the same required class of service. Also, the statistical multiplexing gain is enhanced and we can achieve a significant improvement in the throughput. A major issue in ATM networks, is the price charged to the users. This scheme can be used to reduce the connection-cost, since more users will be able to share the allocated resources per connection. The price to be paid, is that we may perceive a slight degradation in the quality of the voice delivered. Interestingly, the degradation is imperceptible over most of the multiplexer utilization levels. This scheme deals with the problem of controlling bursty traffic via a two steps procedure, first it decreases the connection peak rate, second as a result of decreasing the peak arrival rate we can multiplex more connections and in doing so enhance the smoothing effect of the multiplexer

An obvious question now is how do we choose the control threshold levels subject to a certain COS requirement?. The answer is two fold. First it is quite clear that this choice must be function of the input bursty traffic, in the sense that as the input traffic gets more bursty (or correlated),

then the control must be applied sooner. Second, the choice be robust to the short term statistical variations that arise in the stochastic queueing process, in the mean time be sensitive enough to respond control the bursts of the input traffic. In essence the choice is then function of the peak input rate, duration of the burst, coefficient of variation and a control period (t secs.) In the following section we provide the modeling and the analysis. The exact analysis is a quite complicated transient time one, but to gain an insight of the system behavior we provide in the following section a steady state analysis.

III.3. Modeling and Analysis

Several models were introduced to model the burstiness and correlation characteristics of the cell generation rate from a voice source. The basic model is to represent that process by a periodic process alternating between a talkspurt and a silent period (fig.III.2). Each period is approximated by an exponential distribution of means $1/\alpha$ and $1/\beta$ secs. respectively [58], [59]. The number of cells generated within the talkspurt period is then a geometric multiple of the cell length. Each voice source is sampled at 8 KHz and encoded using Embedded ADPCM [67]. At a coding rate of 4 bits/sample we have a source rate of 32 Kbits/sec. However we can change

this rate by decreasing the coding word length from 4 bits/sample to 3 bits/sample, hence decreasing the effective arrival rate to 24 Kbits/sec. We can further decrease this rate by reducing the coding word length to 2 bits/sample only. The average arrival rate per source S (cells/sec.) is then given by

$$S = \frac{1/\alpha}{T(1/\alpha + 1/\beta)} \qquad (III.1)$$

The periodic process is then 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 (fig.III.3). The inherent correlations in the aggregate arrival process from R input sources, make it a complex nonrenewal process, hence we render to approximations. There has been a considerable amount of work in the literature in this area, and several mathematical models were proposed and analyzed (for example see [54]-[57], [75]-[76]). One approximation is to represent the aggregate arrival process by a continuous time phase process (fig.III.4), where the state of the chain represents the number of active sources [57]. Another possible representation of the aggregate arrival process is by a two state Modulated Markov Poisson Process (MMPP) (fig.III.5). 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 [77]-[81], there has been a significant development in the area of stochastic modeling of multimedia traffic sources, in [78] and [79] 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. In [80], [81] the Batch Markovian Arrival Process was introduced to model a variety of the versatile Markovian Point Process.

To model the multiplexer queueing process, 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. We used three different approximations, 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 [81]. In [75] and [76] it was shown that the randomness introduced by the approximation of the deterministic service time by an exponential one does not affect the queueing process, this is because the correlation effects resulting from the interaction between the cells interarrival times dominate the stochastic queue length behavior, especially in infinite buffer sizes. However in the ATM multiplexer under study, 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 we do not allow the build up of lengthy queues and the correlation effects of the cells interarrival times becomes less significant [51], [52], 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, we used six stages to limit the size of the resulting matrices. Finally in [51], the Poisson approximation was used based upon the above mentioned reasons, which are applicable in our case.

III.4. The $\hat{M}$/D/1/K Model

Let Q_j denotes the number of cells in the system at departure epoches (just after a departure). Let A_j be the number of arrivals in the jth service time. Then the following holds for the number of cells Q_j

$$Q_{j+1} = \text{Min}(N-1, Q_j - 1 + A_{j+1}), \quad Q_j > 0 \qquad (III.2)$$

$$Q_{j+1} = A_{j+1} \qquad Q_j = 0 \qquad (III.3)$$

The buffer size is N and the state space is limited to N-1 only, since at the departure epoches we can have at most N-1

packets in the system. Let the sequences a_i , b_i , c_i , denote the probabilities of number of arrivals during the deterministic service time D where

$$a_i = (\lambda_1 D)^i e^{-\lambda_1 D}/i!, \qquad i \geq 0 \qquad (III.4)$$

$$b_i = (\lambda_2 D)^i e^{-\lambda_2 D}/i!, \qquad i \geq 0 \qquad (III.5)$$

$$c_i = (\lambda_3 D)^i e^{-\lambda_3 D}/i!, \qquad i \geq 0 \qquad (III.6)$$

Where λ_i denotes the average arrival rate from R sources.

Let K1 and K2, be the threshold levels at which the flow control is activated 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. Define π_n to be π_n = Prob. $Q_j = n$ for (N-1 $\geq$ n $\geq$ 0). Then, the steady state transitional probability matrix is

	0	1	2	...	$K1$	$K1+1$	...	$K2$	$K2+1$	...	$N-2$	$N-1$
0	a_0	a_1	a_2	...	·	·	...	·	·	...	·	$1-\sum$
1	a_0	a_1	a_2	...	·	·	...	·	·	...	·	$1-\sum$
2	0	a_0	a_1	...	·	·	...	·	·	...	·	$1-\sum$
$K1$					a_1	a_2	...	·	·	...	·	·
$K1+1$					b_0	b_1	...	·	·	...	·	·
$K2$								b_1	b_2	...	·	·
$K2+1$								c_0	c_1	...	·	·
$N-2$											c_1	$1-\sum$
$N-1$											c_0	$1-c_0$

(*III*.7)

The $\sum$ is the sum of the elements of its respective row. The steady state probabilities can be easily computed from the given matrix. We compute the average arrival rate $\bar{\lambda}$ to be

$$\bar{\lambda} = \lambda_1 \sum_{n=0}^{K1} P_n + \lambda_2 \sum_{n=K1+1}^{K2} P_n + \lambda_3 \sum_{n=K2+1}^{N-1} P_n \qquad (III.8)$$

The transitional probability matrix describes, completely, the embedded Markov chain at the departure epochs. To calculate the probability of cell loss we have to find the steady state probabilities encountered by an arrival. Thus we need to expand the steady state space region to include case of n = N. Note that if an arrival encounters N cells in the system it will be turned away). Let P_n , $N-1 \geq n \geq 0$, denotes the steady state probabilities of the queue length encountered by an arrival such that it joins the queue. Thus the following relationship holds between P_n and π_n , see [82]

$$P_n = \pi_n/(1-P_N) \qquad (III.9)$$

$$P_N = (\pi_0+\rho-1)/(\pi_0+\rho) \qquad (III.10)$$

is the cell blocking probability. The system utilization ρ is defined by

$$\rho = \left(\lambda_1 \sum_{n=0}^{K1} P_n + \lambda_2 \sum_{n=K1+1}^{K2} P_n + \lambda_3 \sum_{n=K2}^{N-1} P_n\right)/\mu \qquad (III.11)$$

where $\mu = 1/D$ is the service rate. We calculate the mean bits per sample from

$$\overline{B} = 4\sum_{n=0}^{K1} P_n + 3\sum_{n=K1+1}^{K2} P_n + 2\sum_{K2+1}^{N-1} P_n \qquad (III.12)$$

III.5. The MMPP/M/1/K Model

The arrival process is the two state MMPP explained above where the arrivals are state dependant. Let the two input arrival phases be λ_H and λ_L , where the transitional rates are r_H and r_L respectively. At K1 each arrival rate drops to 0.75 of its value and at K2 drops to 0.5 of its original value. The state of the Markov chain is then denoted by the duple (i,j) where (i) is the level or the number of cells in the buffer and (j) keeps track of the arrival phase. The infinitesimal generator matrix Q is of size (2Nx2N) where N is the buffer size.

$$Q = \begin{pmatrix} A & B & & & & & & & & & \\ C & D & B & & & & & & & & \\ & C & D & B & & & & & & & \\ & & \cdot & \cdot & \cdot & & & & & & \\ & & & C & D & B & & & & & \\ & & & & C & E & B & & & & \\ & & & & & C & E & F & & & \\ & & & & & & \cdot & \cdot & \cdot & & \\ & & & & & & & C & E & F & \\ & & & & & & & & C & Y & F \\ & & & & & & & & & \cdot & \cdot \\ & & & & & & & & & & C \\ & & & & & & & & & & \\ & & & & & & & & & & \\ & & & & & & & & & & \end{pmatrix} \begin{matrix} \\ \\ \\ \\ \\ \\ \\ \\ \\ \\ \vdots \\ Y \\ C \end{matrix} \begin{matrix} \\ \\ \\ \\ \\ \\ \\ \\ \\ \\ \\ Z \\ W \end{matrix} \tag{III.13}$$

where the substochastic matrices are

$$A=\begin{pmatrix} -\lambda_L - r_L & r_L \\ r_H & -\lambda_H - r_H \end{pmatrix} \qquad (III.14)$$

$$B=\begin{pmatrix} \lambda_L & 0 \\ 0 & \lambda_H \end{pmatrix} \qquad (III.15)$$

$$C=\begin{pmatrix} \mu & 0 \\ 0 & \mu \end{pmatrix} \qquad (III.16)$$

$$D=\begin{pmatrix} -\lambda_L - r_L - \mu & r_L \\ r_H & -\lambda_H - r_H - \mu \end{pmatrix} \qquad (III.17)$$

$$E=\begin{pmatrix} -0.75\lambda_L - r_L - \mu & r_L \\ r_H & -0.75\lambda_H - r_H \end{pmatrix} \qquad (III.18)$$

$$F=\begin{pmatrix} 0.75\lambda_L & 0 \\ 0 & 0.75\lambda_H \end{pmatrix} \qquad (III.19)$$

$$Y=\begin{pmatrix} -0.5\lambda_L - r_L - \mu & r_L \\ r_H & -0.5\lambda_H - r_H - \mu \end{pmatrix} \qquad (III.20)$$

$$Z=\begin{pmatrix} 0.5\lambda_L & 0 \\ 0 & 0.5\lambda_H \end{pmatrix} \qquad (III.21)$$

$$W=\begin{pmatrix} -r_L - \mu & r_L \\ r_H & -r_H - \mu \end{pmatrix} \qquad (III.22)$$

At buffer length K1, K2 the control is activated and is accounted for by the change in the arrival rates in the sub-matrices as shown above. The matrix is then solved using matrix-geometric techniques, which are modified to suit the overload control case here, see [81]. Performance parameters is then evaluated and in this case the cell blocking probability is given by

$$P_N = \sum P_{N,j}, \qquad where\, j \in (L,H) \qquad (III.23)$$

III.6. The MMPP/Er/1/K Model

In this model, the service time is approximated by a 6-stage Erlang distribution with the same arrival process described above. The triplet (i,j,l) forms a Markov chain where (i) is the number of cells, (j) and (l) are arrival and service phases respectively. An infinitesimal generator matrix is then constructed with the same structure as the matrix given in the previous section where now we have

$$A = \begin{pmatrix} -\lambda_L - r_L & r_L \\ r_H & -\lambda_H - r_H \end{pmatrix} \qquad (III.24)$$

$$B=\begin{pmatrix}\lambda_L & 0 & 0 & 0 & 0 & 0 & & & & & & \\ & & & & & & \lambda_H & 0 & 0 & 0 & 0 & 0\end{pmatrix} \qquad (III.25)$$

$$C=Transpose\begin{pmatrix}0 & 0 & 0 & 0 & 0 & \mu & & & & & & \\ & & & & & & 0 & 0 & 0 & 0 & 0 & \mu\end{pmatrix} \qquad (III.26)$$

$$D=\begin{pmatrix}
-\lambda_L-r_L-\mu & \mu & 0 & 0 & 0 & 0 & r_L & & & & & \\
0 & -\lambda_L-r_L-\mu & \mu & 0 & 0 & 0 & 0 & r_L & & & & \\
0 & 0 & \cdot & \cdot & 0 & 0 & 0 & 0 & \cdot & & & \\
0 & 0 & 0 & \cdot & \cdot & 0 & 0 & 0 & 0 & \cdot & & \\
0 & 0 & 0 & 0 & \cdot & \cdot & 0 & 0 & 0 & 0 & \cdot & \\
0 & 0 & 0 & 0 & 0 & -\lambda_L-r_L-\mu & 0 & 0 & 0 & 0 & 0 & r_L \\
-\lambda_H-r_H-\mu & \mu & 0 & 0 & 0 & 0 & r_H & & & & & \\
0 & -\lambda_H-r_H-\mu & \mu & 0 & 0 & 0 & 0 & r_H & & & & \\
0 & 0 & \cdot & \cdot & 0 & 0 & 0 & 0 & \cdot & & & \\
0 & 0 & 0 & \cdot & \cdot & 0 & 0 & 0 & 0 & \cdot & & \\
0 & 0 & 0 & 0 & \cdot & \cdot & 0 & 0 & 0 & 0 & \cdot & \\
0 & 0 & 0 & 0 & 0 & -\lambda_H-r_H-\mu & 0 & 0 & 0 & 0 & 0 & r_H
\end{pmatrix}$$

$$(III.27)$$

The **E, Y** matrices have the same structure as the **D** matrix with the appropriate arrival rates at threshold levels K1 and K2. The **W** matrix has the same structure as the **D** matrix, however the diagonal elements do not contain the λterms. Similarly, the **F** and **Z** matrices have the same structure as the **B** matrix with the appropriate λs. The above matrix is then solved using same techniques, outlined before, and the cell blocking probability is computed from

$$P_N = \sum P_{Njl}, \qquad j \in (H, L), l \in (1-6) \qquad (III.28)$$

III.7. Numerical Results and Conclusions

In this chapter, the cell size is the ATM standard (48 bytes of information plus 5 bytes for the header). The source rate is 32 Kbits/sec, while the line capacity is assumed to be 150 Mbits/sec. The cell time is then 12 msec. while the service time is approximately 3 μ secs and the buffer size is set to 20. At K1 control level, the number of bits /sample drops from 4 to 3 bits/sample. At K2, the number drops further from 3 to 2 bits/sample only, thus throttling the peak arrival rate. We studied several burstiness parameters in order to examine the effects on the choice of the control thresholds K1 and K2. The first burstiness parameters set had $1/\alpha = 352ms.$ and $1/\beta = 650ms.$ which corresponds to a 35% activity factor. The second set had an activity factor of 49%, while the third set had a 60% activity factor. The fourth set had a peak bit rate of 64 Kbits/sec. with the same activity factor as set (1). Fig. (III.6) shows a comparison of the different approximation models, as explained in section III the MMPP/M/1/K overestimates the blocking probability while the Poisson assumption yields quite accurate results. The results get even more appreciable when the flow control is applied.

Fig.(III.7) shows a significant improvement in the system performance when the control is applied. As expected the performance gets better as the control thresholds gets smaller, i.e. the control algorithm is invoked earlier. The results confirm our discussions in sections II and III. The multiplexer can support sources such that the utilization is more than one and with the required COS. This is, merely, due to the enhancement of the statistical multiplexing gain through decreasing the input source peak rate, hence making it possible to support more sources at the same bandwidth and avoiding possible congestion and severe cell loss. Fig.(III.8) shows that the price to be paid is a very slight degradation in the voice quality expressed through the drop of the mean bits/sample. This effect, will set a limit on the gained performance since the COS determines the minimum voice quality. Figs.(III.9) to (III.14) show the effect of the traffic burstiness, as the burstiness increases the network can not support the same number of sources at the same quality. To achieve that, the control threshold levels must be reduced, however there is a limit set by the COS. If the COS is not met, then the bandwidth allocated must be increased. The burstiness effect is very effective, when the peak rate is high (compare the results of burstiness sets 1 and 4).

We, now, summarize a two step flow control action taken by the node controller:

1. The controller, will check the traffic burstiness level, through the declared traffic descriptors. The required COS is also declared by the user. Based upon monitoring the current node utilization and link traffic status, the controller will then activate (or not activate) the control thresholds. If the user violates the declared statistics, the network tries to accommodate the new increase in the load by activating the flow control algorithm.

2. If the COS can not be met, then the controller asks the call level control function for an increase in the bandwidth assigned (this increase is quite small since the control algorithm is activated). If the increase is not honored then the network management level tries to reassign the logical link capacity extra bandwidth. Finally, alternate routing is also checked, after which the call is disconnected.

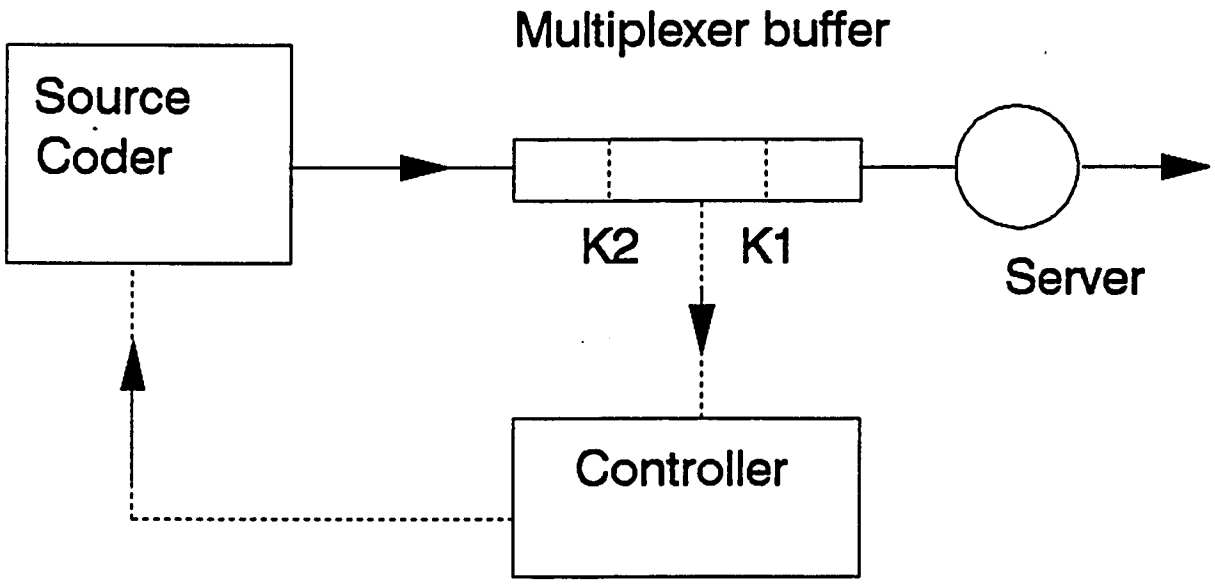


Fig.(III.1) Multiplexer with Feedback Rate Control

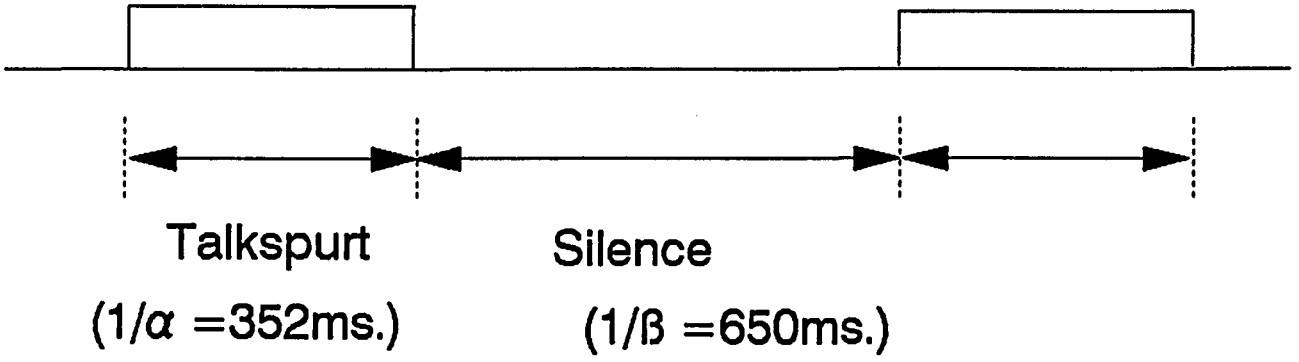


Fig.(III.2) Single Voice Source Model

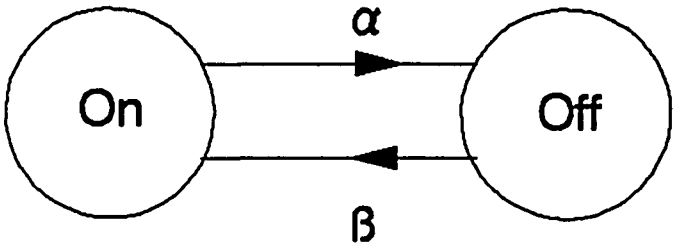
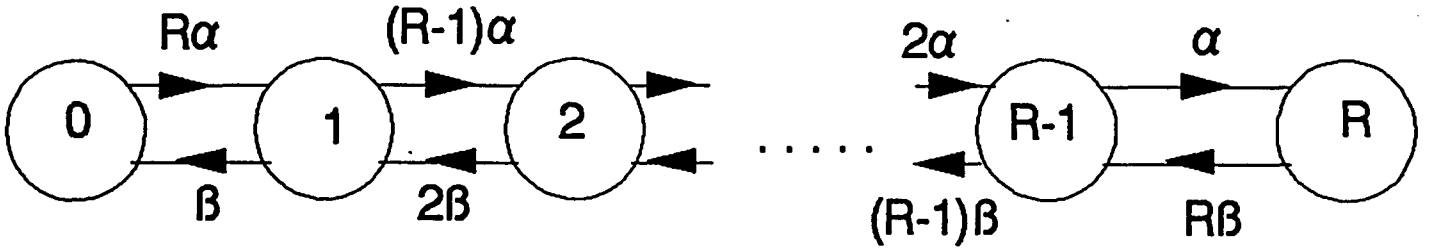


Fig.(III.3) Two State Continuous Time Markov Chain



R is the Number of Active Sources

Fig.(III.4) Superposition Arrival Process Model

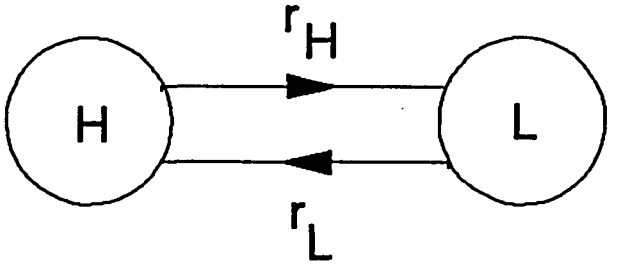


Fig.(III.5) Equivalent Two State MMPP Model

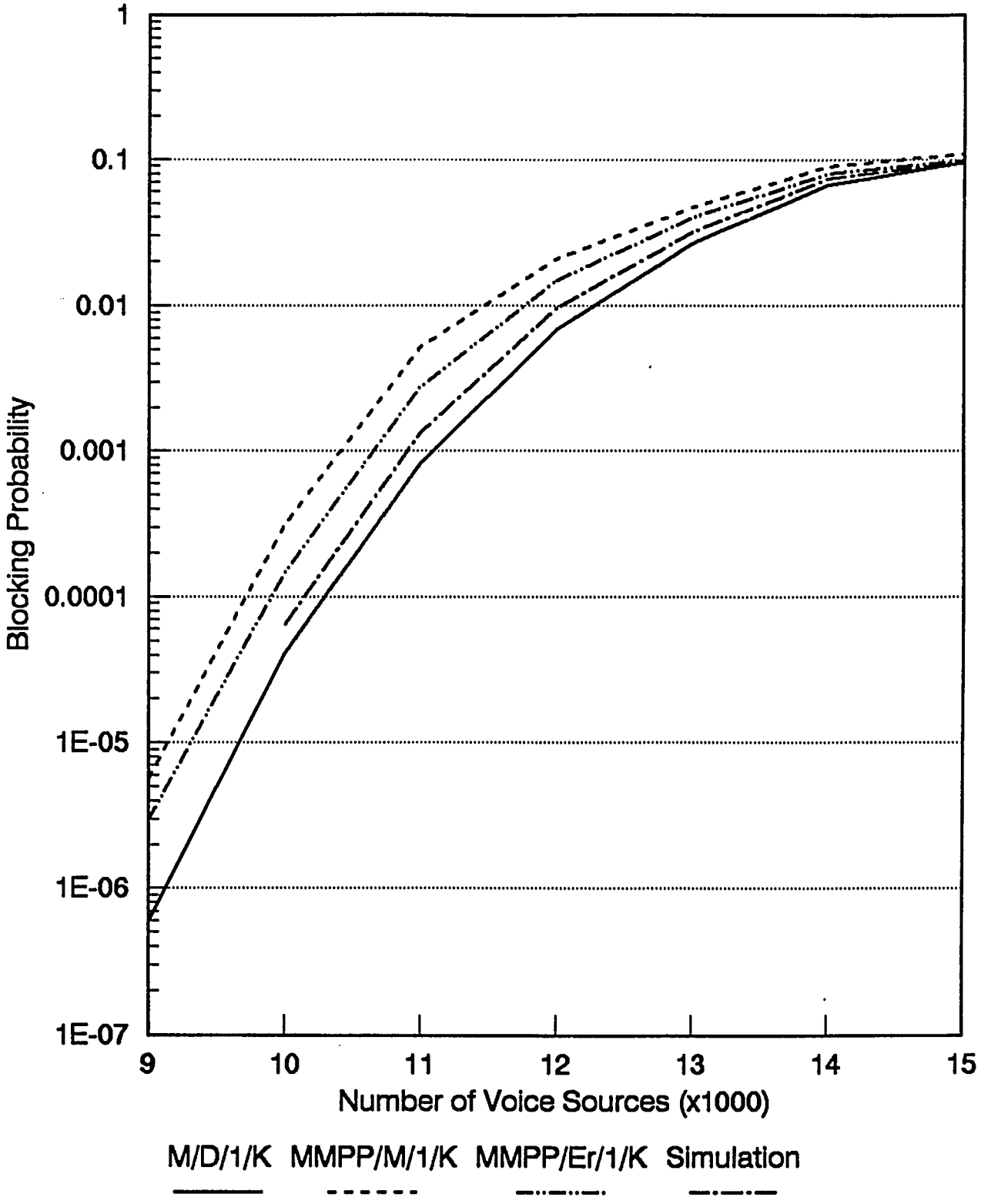


Fig.(III.6) Comparison of Approximation Methods
No Control

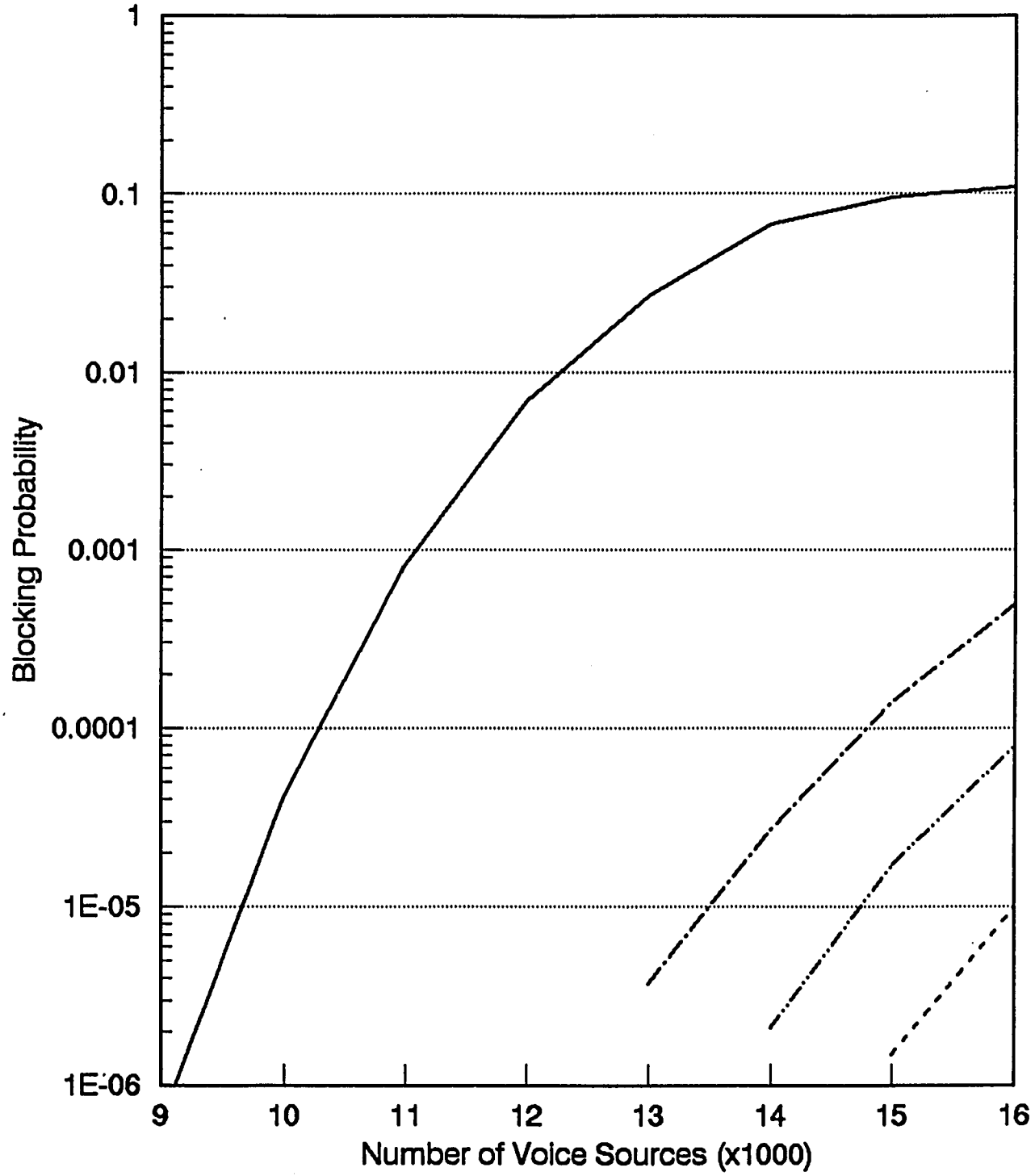


Fig.(III.7) Blocking Probability Vs. Load Burstiness Set (1)

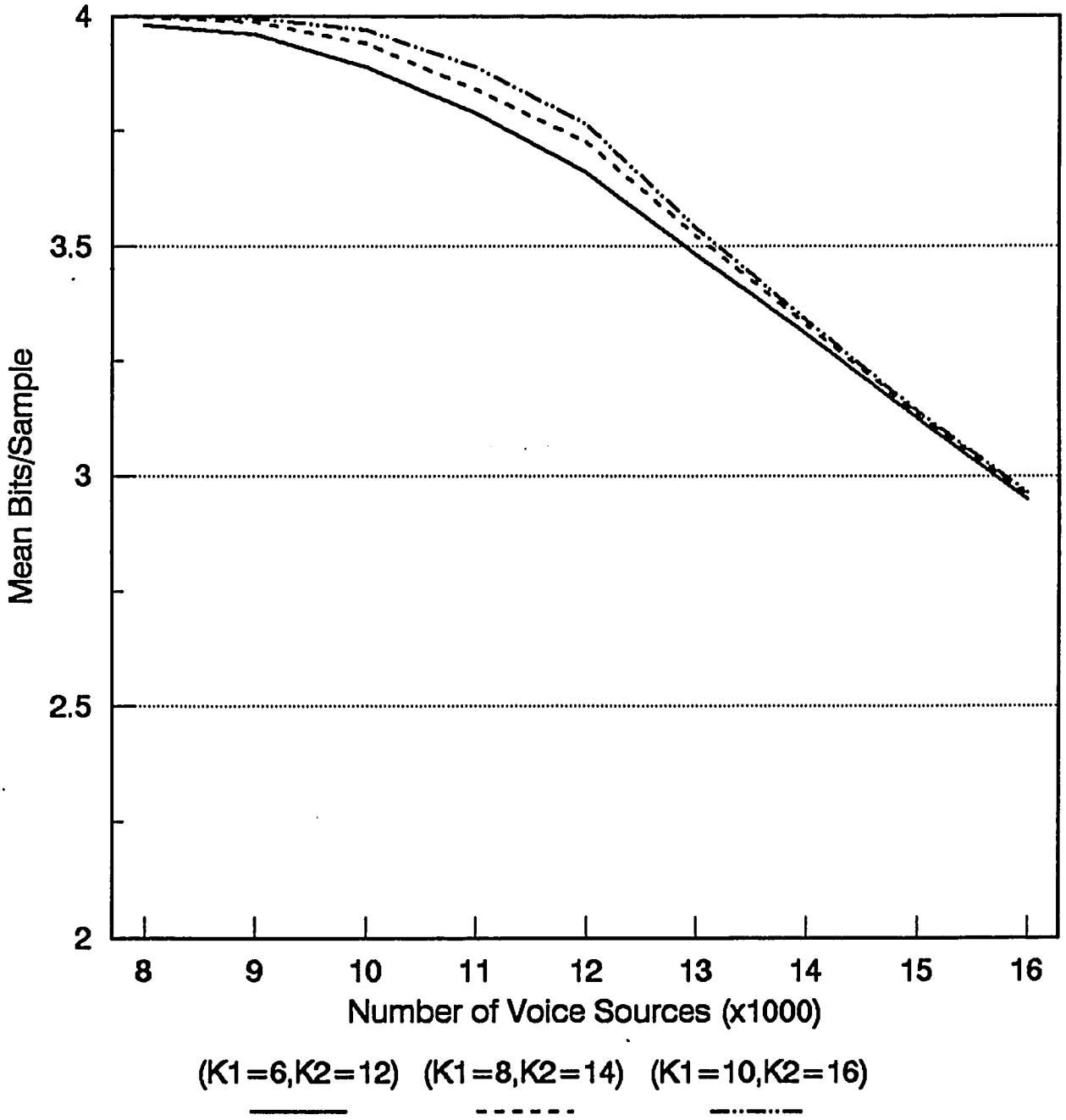


Fig.(III.8) Voice Quality Vs. Load Burstiness Set (1)

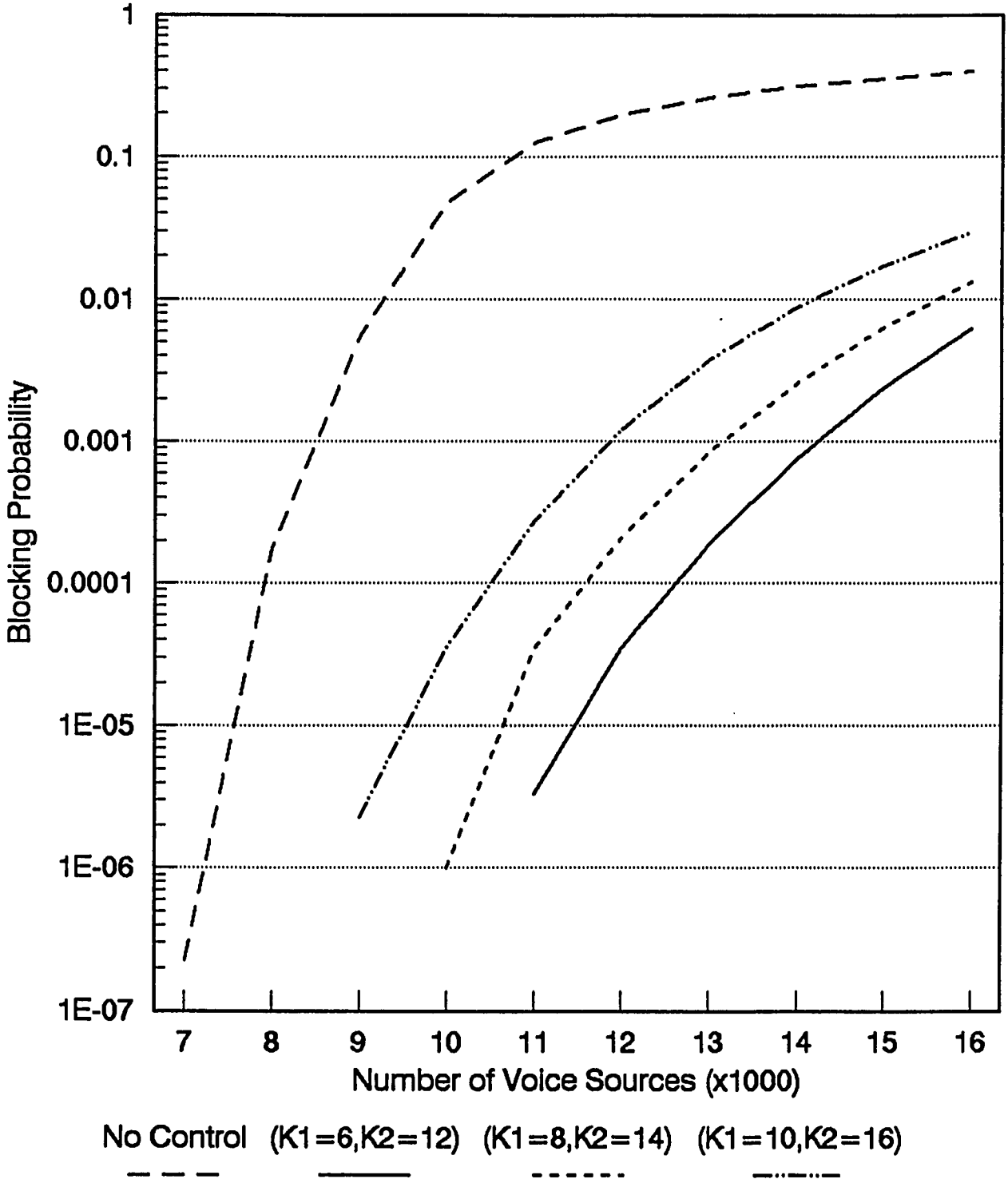


Fig.(III.9) Blocking Probability Vs. Load Burstiness Set (2)

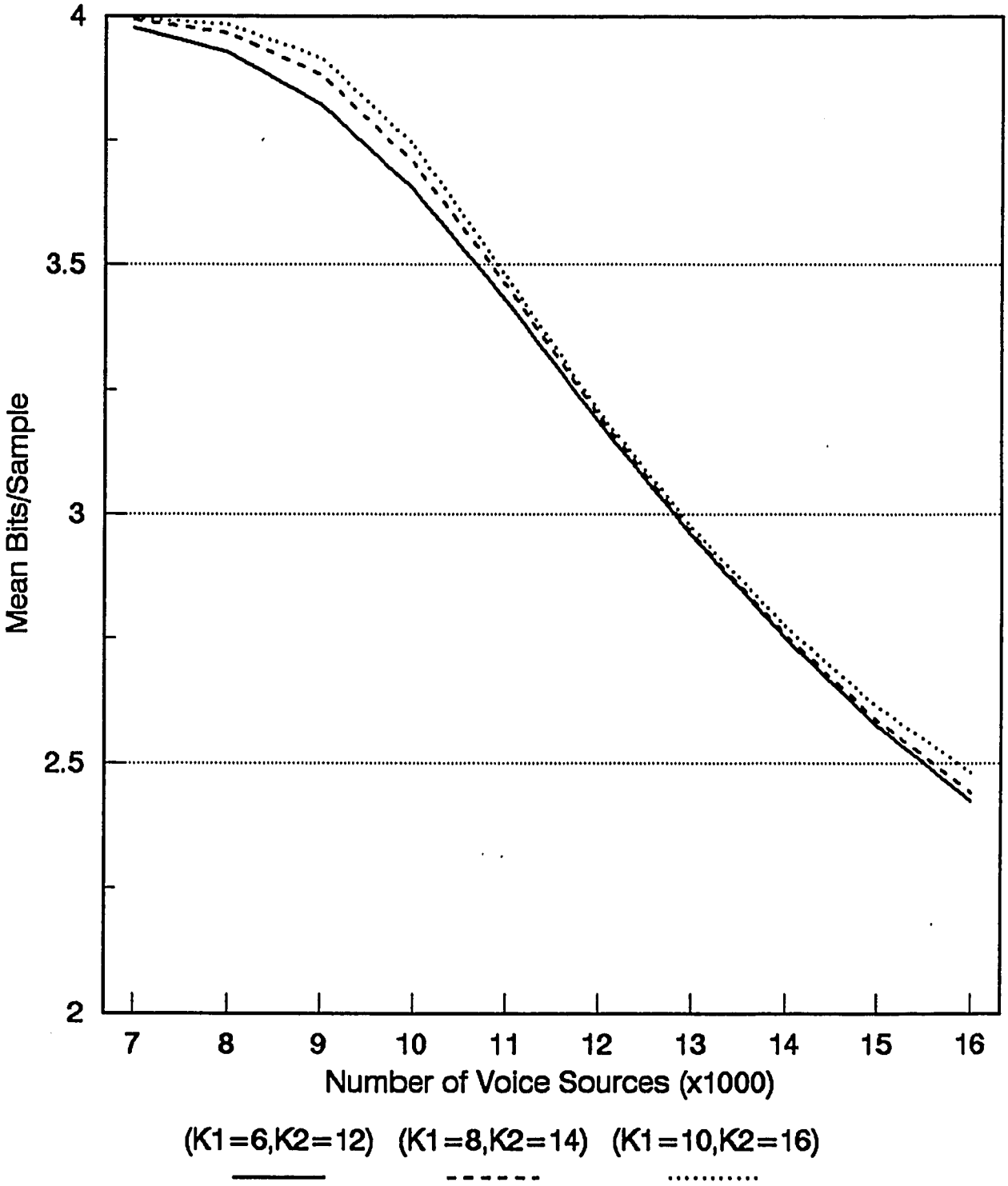


Fig.(III.10) Voice Quality Vs. Load Burstiness Set (2)

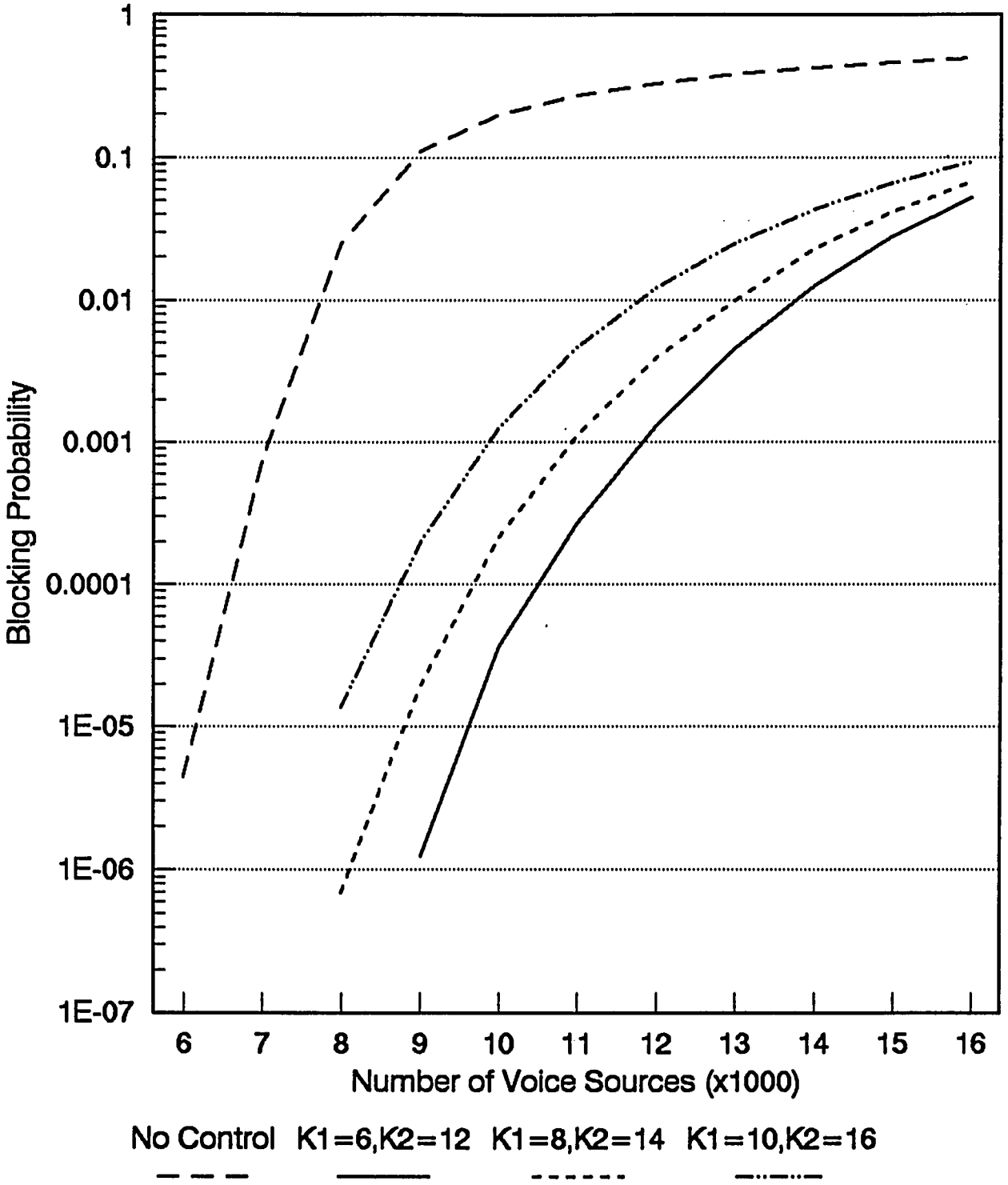


Fig.(III.11) Blocking Probability Vs. Load Burstiness Set (3)

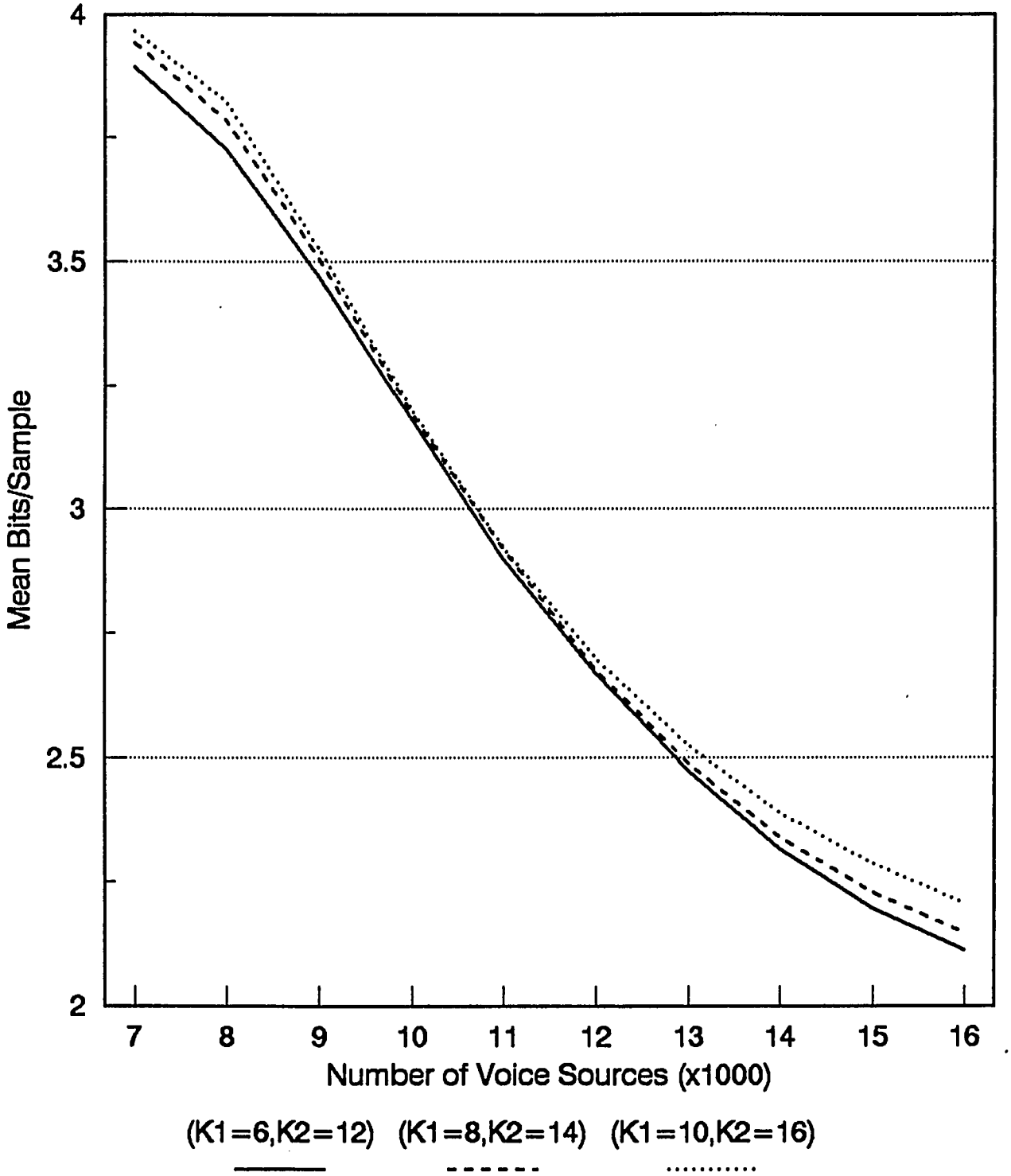


Fig.(III.12) Voice Quality Vs. Load Burstiness Set (3)

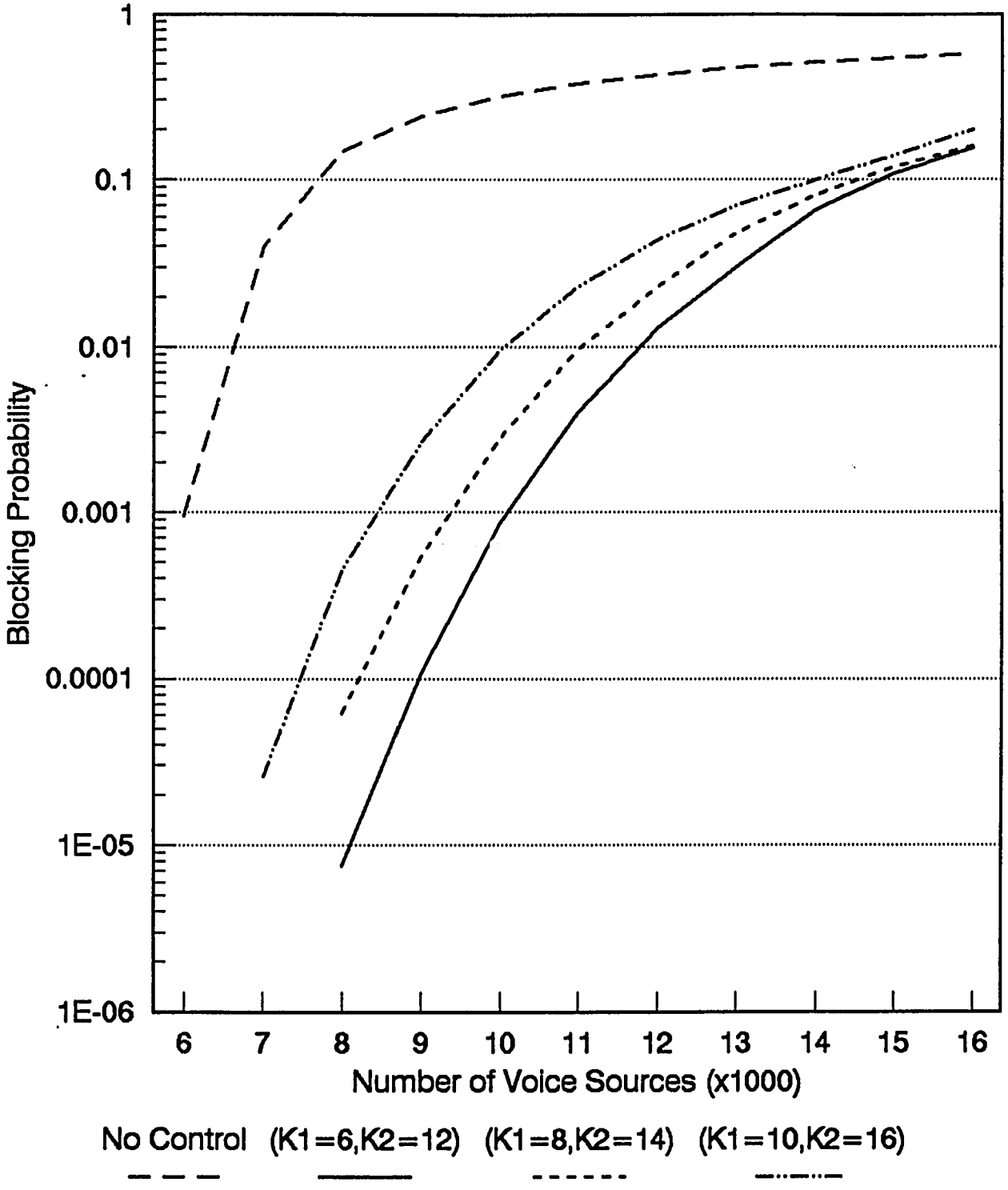


Fig.(III.13) Blocking Probability Vs. Load Burstiness Set (4)

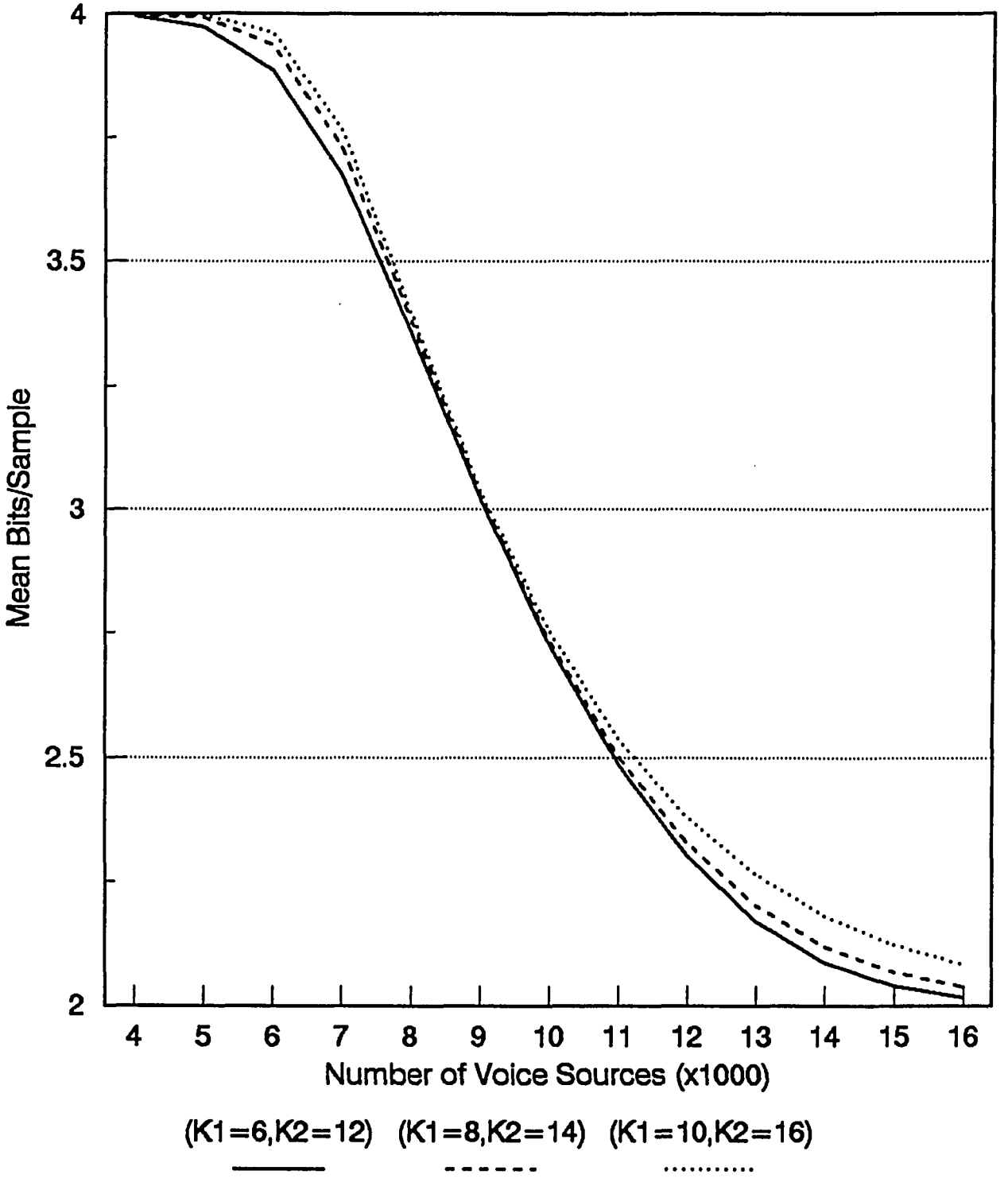


Fig.(III.14) Voice Quality Vs. Load Burstiness Set (4)

IV ACCESS FLOW CONTROL OF VIDEO TRAFFIC

IV.1. Variable Bit Rate Video Coders

In Broadband Networks, flow control functions are shifted to the edges of the network, and implemented on an end to end basis. Access flow control is then essential to avoid congestion. In this chapter, we apply the access flow control algorithm, described in chapter III, to the video multiplexer case. In the video multiplexer case, we have to model the video source and understand its functionality. In our analysis, we assume a video source with a small scene changes such as video conference or head and shoulder videophone types. Variable bit rate (VBR) coding is quite attractive to use in the ATM environment. ATM provides us with the flexibility needed to support variable bit rate codes. VBR coding can utilize the flexibility in bandwidth assignment and produce selectable picture quality irrespective of the rapid scene motion. VBR delivers an overall improved picture quality when compared to fixed bit rate (FBR) codes at the same average rate. The reason is that FBR coding requires a buffer which causes delay and quality degradation during active motion scenes.

VBR techniques, also, improve the channel transmission efficiency through the statistical multiplexing gain and the possibility of sharing the channel among several users. VBR techniques are very powerful in compressing the data transmission rates down to reasonable values suitable for transmission over high speed networks. Sub-band coding is very attractive method to be used, since it is easier to retrieve information loss in this method than any other lossy compression technique. The problem still persists, which is the high degree of burstiness that the traffic exhibit using VBR techniques. In VBR methods, the difference in information, between one frame and the previous one, is transmitted as a burst at the beginning of the following consecutive frame. Also, the peak bit rate of the traffic is rather high, and thus does not allow for the superposition of a large number of sources. It is well known that the multiplexing of a large number of video sources would yield a superposition stream that is more well well behaved and smoother than that of a single video source.

The design of large buffers to absorb the long bursts of the video traffic does not solve the problem, since we can not guarantee a certain burst duration. Consequently, the best solution is to smooth down the characteristics, via controlling the peak rate and hence the possibility of multiplexing a relatively larger number of sources such that the multiplexing

gain is enhanced. In the following sections, we shall apply the same method of controlling the peak arrival rate. It is worth mentioning that the control signal uses out-of-band signalling, hence making it possible to apply it to video sources connected to the ATM network via high speed local area networks.

IV.2. Modeling and Analysis

Fig.(IV.1) presents the continuous time markov chain model of the video source, also known as the phase process. In [53], this model was presented and analyzed to match the statistics of the continuous state autoregressive model. The autoregressive model is quite accurate in modeling the video source statistics. However , it is quite complicated if we try to use it in analytical studies. In [66] a similar model was used to analyze the video statistics. In [60]-[63] a (J-state) Markov modulated process was used to study the performance of the video multiplexer.

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. These parameters were evaluated in [53] by fitting them to the average, variance and the autocovariance function of the original measured data of the source statistics. The results are

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

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

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

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 autocovariance curve, was proved by several authors, to follow an exponential fit. The value (3.9) was found to match the variations of this specific video experiment. The number of states M was set to be 10 R. It was found in [53] that this value of M had yielded reasonable results that were close enough to the measured data. In this paper, all the variables

used were normalized to cells/millisec.

The multiplexer buffer has a fixed buffer length N cells and is fed by the process described by the system of equations in (IV.1)-(IV.3). Fig.(IV.2), shows the two dimension continuous time Markov chain model which describes the system behavior. The queue length stochastic process is a Markov process one, at instants of state changes. Each discrete level arrival to the queue, is a Poisson process with exponential service time with mean μ, where $\mu = L/C$. C is the link capacity in bits/sec., and L is the cell length in bits. The cell fixed service time is replaced by an exponential service time. The work done in [76], proved that replacing the service fixed time by an exponential one does not affect the queueing process, since the correlations effects of the arrival process dominates those of the service time process. Each state of the phase process iA, ($(0 \leq i \leq M)$), is therefore the equivalent of i sources each has a Poisson arrival process of rate A cells/millisec. The transitional rates, between the system states, are thus Poisson, of rate iA. Let K1, K2, be the queue lengths at which the flow control is activated. If the queue length reaches the threshold limit K1, the rates drop to iB, and at queue length K2, the rates drop to iC. The rates B, C, represent the arrival rates after decreasing the number of bits/sample of the source coder. In our analysis, we set those values to be 0.75A and 0.5A respectively.

Let the duple $\{Q, i\hat{A}\}$, where $(\hat{A} \in A, B, C)$, denote the number of cells in the queue and the phase of the arrival process respectively. Then the stochastic equilibrium probability of the system is

$$P_{x,y} = Pr.\{Q = x, i\hat{A} = y\}, \quad (0 \le x \le N, iC \le y \le iA), \quad (0 \le i \le M) \quad (IV.4)$$

We can write the following equations for the system

$$M\alpha P_{0,0} = \mu P_{1,0} + \beta P_{0,A} \qquad for(x = 0, y = 0) \qquad (IV.5)$$

$$\{(M-i)\alpha + iA + i\beta\} P_{0,iA} =$$

$$\mu P_{1,iA} + (M-i+1)\alpha P_{0,(i-1)A} + (i+1)\beta P_{0,(i+1)A}$$

$$for(1 \le i \le M-1) \qquad (IV.6)$$

$$(M\beta + MA) P_{0,MA} = \mu P_{1,MA} + \alpha P_{0,(M-1)A} \qquad (IV.7)$$

for $(1 \le x \le Kl)$

$$(M\alpha + \mu) P_{x,0} = \mu P_{x+1,0} + \beta P_{x,A} \qquad (IV.8)$$

$$\{(M-i)\alpha + i\beta + iA + \mu\} P_{x,iA} =$$

$$(M-i+1)\alpha P_{x,(i-1)A} + \mu P_{x+1,iB} + (i+1)\beta P_{x,(i+1)A} + iA P_{x-1,iA},$$

$$for(1 \le i \le M-1) \qquad (IV.9)$$

$$(M\beta + \mu) P_{x,MA} = \alpha P_{x,(M-1)A} + MA P_{x-1,MA} + \mu P_{x+1,MB} \qquad (IV.10)$$

for $(K1+1 \le x \le K2)$

$$(M\alpha+\mu)P_{x,0} = \mu P_{x+1,0} + \beta P_{x,B} \qquad (IV.11)$$

$$\{(M-i)\alpha + i\beta + iB + \mu\}P_{x,iB} =$$

$$(M-i+1)\alpha P_{x,(i-1)B} + \mu P_{x+1,iC} + (i+1)\beta P_{x,(i+1)B} + iB P_{x-1,iB},$$

$$for(1 \le i \le M-1) \qquad (IV.12)$$

$$(M\beta+\mu)P_{x,MB} = \alpha P_{x,(M-1)B} + MB P_{x-1,MB} + \mu P_{x+1,MC} \qquad (IV.13)$$

for $(K2+1 \le x \le N-1)$

$$(M\alpha+\mu)P_{x,0} = \mu P_{x+1,0} + \beta P_{x,C} \qquad (IV.14)$$

$$\{(M-i)\alpha + i\beta + iC + \mu\}P_{x,iC} =$$

$$(M-i+1)\alpha P_{x,(i-1)C} + \mu P_{x+1,iC} + (i+1)\beta P_{x,(i+1)C} + iC P_{x-1,iC}$$

$$for(1 \le i \le M-1) \qquad (IV.15)$$

$$(M\beta+\mu)P_{x,MC} = \alpha P_{x,(M-1)C} + MC P_{x-1,MC} + \mu P_{x+1,MC} \qquad (IV.16)$$

for $x = N$

$$(M\alpha+\mu)P_{N,0} = \beta P_{N,C} \qquad (IV.17)$$

$$\{(M-i)\alpha + i\beta + iC + \mu\}P_{x,iC} =$$

$$(M-i+1)\alpha P_{x,(i-1)C} + \mu P_{x+1,iC} + (i+1)\beta P_{x,(i+1)C} + iC P_{x-1,iC}$$

$$for(1 \le i \le M-1) \qquad (IV.18)$$

$$(M\beta+\mu)P_{x,MC} = \alpha P_{x,(M-1)C} + MC P_{x-1,MC} + \mu P_{x+1,MC} \qquad (IV.19)$$

The above equations can be written in the matrix form $PQ = 0$ where P is the steady state probabilities vector, Q is the transitional rates matrix, then we have

$$Q = \begin{pmatrix} T(0) & \Lambda & & & & & & & & & \\ S & T(1) & \Lambda & & & & & & & & \\ & \ddots & \ddots & \ddots & & & & & & & \\ & & S & T(K1) & \Lambda & & & & & & \\ & & & S & T(K1) & \Lambda(K1) & & & & & \\ & & & & \ddots & \ddots & \ddots & & & & \\ & & & & & & S & T(K2) & \Lambda(K1) & & \\ & & & & & & & S & T(K2) & \Lambda(K2) & \\ & & & & & & & & \ddots & \ddots & \ddots \\ & & & & & & & & & S & T(N) \end{pmatrix} \qquad (IV.20)$$

$T(0), T(1), T(K1), T(K2), T(N), S, \Lambda, \Lambda(K1), \Lambda(K2)$ are all square matrices of the dimension M, where

$$T(0) = \begin{pmatrix} -M\alpha & M\alpha & & & & \\ \beta & -(M-1)\alpha-\beta-A & (M-1)\alpha & & & \\ & \ddots & \ddots & \ddots & & \\ & & \ddots & \ddots & \ddots & \\ & & & \ddots & \ddots & \ddots \\ & & & & M\beta & -M\beta-MA \end{pmatrix} \qquad (IV.21)$$

$$T(1)=\begin{pmatrix} -M\alpha-\mu & M\alpha & & & & \\ \beta & -(M-1)\alpha-\beta-A-\mu & (M-1)\alpha & & & \\ & \cdot & \cdot & \cdot & & \\ & & \cdot & \cdot & \cdot & \\ & & & \cdot & \cdot & \cdot \\ & & & & M\beta & -M\beta-MA-\mu \end{pmatrix}$$

(*IV*.22)

$$T(N)=\begin{pmatrix} -M\alpha-\mu & M\alpha & & & & \\ \beta & -(M-1)\alpha-\beta-\mu & (M-1)\alpha & & & \\ & \cdot & \cdot & \cdot & & \\ & & \cdot & \cdot & \cdot & \\ & & & \cdot & \cdot & \cdot \\ & & & & M\beta & -M\beta-\mu \end{pmatrix}$$

(*IV*.23)

S is a diagonal matrix where μ is its diagonal element
Λ is also a diagonal matrix, where iA is its diagonal element. The matrices $T(K1),T(K2)$ are of the same structure as $T(1)$ with the appropriate scaled λs. Similarly $\Lambda(K1),\Lambda(K2)$, are the same as Λ but with the appropriate scaled λs at each corresponding threshold levels K1 and K2.

The above 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. We used matrix-geometric techniques, introduced in [81], 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 our case. The details are not repeated here and the reader is referred to [81]. Finally, the above equations were solved numerically for $P_{x,y}$. The blocking probability P_B is calculated from

$$P_B = \sum_{i=0}^{M} P_{N,iC} \qquad (IV.24)$$

IV.3. Numerical results and Conclusions

The video source characteristics reported in sec. (IV.2), had an average arrival rate of 3.9 Mbits/sec. and a peak rate of 11 Mbits/sec. We used a buffer length of 20 cells in order to limit the delay to 50 μsecs., where the cell length is the ATM standard of 53 bytes and C is 150 Mbits/sec. Fig.(IV.3), reflects the improvement in the multiplexer performance for different number of sources. The blocking probability has dropped significantly, as a result of applying the flow control technique. As the buffer size is relatively small, the statistical multiplexing gain is not very appreciable. Figs.(IV.4) and (IV.5), illustrate more clearly this effect. The trend

is clear, as the buffer size is increased, the statistical multiplexing gain becomes more effective. However, as expected, the flow control technique has significantly enhanced the multiplexing gain.

Figs.(IV.6) to (IV.9), compare the performance of the multiplexer for different flow control thresholds. As the threshold levels are decreased, the performance improves, however the price will be a slight degradation of the image quality. The results that are reported here, confirm our earlier discussions. It proves the importance of applying this type of flow control technique in order to accommodate sources with high peak rates without sacrificing the efficiency. Figs.(IV.10) to (IV.15) report the results of the performance analysis for different number of sources. It is clear from comparing the reported results that the statistical multiplexing is enhanced as the number of sources is increased. Although it is rather difficult to assess the effect of the access flow control on the image quality without subjective quality tests. We reported the variation of the mean number of bits/sample over the utilization load. The drop in the image quality is more perceived in the video multiplexer case than the voice multiplexer case. The reason, of course, is the increase in the correlation effects between the successive interarrival cell times in the video case, and the higher degree of burstiness.

In our work, we have placed a stringent delay limit of 50 μ secs. However, we can relax this value to 100 μ secs without having a major effect on the delay requirement. Therefore, we can double the buffer size, and the statistical multiplexing gain will be more effective. The obvious result of decreasing the blocking probability will have the direct impact of minimizing serious congestion problems. Thus, we can efficiently, utilize the network valuable resources such as bandwidth and, in the mean time, provide different users with the required performance. Another impact, is that we can accommodate more sources at the same bandwidth, when flow control is not used. It had been suggested, in [25], that for bursty traffic with high peak to link ratio, the non-statistical operation mode could be more effective. We believe that, with our proposed flow control algorithm, we can utilize the VBR coding techniques to operate within the statistical multiplexing region and with high efficiency.

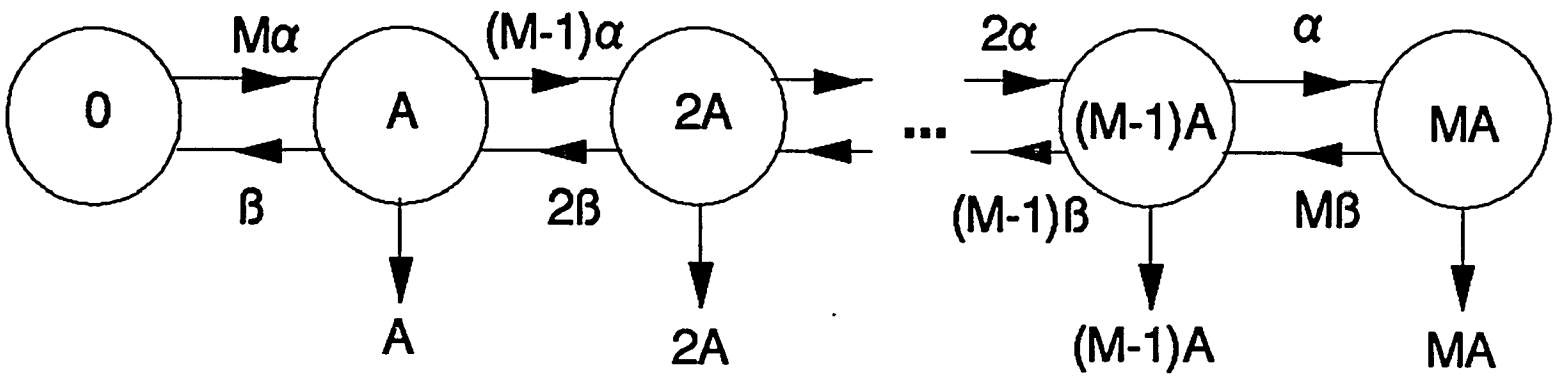


Fig.(IV.1) Single Video Source Model (Phase Process)

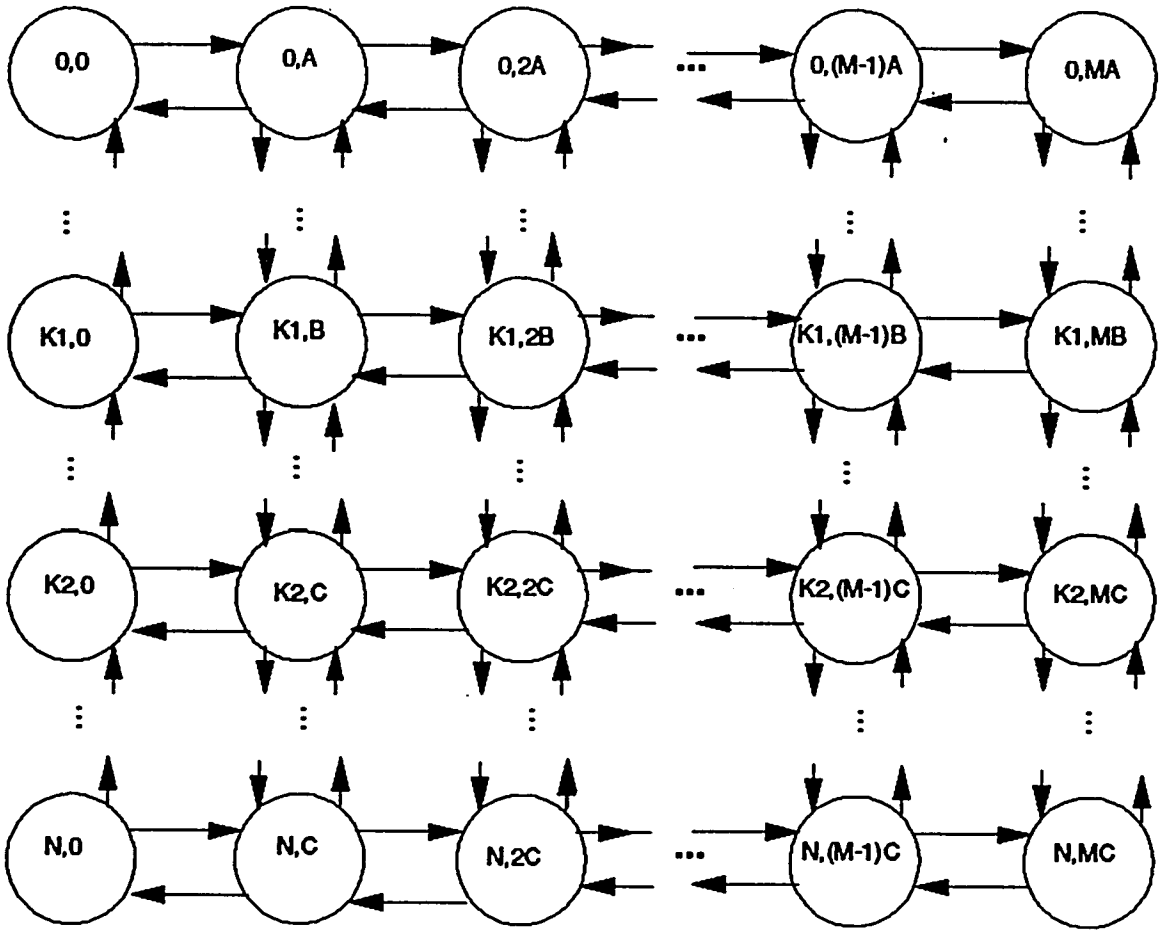


Fig.(IV.2) Continuous Time Markov Chain Model

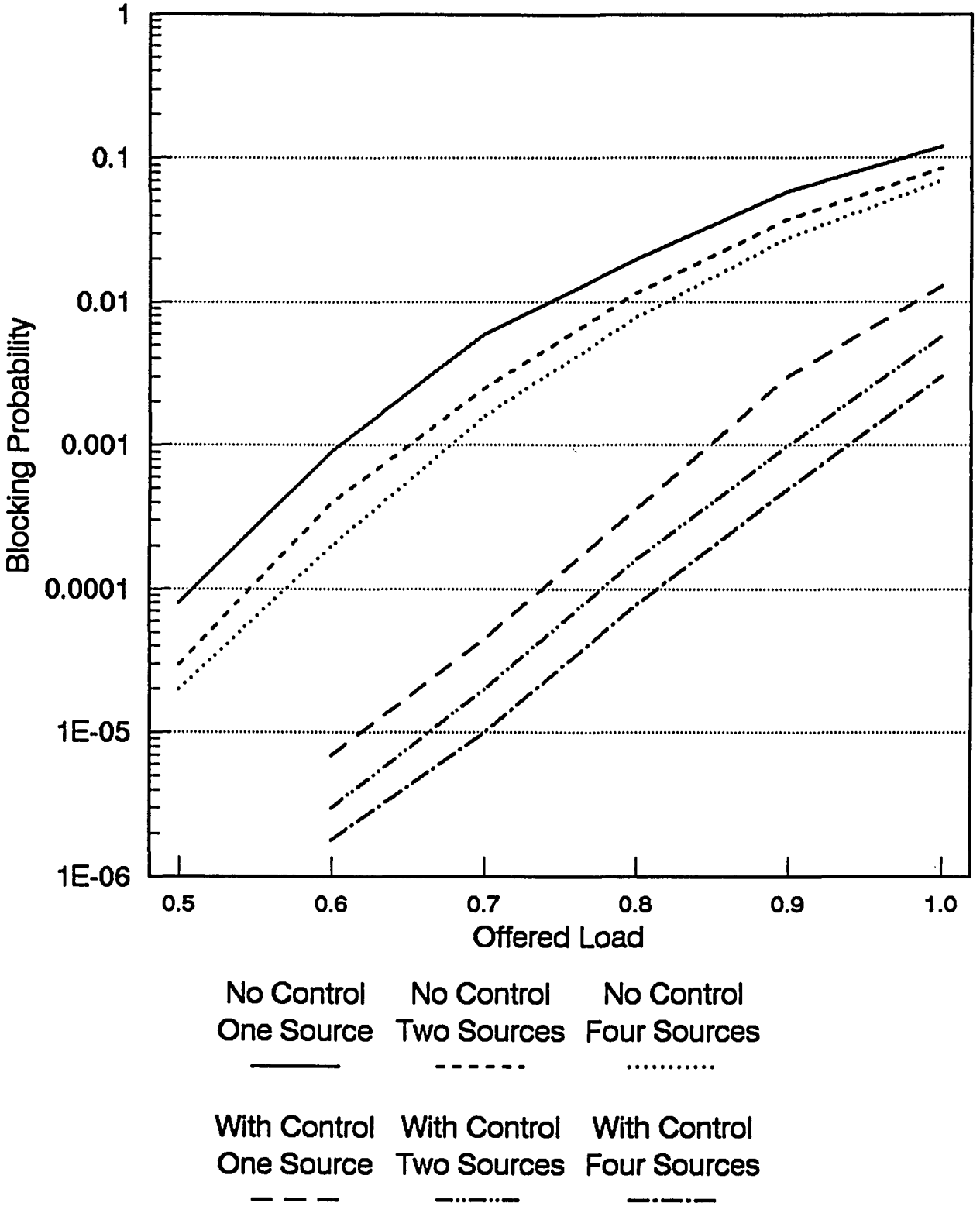


Fig.(IV.3) Blocking Probability Vs. Load
Buffer size=20 cells

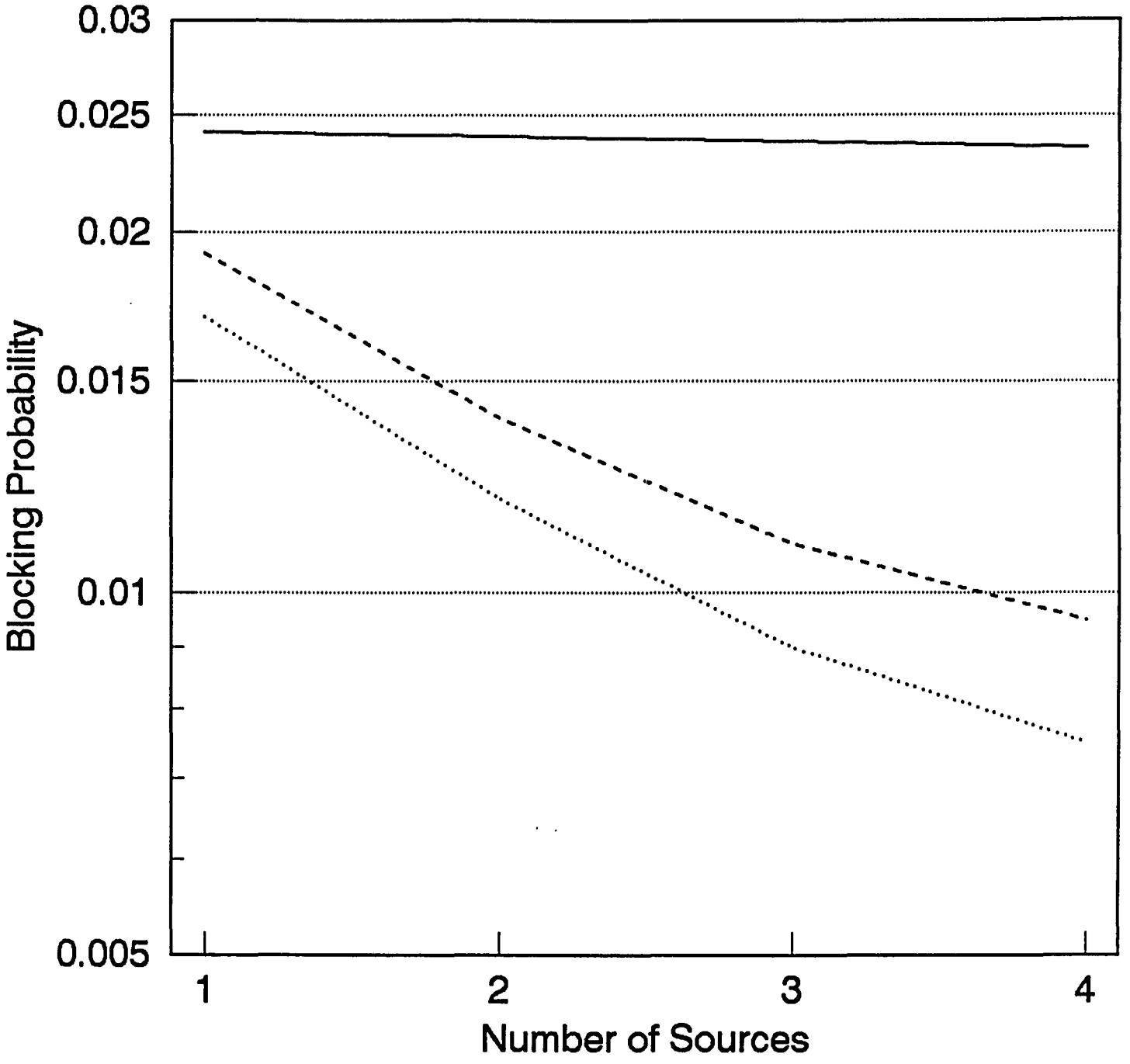


Fig.(IV.4) Statistical Multiplexing Gain
Uitilization=0.8

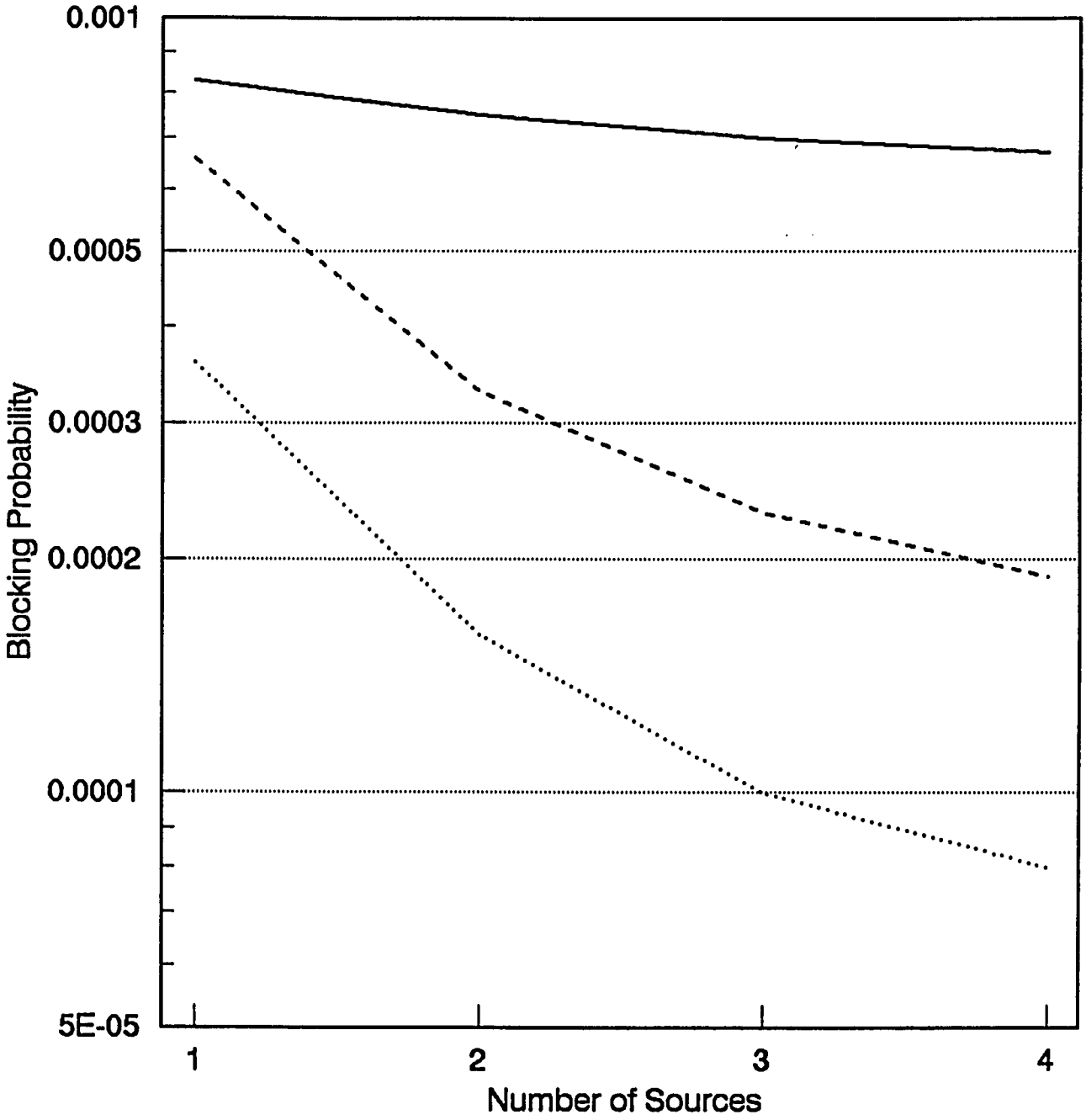


Fig.(IV.5) Statistical Multiplexing Gain
Utilization=0.8

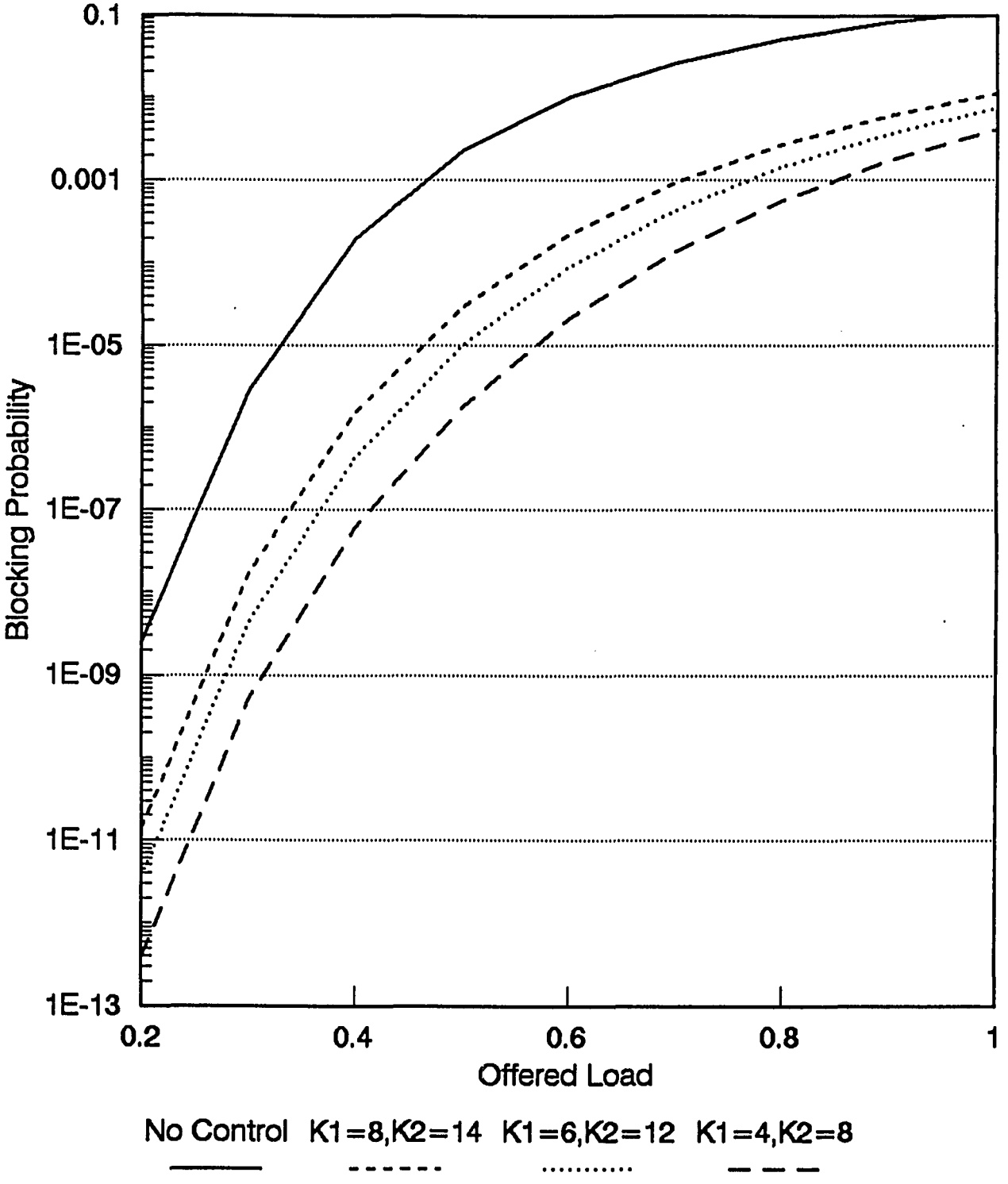


Fig.(IV.6) Blocking Probability Vs. Load
Single Source

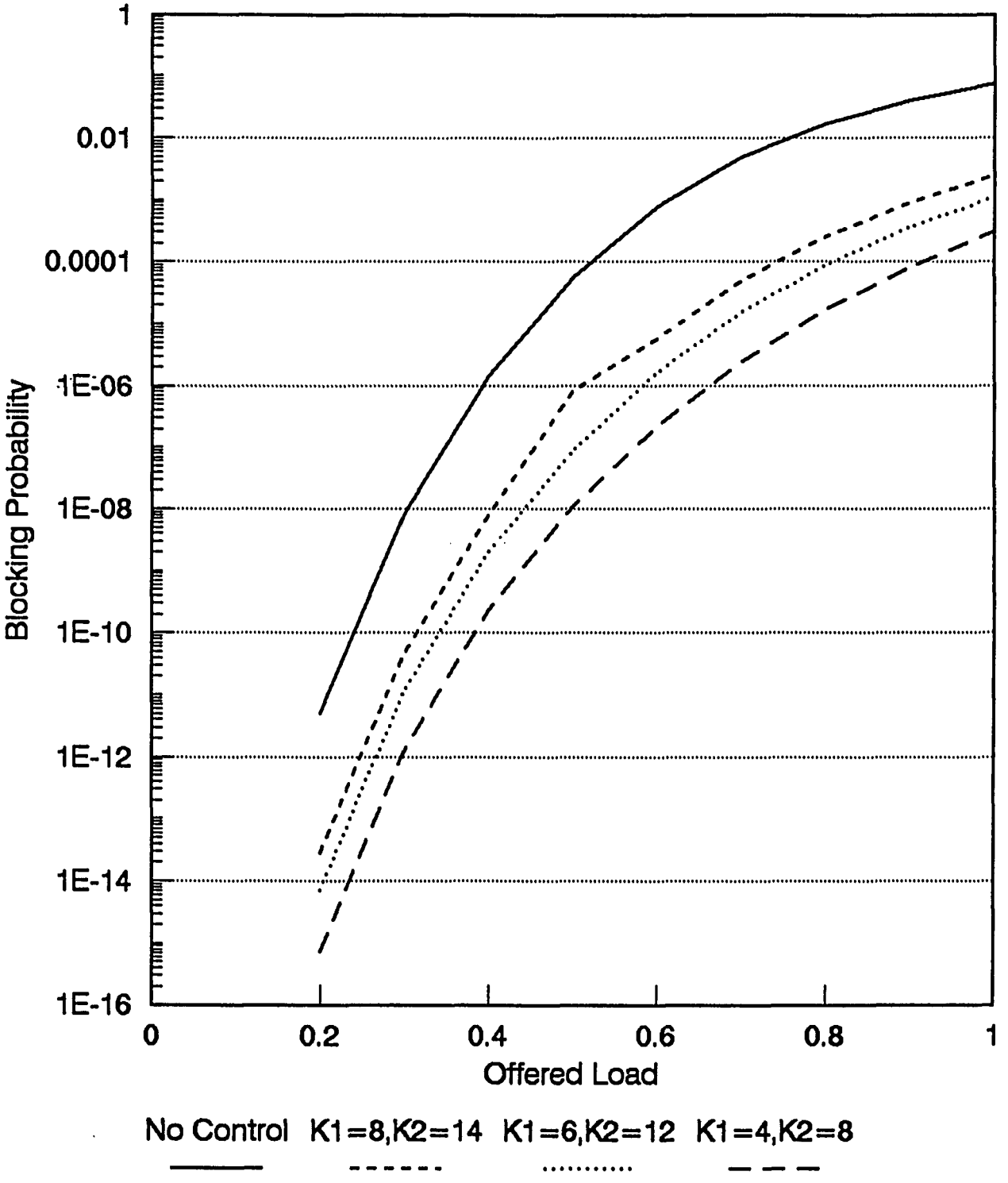


Fig.(IV.7) Blocking Probability Vs. Load Four Sources

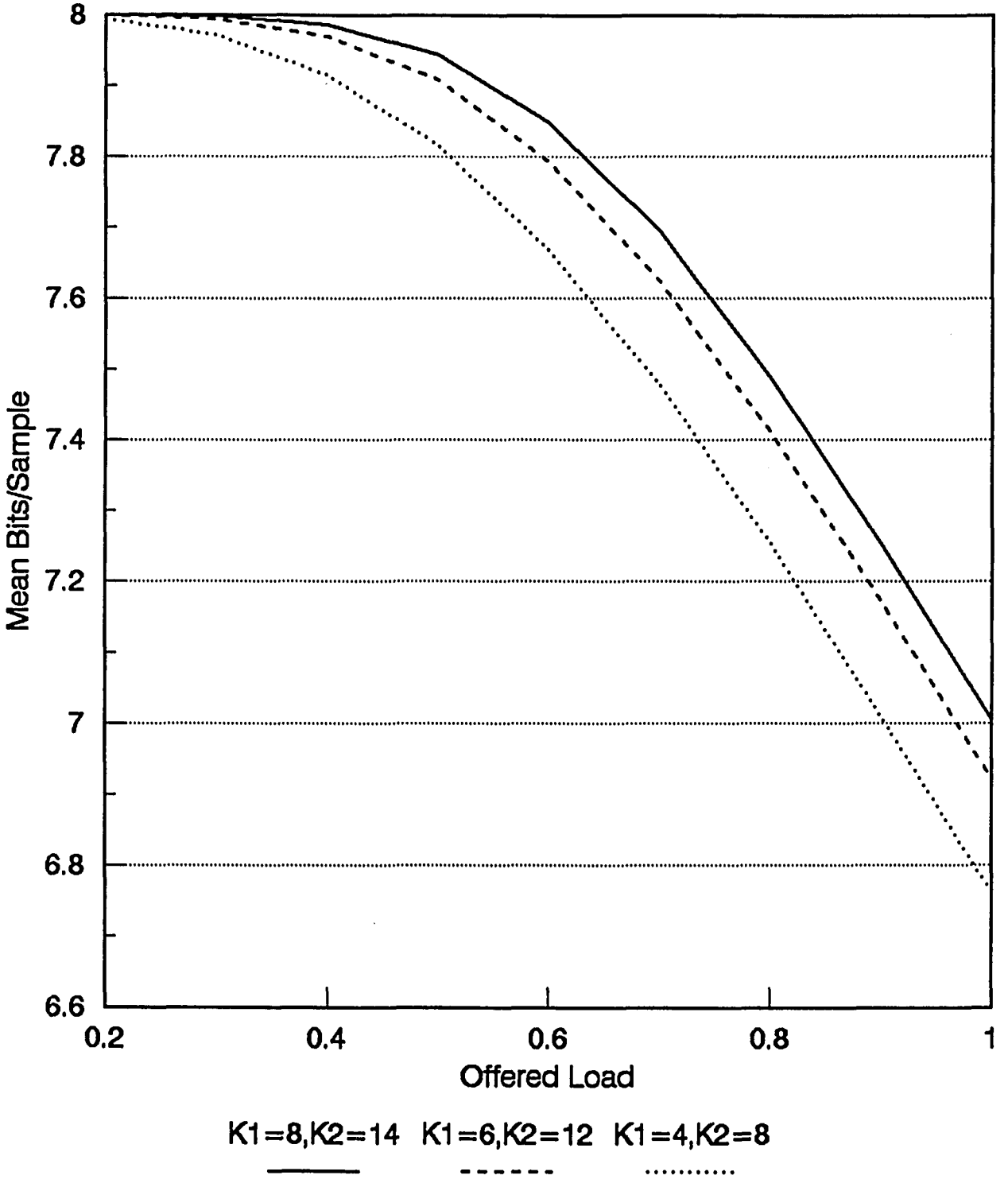


Fig.(IV.8) Mean Bits/Sample Vs. Load
Single Source

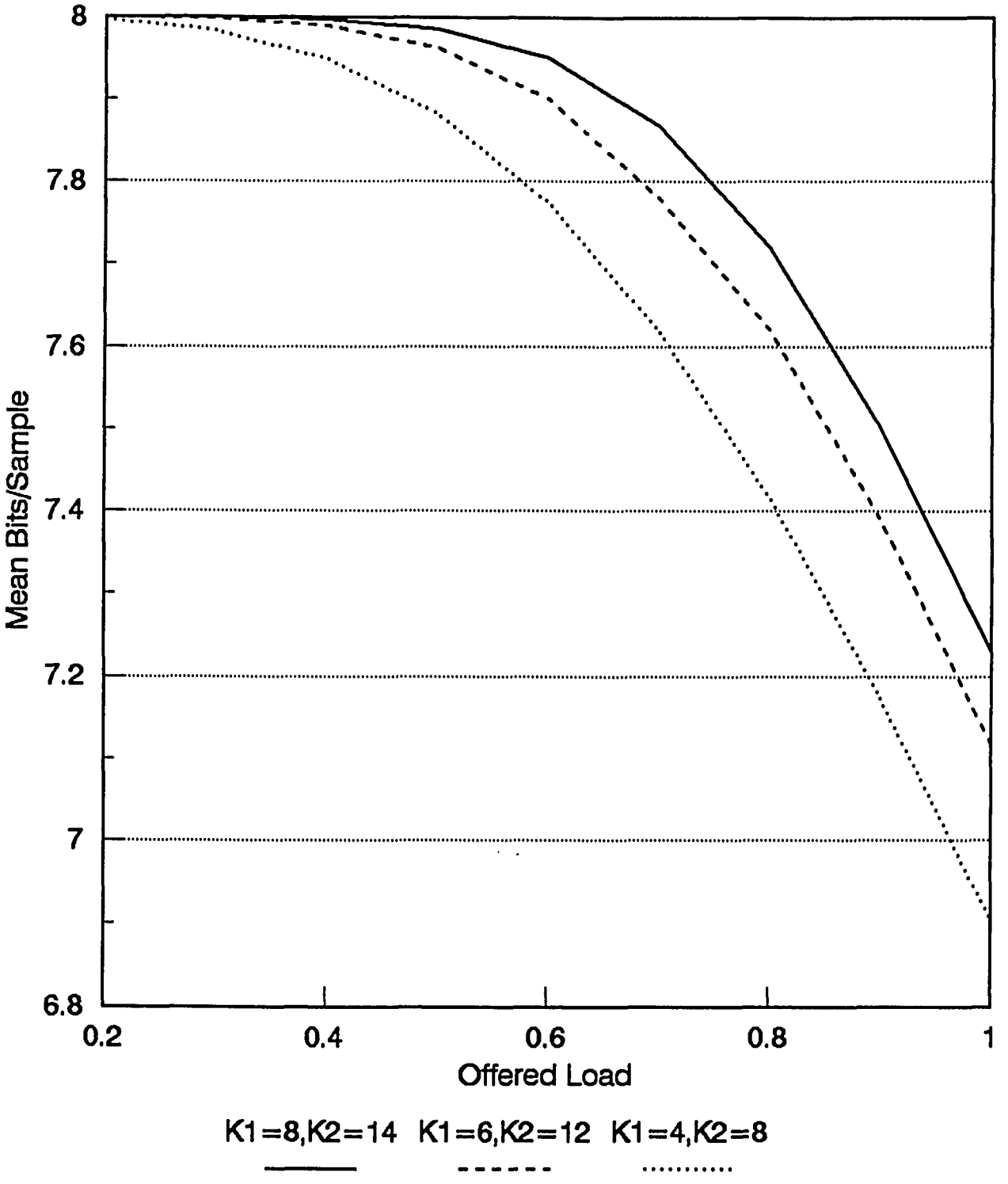


Fig.(IV.9) Mean Bits/Sample Vs. Load
Four Sources

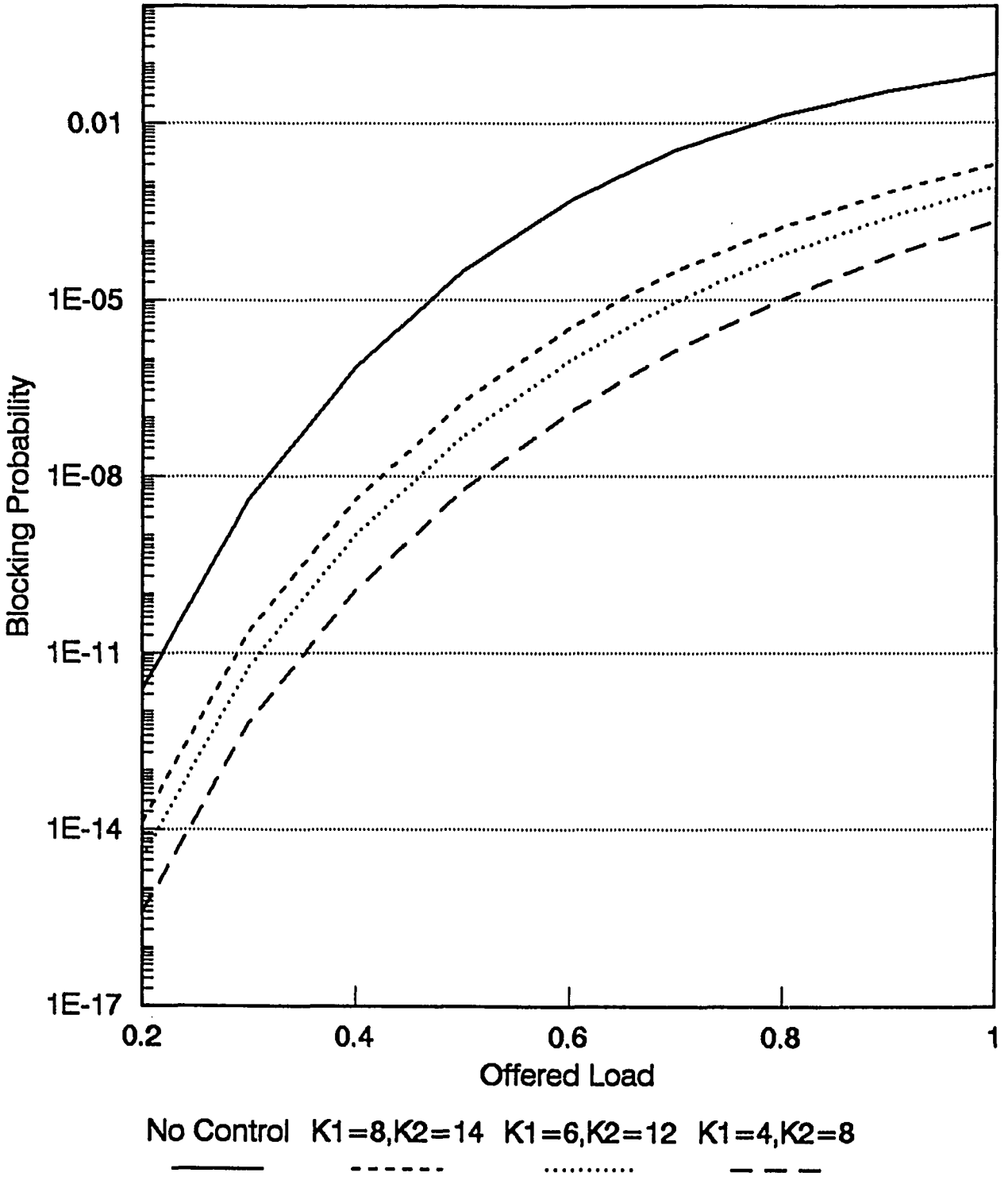


Fig.(IV.10) Blocking Probability Vs. Load
Five Sources

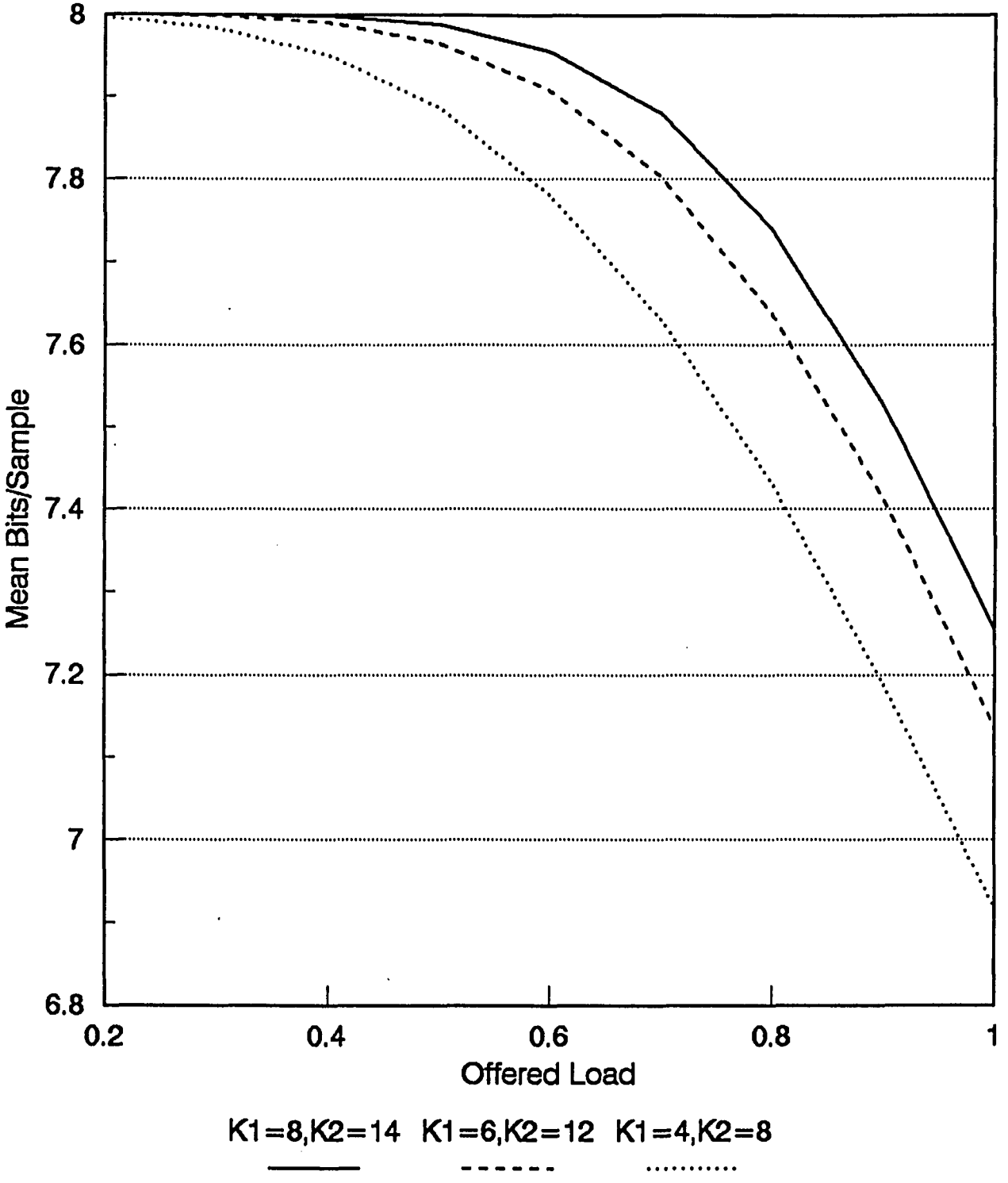


Fig.(IV.11) Mean Bits/Sample Vs. Load
Five Sources

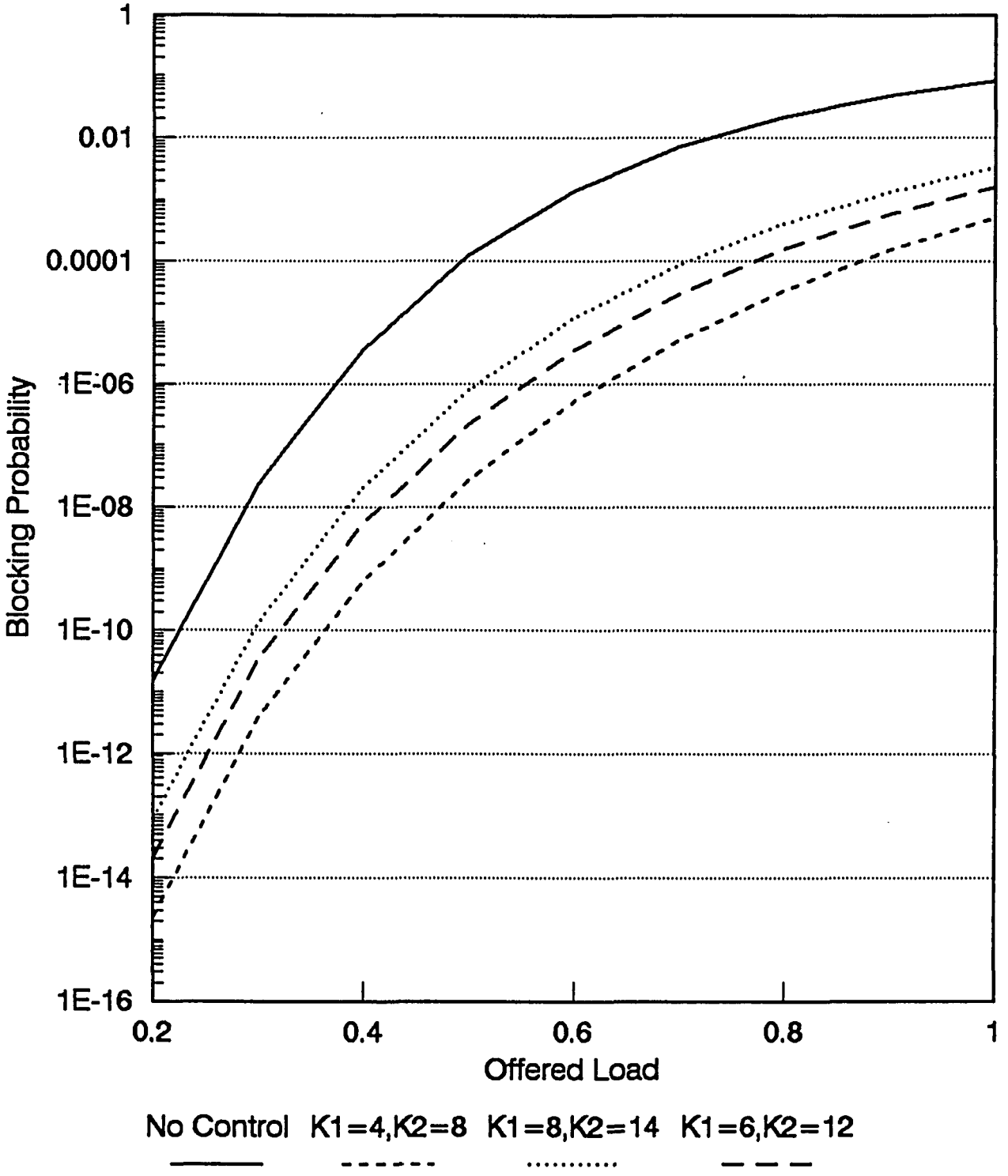


Fig.(IV.12) Blocking Probability Vs. Load
Three Sources

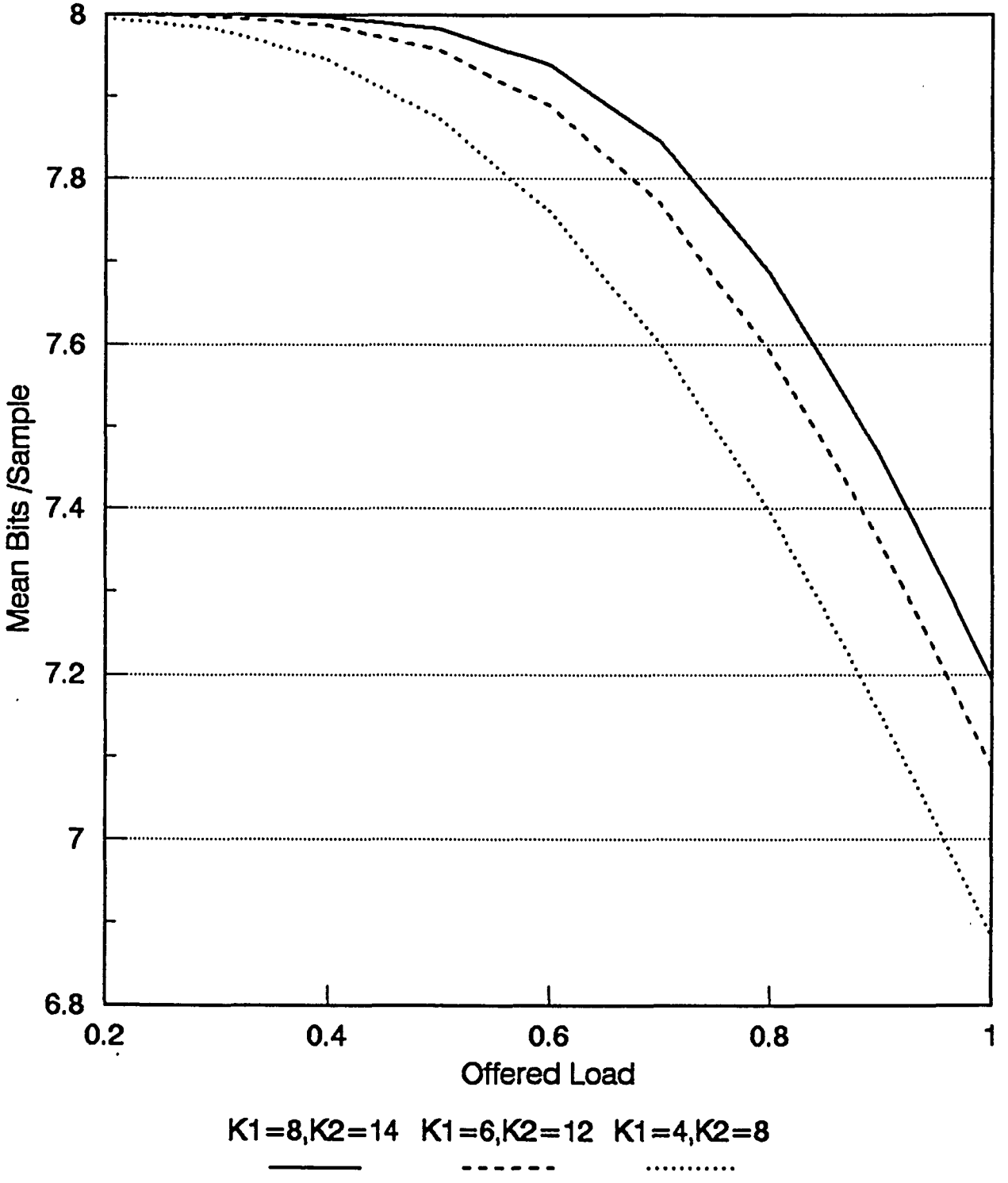


Fig.(13) Mean Bits/Sample Vs. Load
Three Sources

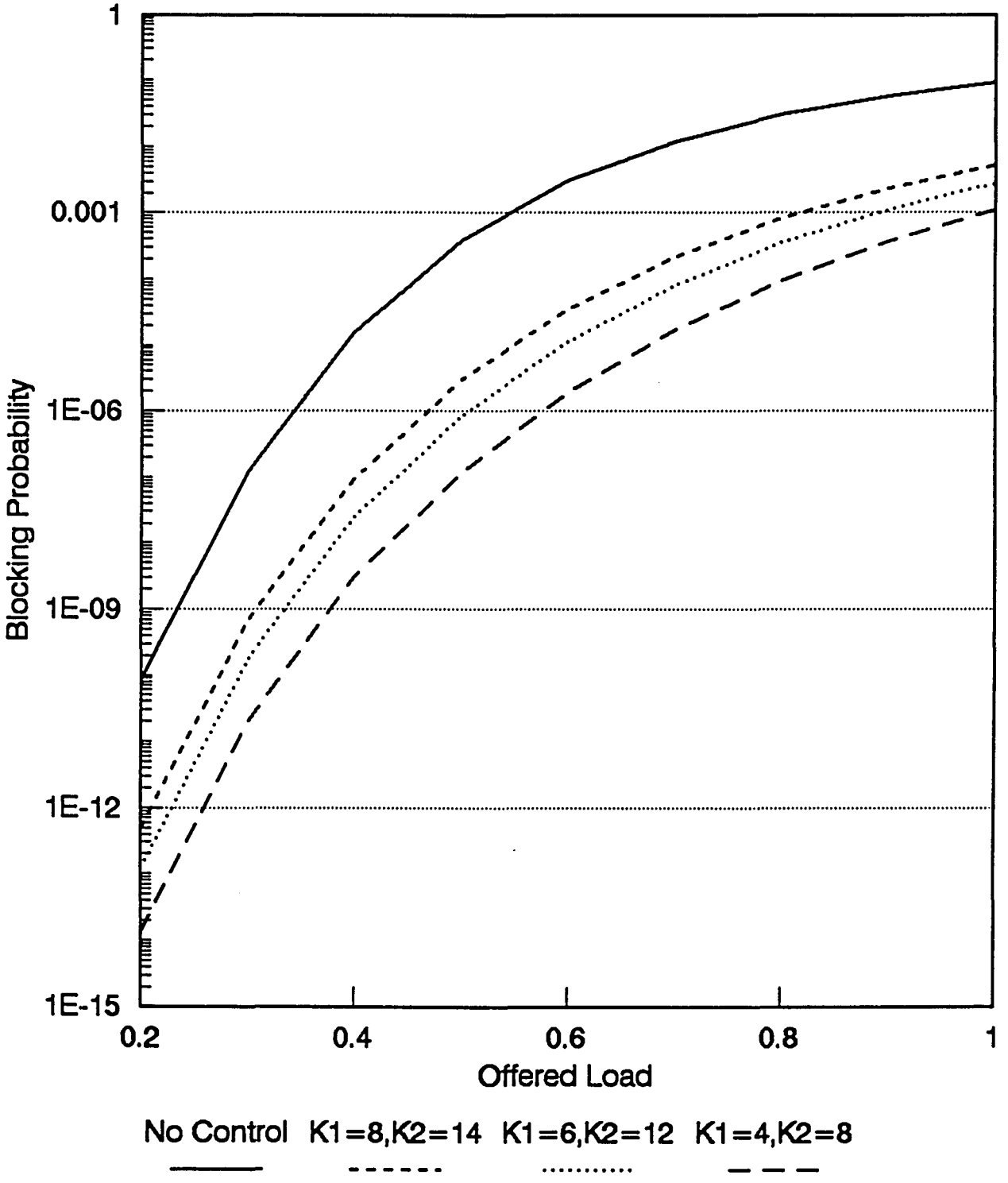


Fig.(IV.14) Blocking Probability Vs. Load
Two Sources

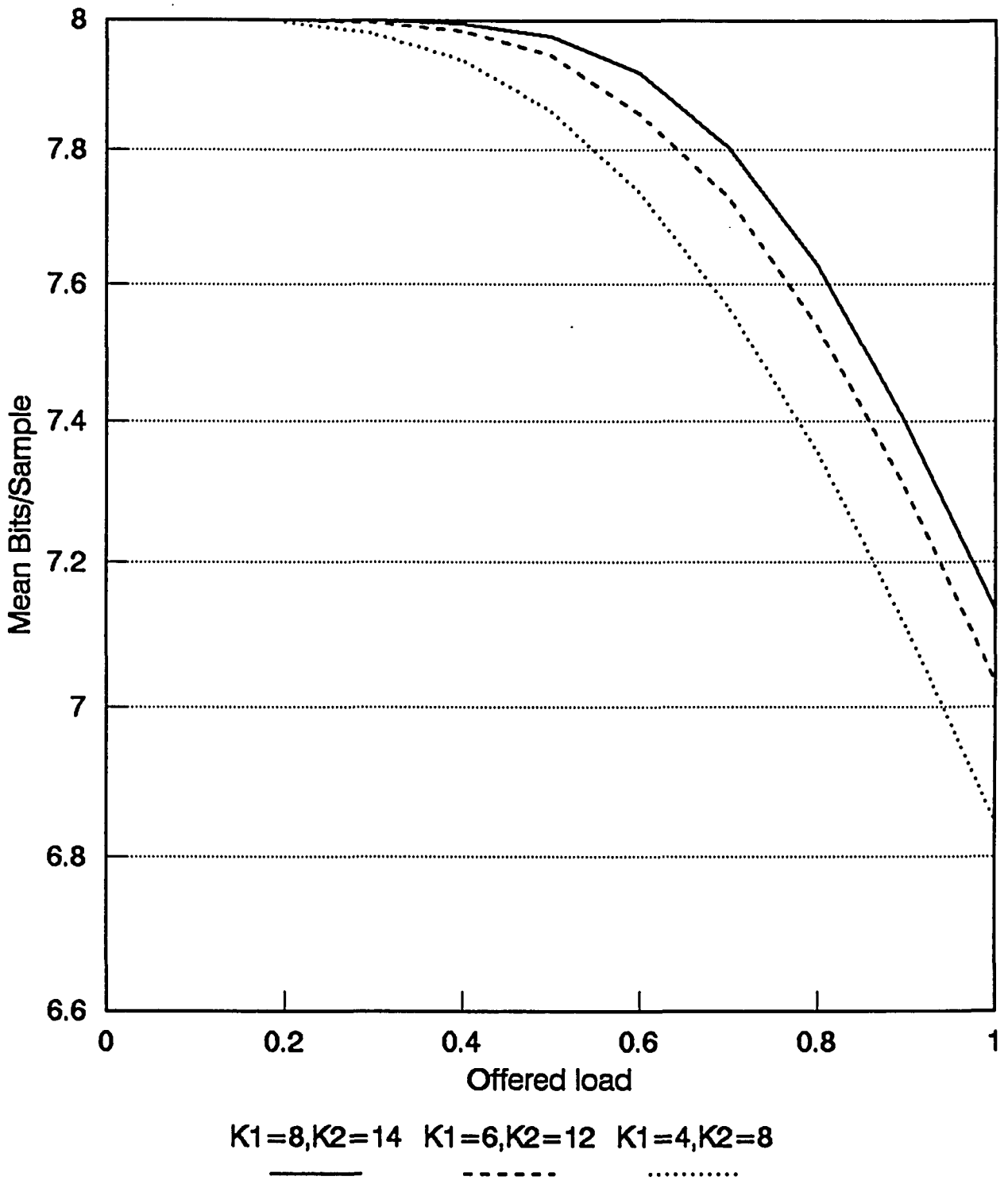


Fig.(IV.15) Mean Bits/Sample Vs. Load
Two Sources

V. Dynamic Bandwidth Allocation of Virtual Paths

V.1. Background

In this chapter, we propose and analyze a dynamic bandwidth allocation and control scheme based upon the virtual path principle. The scheme exploits the statistical multiplexing gain in allocating the bandwidth to each virtual path. The bandwidth allocated is function of, not only, the traffic burstiness but also the required cell loss rate as declared by the Class Of Service (COS) of each supported connection. To dynamically control the allocated path-bandwidth, a Bandwidth Control Period (BCP) rule, is proposed to control the scheduling of different classes of traffic that are supported by separate virtual paths. It is shown that with proper choice of the BCP, a path-bandwidth can be allocated such that to guarantee the traffic required COS in terms of its cell loss and quality delivered. Further more, it is shown that access flow control is required to minimize the cell loss rate and enhance the statistical multiplexing gain per virtual path.

We recall that ATM is based upon two principles

1. Out of band signalling and control functions, supported via separate virtual circuits. This principle makes it possible to design access flow control schemes at the User Network

Interface (UNI), where the network controls the input bit rate from the users and throttle it down to avoid congestion and enhance the bandwidth utilization. In chapters III and IV, an access flow control scheme was proposed and analyzed where the source peak rate is throttled by a feedback control signal from the access node buffer controller. The control signal is out of band, hence it is not affected by the data stream and controls the source rate in time to prevent congestion.
2. Connection oriented principle, where no connections are established unless the required resources to support them are available. This principle calls for efficient dynamic bandwidth allocation and control per virtual path and per each connection within each path. This function is performed at the network and call logical levels respectively.

In [33]-[35] and [37], it was shown that dynamic bandwidth control of each virtual path improves the transmission efficiency and increases the bandwidth utilization at the expense of increased load processing per node, which is required to change the bandwidth allocated per connection. The work done there is based upon the Poisson assumption of call arrivals and that the bandwidth allocated is deterministic and is varied in fixed steps. These assumptions are not valid in the ATM networks. In [49] a bandwidth allocation scheme called the (T1,T2) scheme was presented, in the context of wide band networks, where the bandwidth is allocated to data and voice

traffic queues and the T1 and T2 are time limits set to limit the delay per each traffic. The scheme is dynamic in the sense that once the service in one queue reaches the time limit, the service is switched to the other queue, however it did not consider the issue of the bandwidth required to allocate to each type of traffic such that a certain COS is met which is essential in ATM networks. In [39]-[41], a simulation analysis of a bandwidth allocation scheme called Class Related Rule (CRR) was presented. The study was based upon hypothetical two classes of bursty traffic with different peak, average rates and active periods. Thus the results reported there can not be applied to the real ATM traffic situation.

In this chapter, we consider the bandwidth allocation problem in ATM networks where the input traffic are variable rate video and audio traffic. The bandwidth allocation scheme we present and analyze is based upon the statistical multiplexing gain achieved per each virtual path and is dramatically enhanced by our access control scheme reported in chapters III and IV. We investigate the case of multiplexing several virtual paths carrying different traffic with different correlations and burstiness, such as video and voice. We then present the BCP rule and prove that the traffic with higher correlations (the video traffic in our analysis) dominates the queueing behavior. To the best of the authors knowledge, this problem has never been analyzed before and our results reported here

are a first step in this direction. In section V.2, the dynamic bandwidth allocation and control scheme based upon our BCP rule is presented. In section V.3, we provide the analytical performance evaluation which is based upon the Quasi Birth Death queueing process. In section V.4, numerical analysis and conclusions are given.

V.2. Dynamic Bandwidth Control and the BCP rule

In section V.I, we have elaborated on the concept of virtual paths. The advantages of the virtual paths are numerous and include, direct multiplexing of virtual paths with different bandwidths with a simplified network architecture, statistical bandwidth allocation per call and per path increases efficiency of the link capacity. Two possibilities exist in implementing the virtual paths, one is to support traffic with similar characteristics and COS over the same path. The second alternative is to support traffic with different characteristics and COS on the same virtual path. In this paper we employ the first alternative for several reasons. First, it is easier to enforce the COS function. Secondly, it is also easier to apply traffic enforcement and access control functions at the UNI where such functions are function of the input traffic characteristics, such as the, peak, average rates and coefficient of variation. Finally, multiplexing traffic with

different burstiness does not provide any gain in terms of bandwidth efficiency, moreover the traffic with the higher burstiness and correlations is not "smoothed out" by the lower burstiness traffic, see chapter II for detailed discussion on the subject.

Figure (V.1) shows a block diagram of the access node multiplexer, where three different types of traffic, video, voice and data 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. The controller reads the input traffic characteristics and its required COS and based upon the bandwidth availability, either accepts the call or rejects it (admission control). As each buffer length reaches a certain threshold, access control is activated and the input rate is compressed. A similar action is repeated when another control threshold is achieved. The feedback control signal uses separate out of band virtual circuit to compress the arrival rate. Although we can use the same signal to mask the least significant bits right at the input node, it is not favorable because of the implicit dependency on the physical structure of the cell in order to separate the least the significant bits, thus the control signal is fed back directly to the source coder. 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. As we shall see in section V.4, the bandwidth utilization decreases significantly in the video case when compared to the voice case.

The bandwidth allocation problem can be segregated into two phases. In the first phase, the controller is designed to allocate the bandwidth according to a predetermined rule which is based upon the statistical multiplexer 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 cell loss rate and the number of multiplexed calls per each virtual path. Let $\overline{R}_i$ be the average rate of traffic per call per virtual path, BW_i is the call required bandwidth where there are i classes of traffic supported by i virtual paths per link, and let x_i be the required bandwidth factor then the following holds

$$BW_i = x_i \overline{R}_i \qquad (V.1)$$

Consider there are N calls per virtual path i, then the total virtual path capacity is

$$C_{vpi} = N BW_i \qquad (V.2)$$

bounded by

$$C_l \geq \sum_i C_{vpi} \qquad (V.3)$$

where C_l is the total link capacity

To find the bandwidth allocation factor, we solve each queue independently for the minimum bandwidth required to achieve a certain cell loss rate for each type of traffic and we obtain a set of curves indicating the numerical values for each x_i (the details are given in section V.4). A bandwidth allocation table, which contains a set of statistically assigned bandwidths, is then stored into the controller where it is used to allocate the required bandwidth per each call.

In the second phase, the scheduler schedules transmission of cells, from each queue, such that the average bandwidth allocated to each type of traffic, and hence the average virtual path capacity, equals the value driven from the bandwidth allocation table in phase I. The scheduling scheme follows a simple ATDM technique, where the total link capacity is allocated to serve the video queue for an average time window T_v, followed by a another time window T_a to serve the voice

queue (see fig. V.2). During each time window, a certain number of cells n_v or n_a is transmitted on the link where $n_v = T_v/\mu$, $n_a = T_a/\mu$, and μ is the cell transmission time.

Each of the values of T_a and T_v are averages values drawn from an exponential distributed random variable. The scheduling of cells from separate queues, is controlled via the Bandwidth Control Period (BCP) rule. The BCP, is set to be the scheduler maximum switching period of time required to support both types of traffic (video and voice), such that their respective cell loss rates are delivered. Fig. (V.3), shows a flow chart of the algorithm. As explained above, the controller selects the required bandwidth to support each class of traffic according to the predetermined COS. 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. Accordingly, the average time windows T_a and T_v are set. The controller monitors both the arrival statistics and the buffers' lengths to maintain the required COS and avoid congestion. Whenever the controller detects a change in the arrival statistics or a possible congestion, then the BCP value is changed to provide the required control. The initial BCP value, in this case, will be the sum of the individual time windows allocated to each queue, where

$$BCP = \sum_{i} T_i \qquad i \in (1,2,3...) \qquad (V.4)$$

So that the average bandwidth allocated to the voice queue is

$$BW_a = \sum_{i} C_{vpi} T_a / (T_a + T_v) \qquad (V.5)$$

and the average bandwidth allocated to the video queue is

$$BW_v = \sum_{i} C_{vpi} T_v / (T_a + T_v) \qquad (V.6)$$

The bandwidth allocated, is controlled via the BCP value. It is clear that there is a possible set of values T_a and T_v that can satisfy the bandwidth requirement assigned to each queue. However there is an optimum value of the BCP parameter that bounds the time window allocated to each queue. To start with, the maximum value of the BCP must be less than or at most equals the sum of the sizes of the voice and video queues in cells. This condition is required in order to avoid the possibility of one queue overwhelming the other, and to limit the maximum allowed delay. As the BCP period gets smaller, the switching speed of the scheduler gets higher and hence the mean number of cells per each buffer gets smaller, consequently the cell loss rate decreases. In sections V.3 and V.4, we prove these results analytically. The BCP value depends upon the arrival statistics and the required cell loss rate, it changes dynamically with both the bandwidth utilization and

the traffic burstiness (fig. V.4), hence the bandwidth allocated also changes dynamically. Because the multiplexer accommodates heterogeneous traffic mix, there are different possible BCP values to allocate bandwidth for different traffic mixtures. For example, the BCP required to support voice calls at a utilization of 0.8 and cell loss rate of 10^{-4} is much greater than that required to support video calls at the same utilization and cell loss rate of 10^{-9}.

V.3. Modeling and Performance Analysis

Recently several stochastic models has been introduced to model the superposition of a number of voice or video independent sources, that comprises the arrival process to the voice or the video queue respectively [77]-[81]. We recall, from chapter III, that the voice source is represented by a two state continuous time markov chain, alternating between active and idle periods where the duration of the active period is $1/\alpha_a$ secs. and the duration of the idle period is $1/\beta_a$ secs. The superpositon arrival stream can be represented by a phase type continuous time markov chain, where the state of chain is the number of active voice sources. The transitional rate matrix of such birth-death process is given by

$$Q_\alpha = \begin{pmatrix} -N_\alpha\alpha_\alpha & N_\alpha\alpha_\alpha & & & & & \\ \beta_\alpha & -(N_\alpha-1)\alpha_\alpha-\beta_\alpha & (N_\alpha-1)\alpha_\alpha & & & & \\ & \ddots & \ddots & \ddots & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & \ddots & \ddots & \ddots & \\ & & & & \ddots & \ddots & \ddots \\ & & & & & N_\alpha\beta_\alpha & -N_\alpha\beta_\alpha \end{pmatrix} \tag{V.7}$$

which yields extremely high number of states and complicates the solution. However, the same process is equivalent to a two state Modulated Markov Poisson Process (MMPP). The process switches between two states, High and Low, with arrival rates λ_α^H and λ_α^L. The mean sojourn time in each state is exponentially distributed with means $1/r_H$ and $1/r_L$respectively. The transitional rate matrix is given by

$$Q_\alpha = \begin{pmatrix} -r_\alpha^H & r_\alpha^H \\ r_\alpha^L & -r_\alpha^L \end{pmatrix} \tag{V.8}$$

In the case of the superposition of video sources, the model depends upon the contents of the scene, and the type of video coding algorithm implemented. In our work, we used the model introduced in [44] which assumes single motion activity scene. The aggregate video arrival rate is also modeled by

a continuous time birth-death process with exactly the same structure as Q_a above. In this case, $N_v = 10M$, where M is the number of active video sources.

As explained in section V.2, we first solve for the steady state probabilities for each queue independently. The stochastic queueing process is a quasi birth-death process where the service time is replaced by an exponentially distributed with mean $1/\mu$. It was shown in [75],[76] that this replacement had no effect on the queueing process, since the correlations effect introduced by the video arrival process dominates the queueing behavior over the buffer length. In the voice case, however, it was shown, in chapter III, that this approximation does overestimate the cell loss rate for small buffer sizes. This is due to the fact that the effect of correlations between interarrival times is limited because of the limited buffer size and the process approaches the Poisson approximation for large number of input sources. To account for the access control scheme, let K1 and K2 be the buffer control thresholds. As the queue length reaches K1, the arrival rate drops to 0.75 of its uncontrolled level and drops further to 0.5 of its original level at K2. The infinitesimal generator for the queueing process for each of the voice and video queues is

$$Q=\begin{pmatrix}
A & \Lambda_a & & & & & & & & & \\
S & B & \Lambda_a & & & & & & & & \\
 & \cdot & \cdot & \cdot & & & & & & & \\
 & & \cdot & \cdot & \cdot & & & & & & \\
 & & & S & C & \Lambda_a & & & & & \\
 & & & & S & C & \Lambda_a(1) & & & & \\
 & & & & & \cdot & \cdot & \cdot & & & \\
 & & & & & & \cdot & \cdot & \cdot & & \\
 & & & & & & & S & D & \Lambda_a(2) & \\
 & & & & & & & & S & D & \Lambda_a(2) \\
 & & & & & & & & & \cdot & \cdot \\
 & & & & & & & & & \cdot & \cdot \\
 & & & & & & & & & S & E
\end{pmatrix} \qquad (V.9)$$

where

$A, B, C, D, E, S, \Lambda_a, \Lambda_a(1), \Lambda_a(2)$ are all matrices of the dimension 2, in the voice case, and N_v in the case of the video traffic. In the case of the voice queueing process we have

$$A=\begin{pmatrix} -r_a^H-\lambda_a^H & r_a^H \\ r_a^L & -r_a^L-\lambda_a^L \end{pmatrix} \qquad (V.10)$$

$$\Lambda_a=\begin{pmatrix} \lambda_a^H & \\ & \lambda_a^L \end{pmatrix} \qquad (V.11)$$

$$B=\begin{pmatrix} -r_a^H-\lambda_a^H-\mu & r_a^H \\ r_a^L & -r_a^L-\lambda_a^L-\mu \end{pmatrix} \qquad (V.12)$$

$$E=\begin{pmatrix} -r_a^H-\mu & r_a^H \\ r_a^L & -r_a^L-\mu \end{pmatrix} \qquad (V.13)$$

$$S=\begin{pmatrix} \mu & \\ & \mu \end{pmatrix} \qquad (V.14)$$

The matrices C, D are of the same structure as B with the appropriate scaled λ_a s. Similarly $\Lambda_a(1), \Lambda_a(2)$, are the same as Λ_a but with the appropriate scaled λ_as at each corresponding threshold levels K1 and K2.

In the video case, we have the following

$$A=\begin{pmatrix} -N_v\alpha_v & N_v\alpha_v & & & & & \\ \beta_v & -(N_v-1)\alpha_v-\beta_v-\lambda_v & (N_v-1)\alpha_v & & & & \\ & \ddots & \ddots & \ddots & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & \ddots & \ddots & \ddots & \\ & & & & \ddots & \ddots & \ddots \\ & & & & & N_v\beta_v & -N_v\beta_v-\lambda_v \end{pmatrix} \qquad (V.15)$$

$$B=\begin{pmatrix} -N_v\alpha_v-\mu & N_v\alpha_v & & & & & \\ \beta_v & -(N_v-1)\alpha_v-\beta_v-\lambda_v-\mu & (N_v-1)\alpha_v & & & & \\ & \ddots & \ddots & \ddots & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & \ddots & \ddots & \ddots & \\ & & & & \ddots & \ddots & \ddots \\ & & & & & N_v\beta_v & -N_v\beta_v-\lambda_v-\mu \end{pmatrix} \qquad (V.16)$$

$$E=\begin{pmatrix} -N_v\alpha_v-\mu & N_v\alpha_v & & & & & \\ \beta_v & -(N_v-1)\alpha_v-\beta_v-\mu & (N_v-1)\alpha_v & & & & \\ & \ddots & \ddots & \ddots & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & \ddots & \ddots & \ddots & \\ & & & & \ddots & \ddots & N_v\beta_v \\ & & & & & N_v\beta_v & -N_v\beta_v-\mu \end{pmatrix} \qquad (V.17)$$

$$S=\begin{pmatrix} \mu & & & \\ & \mu & & \\ & & \ddots & \\ & & & \mu \end{pmatrix} \qquad (V.18)$$

and Λ_v is a diagonal matrix, with dimension N_v and the elements $\lambda_v = iA$. Similarly, the matrices C, D are the same as B but with corresponding λ_v scaled to 0.75 and 0.5 of its value at K1 and K2 respectively.

The generator matrices Q_a, Q_v are solved independently for the voice and video queue lengths distributions. The minimum required capacity is calculated, where in the video queue case, the cell loss rate is

$$PL_v = \sum_i P_{(K,i)} i\lambda_v \Big/ \sum_{(j,i)} P_{(j,i)} i\lambda_v \qquad 0 \le j \le K, 0 \le i \le 10M \qquad (V.19)$$

while in the voice queue it becomes,

$$PL_a = \sum_{i} P_{(K,i)}\lambda_a^i(2) / \sum_{(j,i,l)} P_{(j,i)}\lambda_a^i(l) \qquad 0 \le j \le K, 0 \le l \le 2, \quad (V.20)$$

To solve for the steady state probabilities, we used matrix geometric techniques reported in [81].

In the second phase of the problem, we have to solve for the maximum BCP period which will guarantee the bandwidth allocation to each type of traffic. The model solves the queueing problem in fig.(V.2), where the mean sojourn time that the server spends at each queue is approximated by an exponentially distributed random variable with mean equals T_a for the voice queue and T_v for the video queue. This approximation fits well with our proposed BCP rule explained in the previous section. It follows, also, from the fact that the steady state probabilities of the number of cells per queue is well known to follow a geometric distribution, hence on the average, the mean sojourn time is approximated by an exponential random variable. It remains to add, that the stochastic process, represented by the server departure times from one queue to the other, is a phase type renewal process [82], where its generator matrix is Q_a for the voice queue and Q_v for the video queue. In fact, the problem is unsolvable without this approximation. It follows that the server stochastic process

can be represented by a two state continuous time markov chain, where in one state the server serves the voice queue for an average time of T_a and in the other state, it serves the video queue for an average time of T_v. Let r^s_{a-v} and r^s_{v-a} be the server transition rates from the voice to the video queue, and from the video to the voice queue respectively, then

$$r^s_{a-v} = 1/T_a \qquad (V.21)$$

$$r^s_{v-a} = 1/T_v \qquad (V.22)$$

The problem, now, is to solve two dependent queues of the type PH/PH/1/K. In the case of the video queue, the transitional rate matrix is

$$Q = \begin{pmatrix}
A & \Lambda_v & & & & & & & & & \\
S & B & \Lambda_v & & & & & & & & \\
 & \ddots & \ddots & \ddots & & & & & & & \\
 & & & \hat{S} & C & \Lambda_v & & & & & \\
 & & & & \hat{S} & C & \Lambda_v(1) & & & & \\
 & & & & & \ddots & \ddots & \ddots & & & \\
 & & & & & & & \hat{S} & D & \Lambda_v(2) & \\
 & & & & & & & & \hat{S} & D & \Lambda_v(2) \\
 & & & & & & & & & \ddots & \ddots \\
 & & & & & & & & & & \hat{S} & E
\end{pmatrix} \qquad (V.23)$$

$$A=\begin{pmatrix} -N_v\alpha_v & N_v\alpha_v & & & & \\ \beta_v & -(N_v-1)\alpha_v-\beta_v-\lambda_v & (N_v-1)\alpha_v & & & \\ & \ddots & \ddots & \ddots & & \\ & & \ddots & \ddots & \ddots & \\ & & & \ddots & \ddots & \ddots \\ & & & & N_v\beta_v & -N_v\beta_v-\lambda_v \end{pmatrix} \qquad (V.24)$$

$$S=transpose\begin{pmatrix} 0 & \mu & & & & \\ & 0 & \mu & & & \\ & & \ddots & \ddots & & \\ & & & \ddots & \ddots & \\ & & & & \ddots & \ddots \\ & & & & & 0 \quad \mu \end{pmatrix} \qquad (V.25)$$

$$B=\begin{pmatrix} -N_s\alpha_s-p\lambda_s-r'_{s-s} & r'_{s-s} & N_s\alpha_s & & & \\ r'_{s-s} & -N_s\alpha_s-q\lambda_s-r'_{s-s}-\mu & 0 & N_s\alpha_s & & \\ \beta_s & 0 & -(N_s-1)\alpha_s-p\lambda_s-r'_{s-s}-\beta_s & r'_{s-s} & (N_s-1)\alpha_s & \\ & \ddots & \ddots & \ddots & \ddots & \ddots \\ & & \ddots & \ddots & \ddots & \ddots \\ & & & \ddots & \ddots & \ddots \\ & & & & N_s\beta_s & r'_{s-s} \quad -N_s\beta_s-q\lambda_s-r'_{s-s}-\mu \end{pmatrix} \qquad (V.26)$$

$$\Lambda_v=\begin{pmatrix} p\lambda_v & q\lambda_v & & & \\ & \ddots & \ddots & & \\ & & \ddots & \ddots & \\ & & & \ddots & \ddots \end{pmatrix} \qquad (V.27)$$

$$\hat{S} = \begin{pmatrix} 0 & & & & \\ & \mu & & & \\ & & 0 & & \\ & & & \mu & \\ & & & & \ddots \end{pmatrix} \qquad (V.28)$$

where

$$p = T_a/(T_a + T_v) \qquad (V.29)$$

$$q = 1 - p \qquad (V.30)$$

The matrix A is of dimension N_v, while the matrices $S \wedge \Lambda_v$ are of dimension $N_v x 2N_v$ and all other matrices are of dimension $2N_v x 2N_v$. Similar structure is solved for the voice queue where the arrival process is an MMPP one, as explained above. The solution of the above matrices involve numerous numerical calculations using iterative matrix geometric techniques as indicated in the independent case.

V.4. Numerical Results and Conclusions

The video source model used in this paper, generates 30 frames per second, each frame contains 512 x 512 pixels. The quantized arrival bit rate (A) is (0.152 bits/pixel), $\beta_v = 2.59 transitions/\sec, \alpha_v = 1.3 transitions/\sec$. These values are

adopted from the work done in [53], however we normalized these values so that all the units used are in cells/msec. The 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 delay to 50 μsecs, although this requirement can be relaxed in the actual ATM network where the maximum tolerable delay can be 0.1 msec per node. The cell loss rate was set to 10^{-9} for the video traffic and 10^{-4} for the voice queue, which are reasonable figures for high speed networks. The values of the MMPP for the voice traffic were evaluated from the work done in [55].

Figs. (V.5) and (V.6) show the statistical bandwidth assignment, required to support the voice and video traffic respectively. We can see that the access flow control has significantly improved the statistical multiplexing gain in both cases, and has significantly reduced the bandwidth requirements at a particular COS. Because of the higher correlations and burstiness, in the video traffic, the access control thresholds (K1, K2) have to be smaller than those for the voice traffic. Also, the statistical multiplexing gain is not as effective as with the voice traffic. One solution would be to increase the video traffic buffer size, however this solution might not solve the problem, since increasing the buffer size would directly cause the delay and the delay

variability to be significantly high. Consequently, out of timing cells, arriving at the receiver, would be considered lost leading to poor image quality. A more effective solution, is to increase the ratio of the link capacity to the input traffic peak bit rate. Another important observation, is that as (K1,K2) values change, in the video queue, there is very little gain achieved by that, contrary to the voice queue where the sensitivity of the control threshold levels is much more perceptible.

Figs. (V.7) and (V.8) show the bandwidth allocation factor (X) for the voice and video traffics. It is also clear that the access flow control has a perceived gain in terms of the bandwidth allocation. This gain can be also viewed as an increase in the number of accepted calls at a certain bandwidth, which implies higher utilization and a decrease in the cost per connection for the users. Notice that this ratio determines the bandwidth allocated to each type of traffic. It is the basis upon which, the controller's look-up table is constructed, and thus by far the most important parameter for network traffic management. The bandwidth utilization per voice and video queues is further elaborated in figs (V.9) and (V.10), where the increase in the utilization is more apparent in the voice traffic than the video traffic, mainly because of the increased correlations effect in the video queue case. For the voice traffic, the maximum utilization is 61% without control and

with 12000 voice lines being multiplexed. The utilization has increased from 61% to 92% with access flow control at the same load, which reflects how crucial is the access flow control in enhancing the statistical multiplexing gain. The same effect holds for the video queue, where the utilization has increased from 50% at a load of 7 video sources without control up to close to 90% at the same load. Notice that the control threshold levels do not show any significant sensitivity, contrary to the voice traffic case.

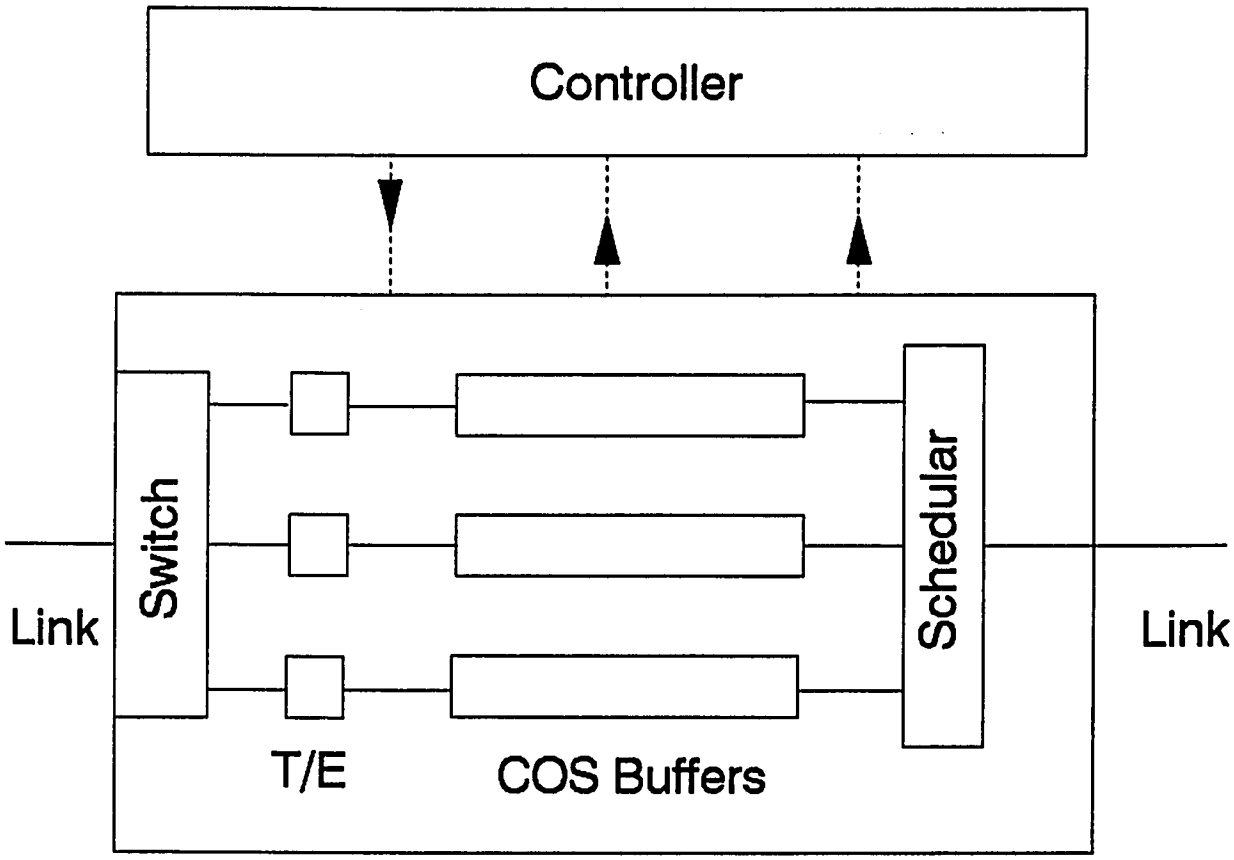
Figs. (V.11) and (V.12) reflect the price paid in terms of the delivered quality. A very graceful degradation in the voice quality over the input traffic load, while the degradation effect is quite clear in the video traffic. It is difficult to assess exactly the image quality without subjective tests, but for this work we have used the mean bits/sample as a factor as it relates to the Signal to Noise ratio. Because of the degradation effect, it is logical to use the threshold values (10,16) in the video traffic case, as there is no high gain achieved by decreasing the threshold levels, moreover the image quality degrades as the load is increased.

Figs. (V.13) to (V.18), show how the BCP changes with bandwidth utilization of both the voice and video traffic. Each BCP value on the curves reflects its maximum possible value in order to support the given video and voice calls,

under the required COS. In fig.(V.13), we loaded the voice queue with only one source, and changed the voice load. We applied access control for the voice queue only, and as expected the BCP value increased significantly from 16 cells transmission time to 25 cells transmission time at a voice load of 4,000 connections. At a voice load of 10,000 connections (no voice access control) and single video source, there is no BCP value to support such load, as the total link capacity is saturated. A BCP value of 6 cell transmission time is required to support at most 9000 voice lines and single video source with no control, this BCP value implies a very fast switching speed at the scheduler. In other words the maximum time that the server can stay at the voice queue would be 4 cells transmission time (12 μsecs. and only 2 cells transmission time (6 μ)secs. at the video queue. Using the access control, only for voice queue, the number of supported voice lines goes up to 14,000 connections at a BCP value of 7 cells transmission time. The statistical multiplexing gain has been enhanced, which can be observed by comparing the slope of the curves for different control thresholds over the no-control case.

Fig.(V.14) shows the same effect, however the video queue has been loaded with three video sources in this case. Because the bandwidth allocation has increased, in order to support three video sources, the BCP value has dropped in this case from 16 cells transmission time (for the single video source

case) to 13 cells transmission time at the same 4,000 voice connections with no control. However, the statistical multiplexing gain is more significant in this case due to the multiplexing of three video sources, whereas in fig.(V.13) only one video source was supported. In figs.(V.15),(V.16) we present the performance curves when video access control is applied. In these cases the gain achieved has increased in terms of the higher values that BCP can acquire to support a certain bandwidth utilization. The same gain is also clear in terms of an increased number of voice and video connections that can be supported at a certain fixed BCP value. Fig.(V.17) elaborates on the statistical multiplexing gain over the voice load spectrum, where the video queue was loaded with only a single source. The curve proves that as the utilization of the video queue increases, the multiplexing gain also increases. Consequently, the multiplexer can support the same number of voice connections at a higher BCP value, and thus reducing the scheduler switching speed. Fig.(V.18) compares the results of supporting both single and three video sources with, and without access control. The difference in the slopes when video access flow control is applied, reflects the enhancement in the statistical multiplexing gain, as explained before.



T/E: Traffic Enforcement

Fig.(V.1) ATM Access Node Multiplexer with COS Multipath Bandwidth Control

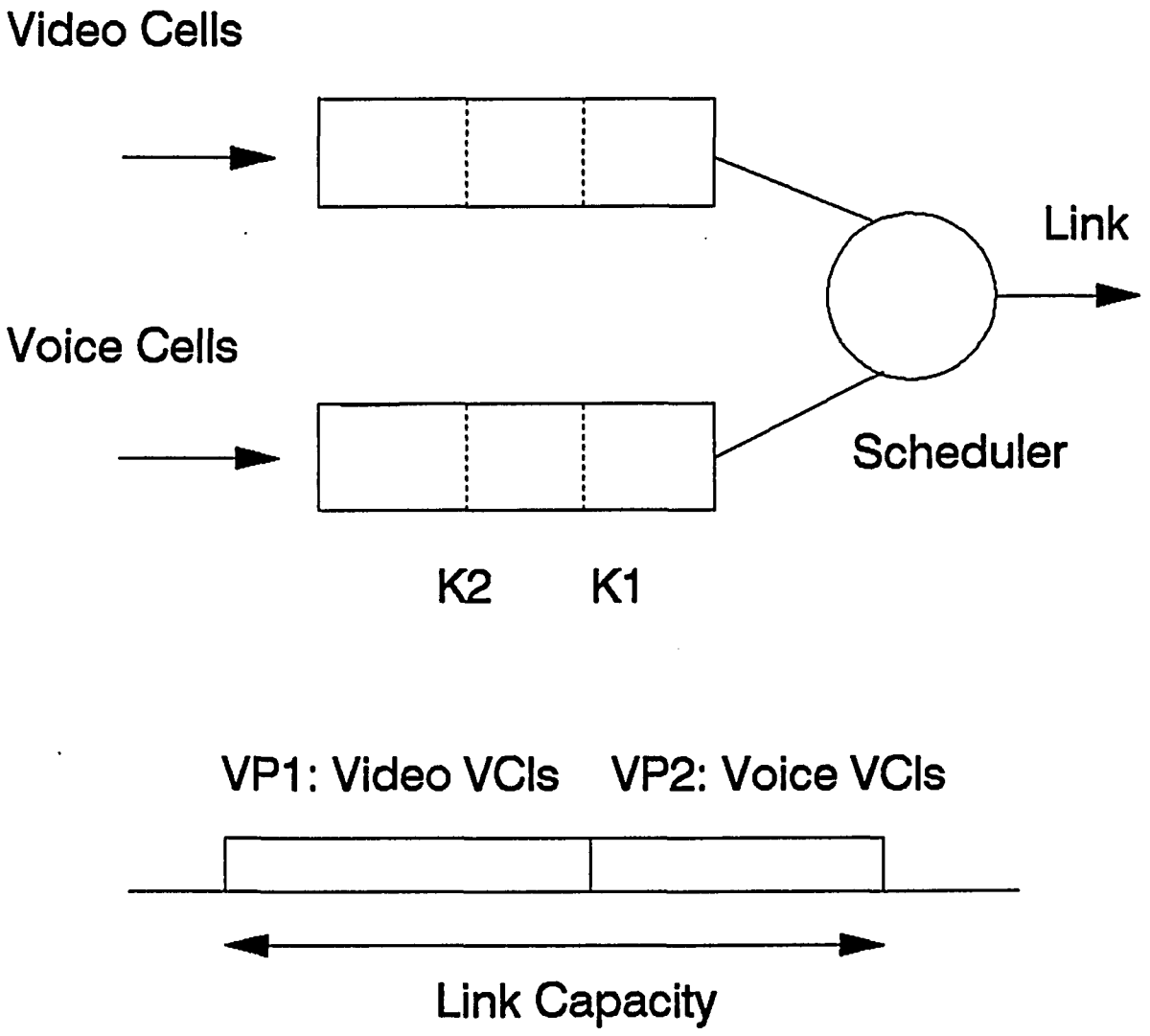


Fig.(V.2) Scheduling of Heterogeneous Traffic

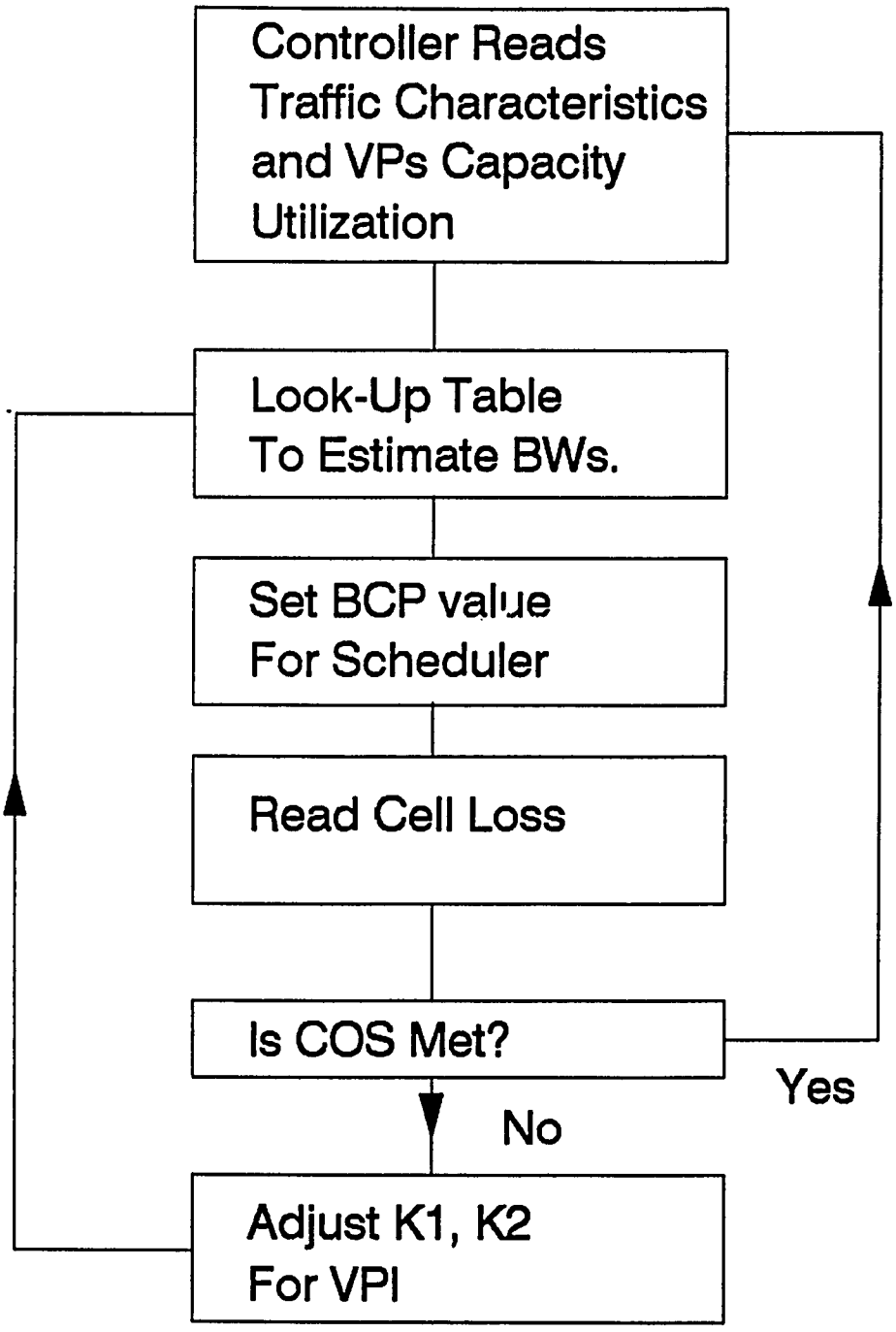
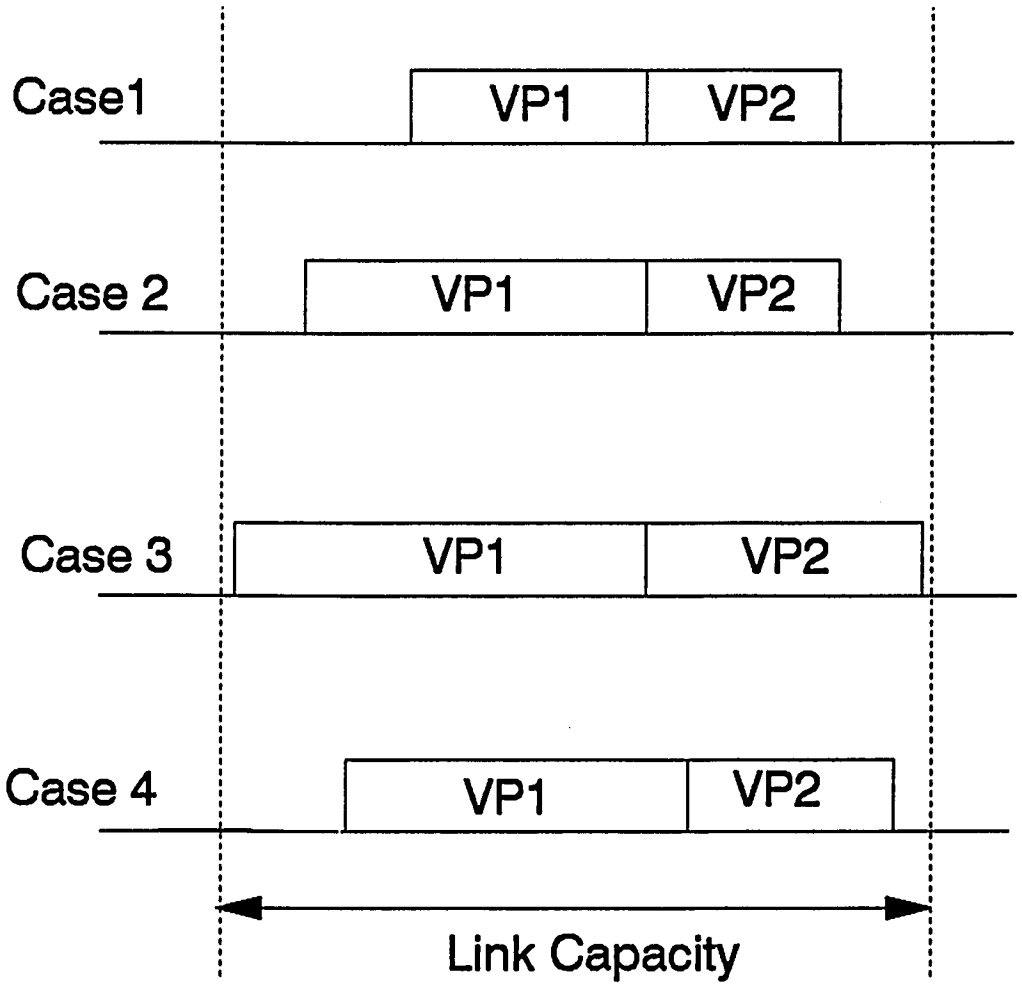


Fig.(V.3) B.W. Allocation Algorithm Flow Chart



Case1: VP1=VP2, BCP=BCP1

Case2: VP1>VP2, BCP2<BCP1

Case3: VP1/VP2 remains as Case2, but BCP3<BCP2

Case4: VP1/VP2 decreased, BCP4>BCP3

Fig.(V.4) Dynamic Bandwidth Control of Virtual Paths

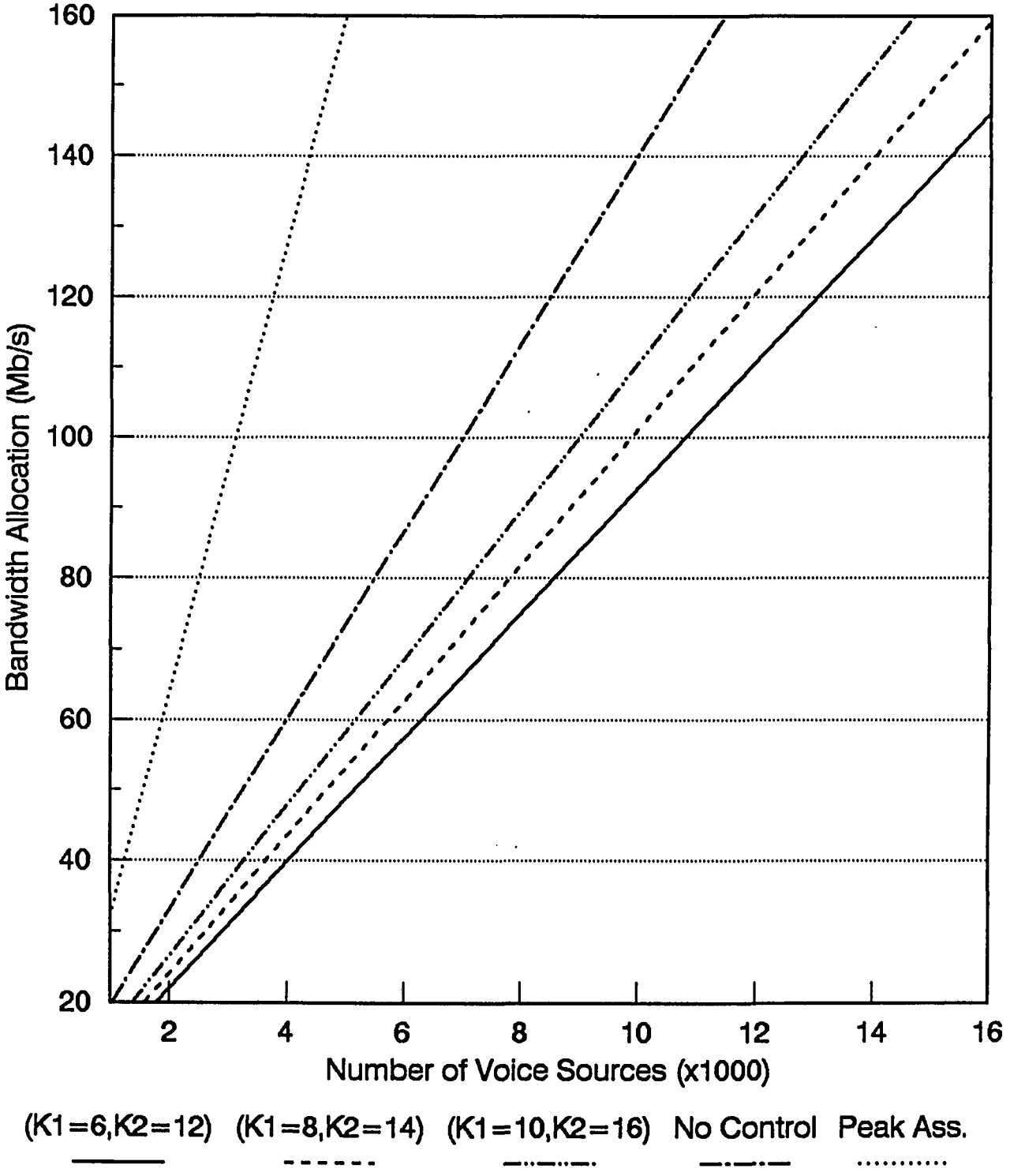


Fig.(V.5) Bandwidth Assignment Vs. Load
Voice Cell Loss Rate =10-4

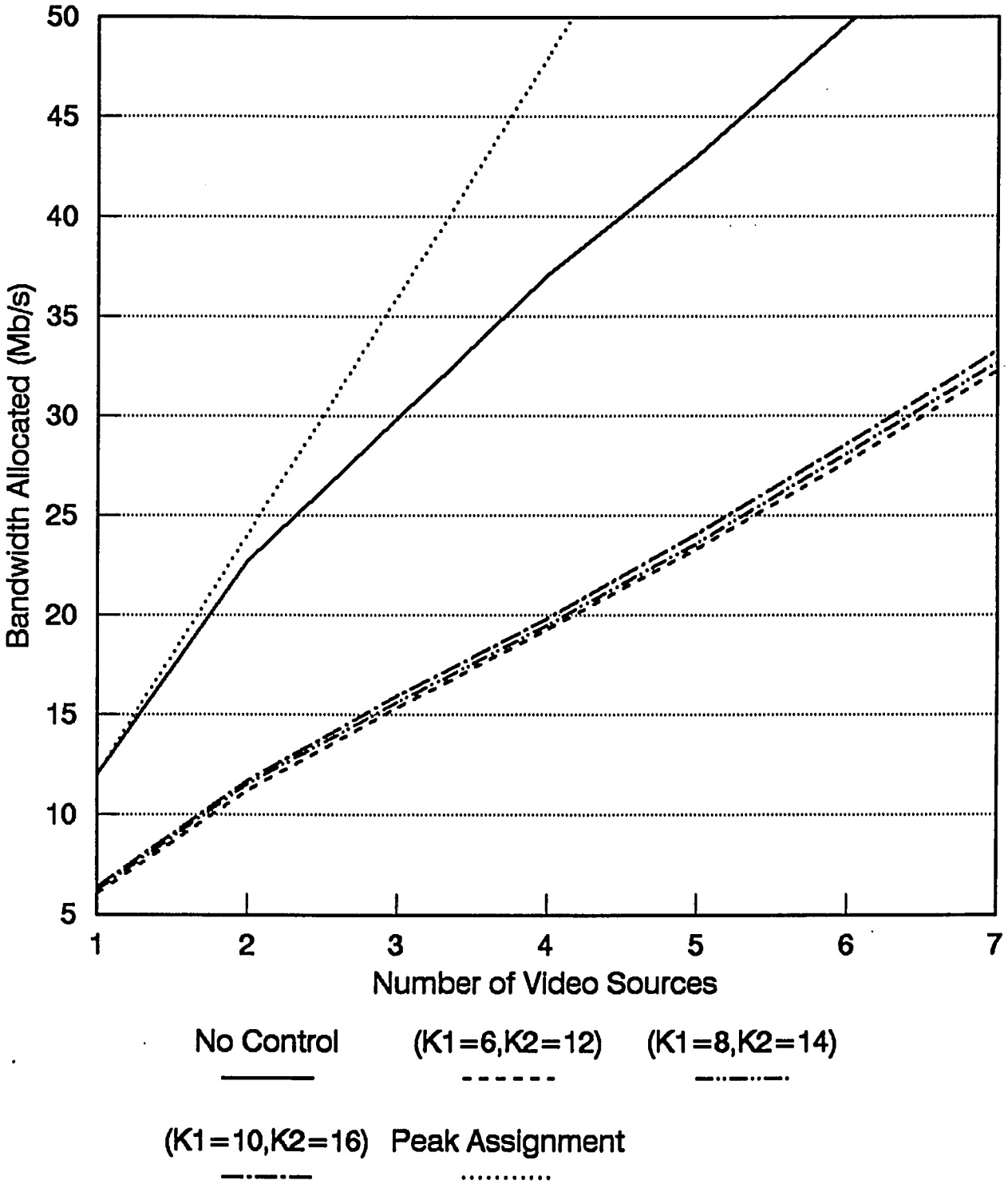


Fig.(V.6) Bandwidth Assignment Vs. Load
Video Cell Loss Rate = 10-9

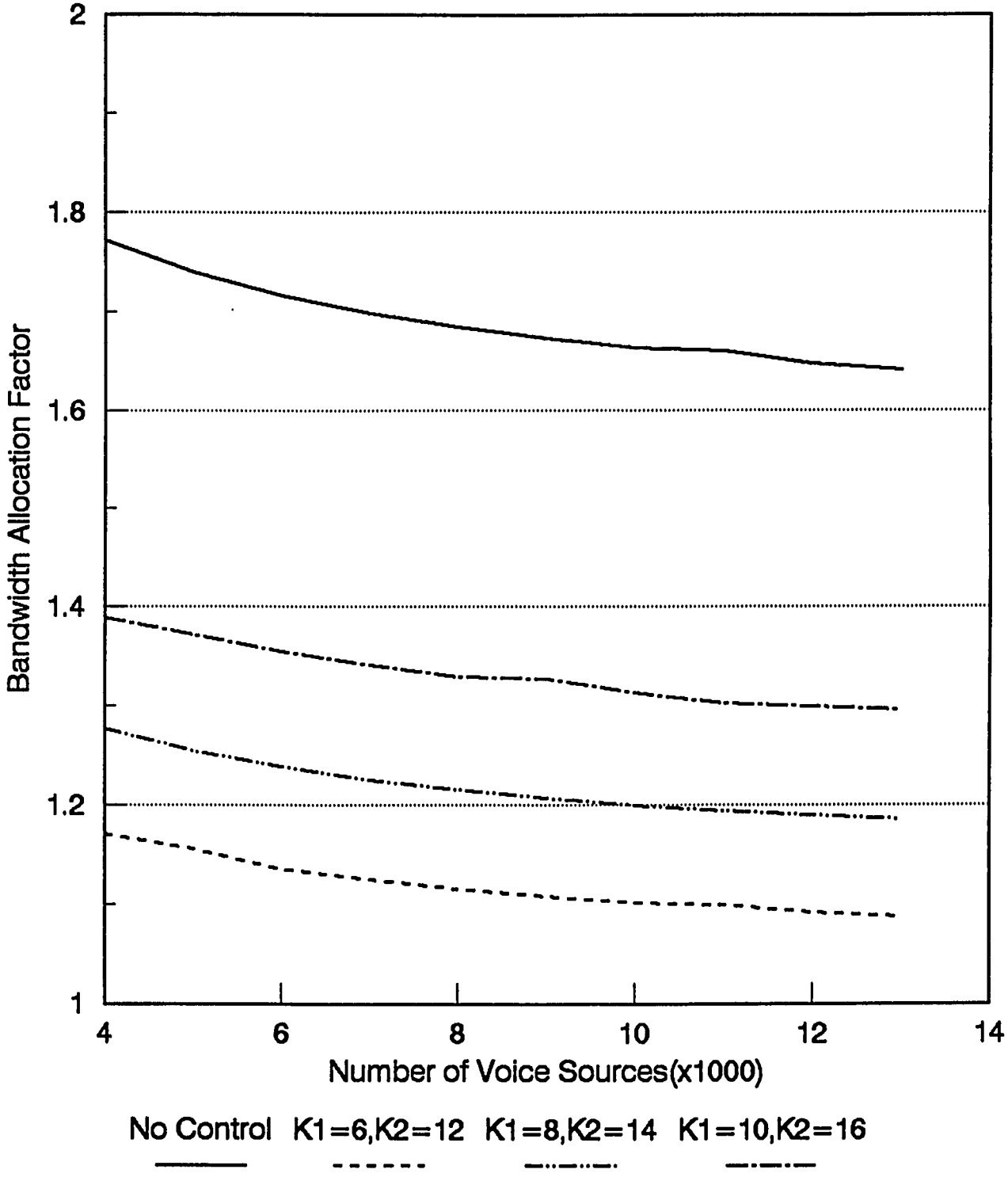


Fig.(V.7) Bandwidth Allocation Factor Vs. Load
Voice Cell Loss Rate = 10-4

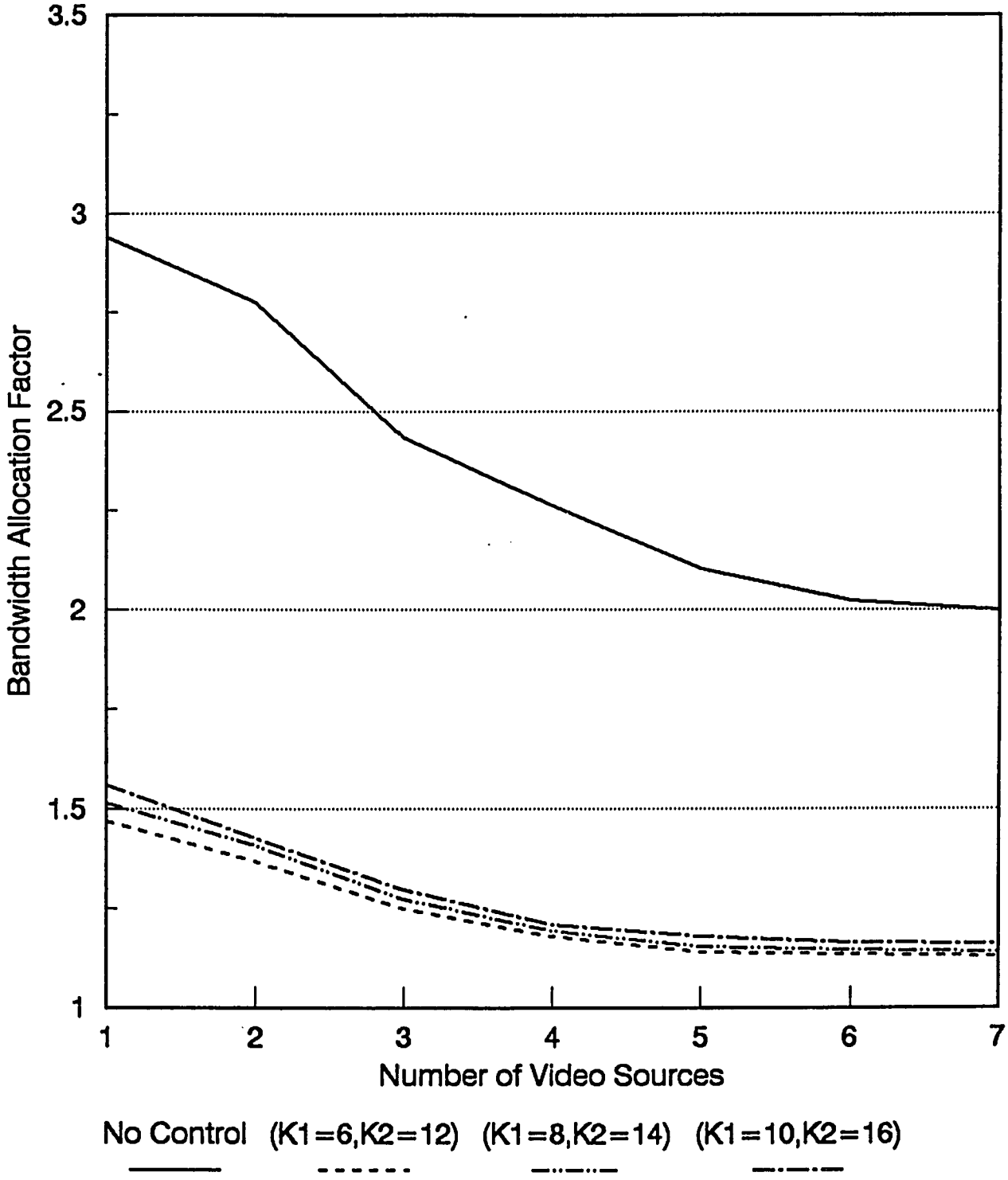


Fig.(V.8) Bandwidth Allocation Factor Vs. Load
Video Cell Loss Rate = 10-9

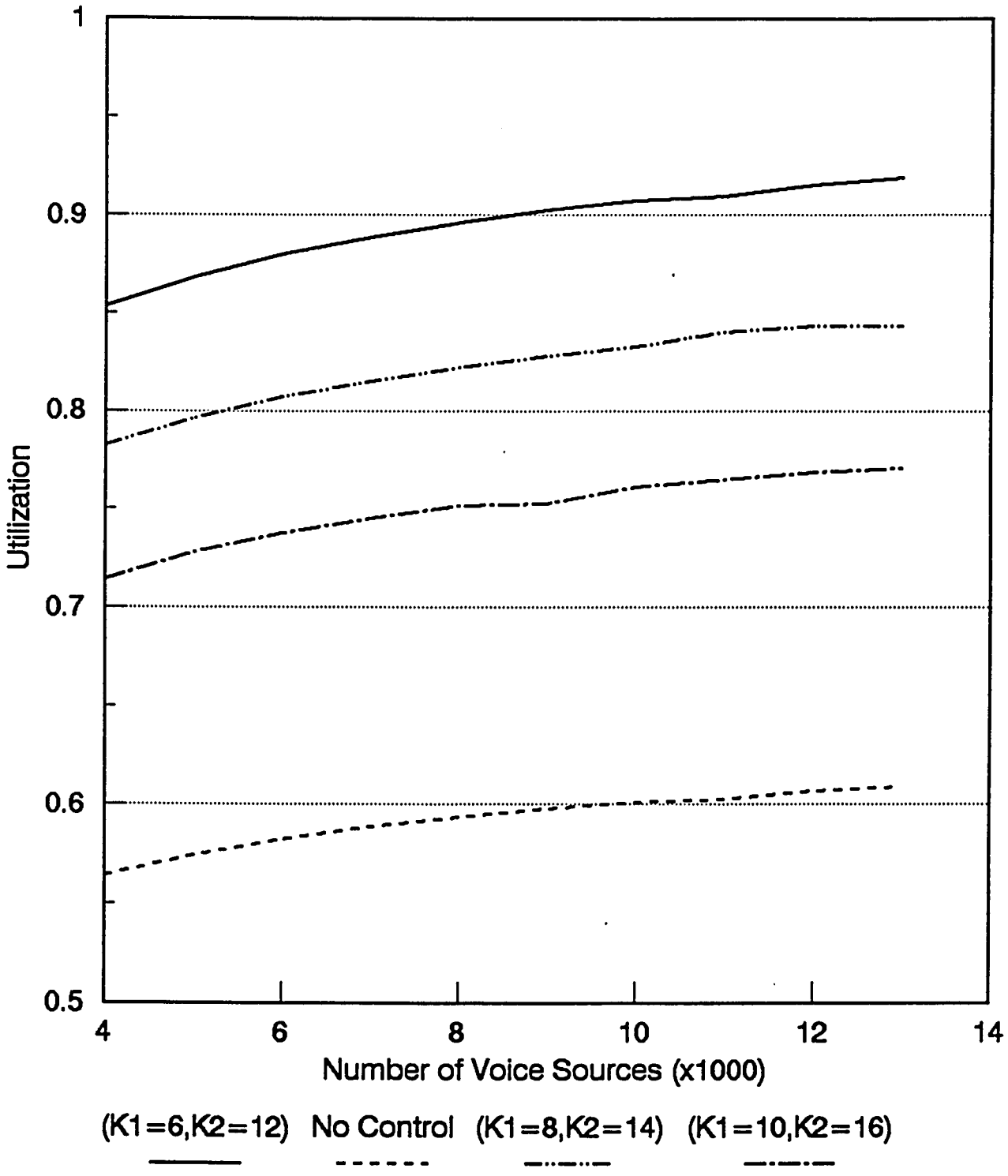


Fig.(V.9) Bandwidth Utilization Vs. Load
Voice Cell Loss Rate = 10-4

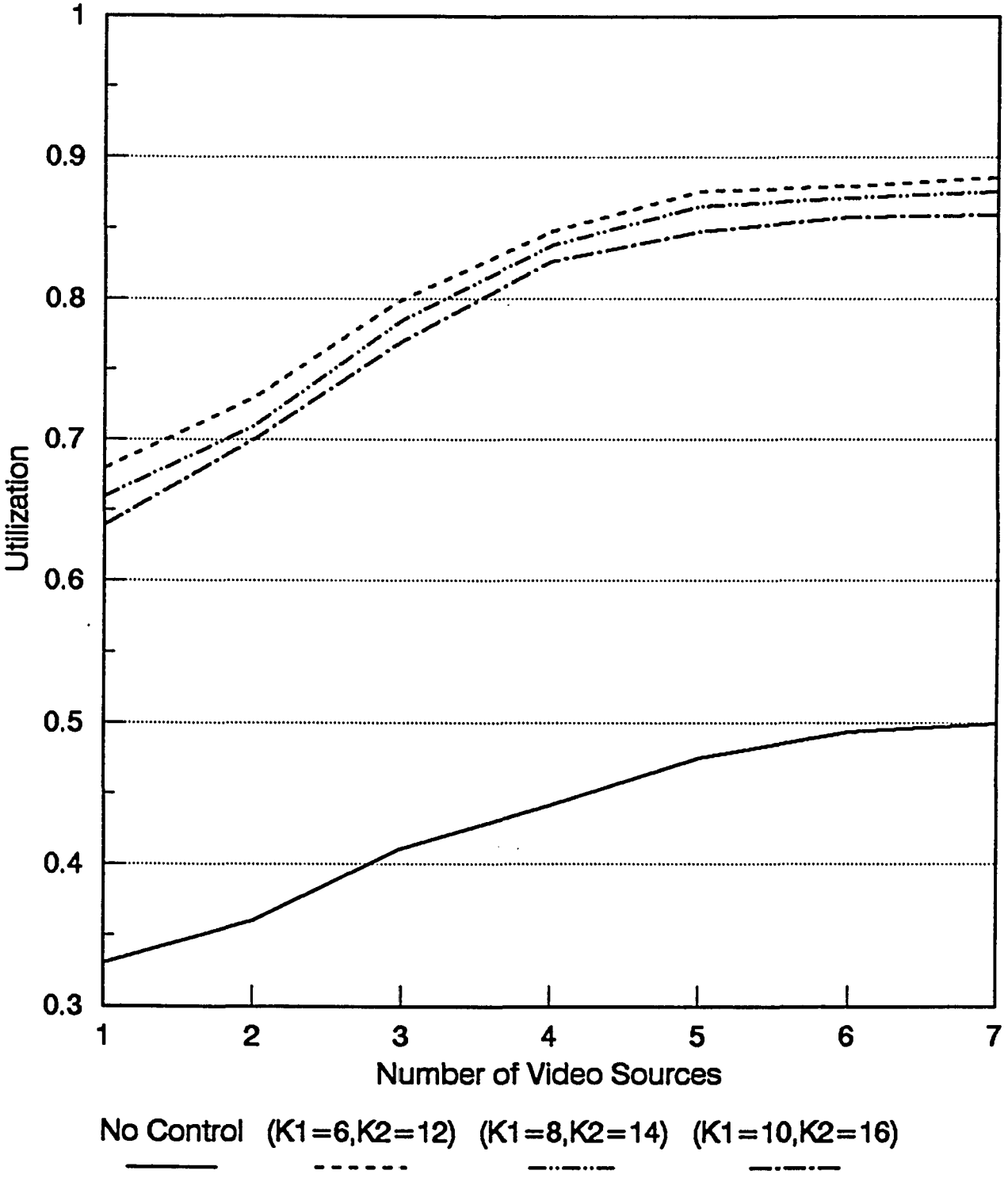


Fig.(V.10) Bandwidth Utilization Vs. Load
Video Cell Loss Rate = 10-9

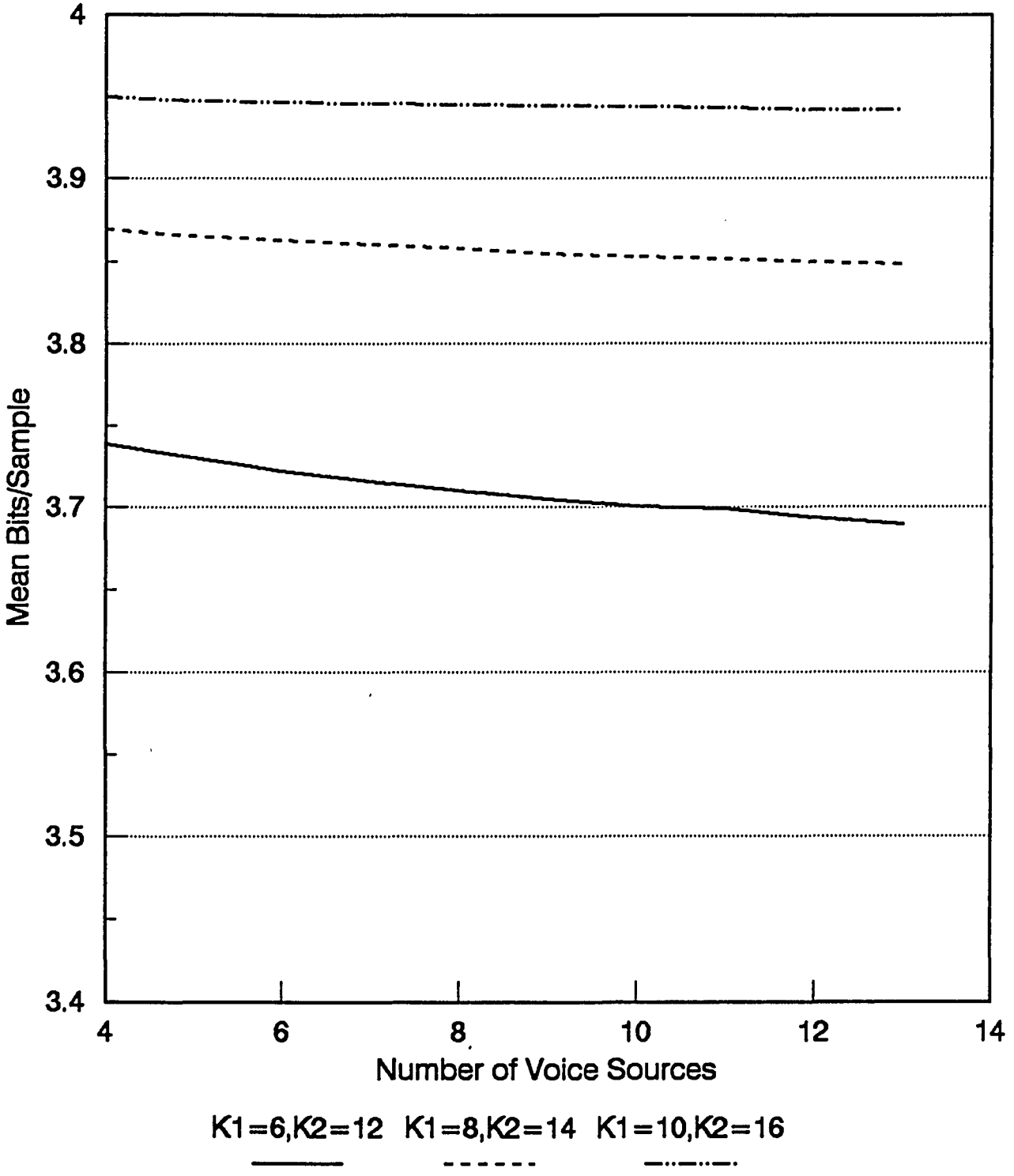


Fig.(V.11) Mean Bits/Sample Vs. Load
Voice Cell Loss Rate = 10-4

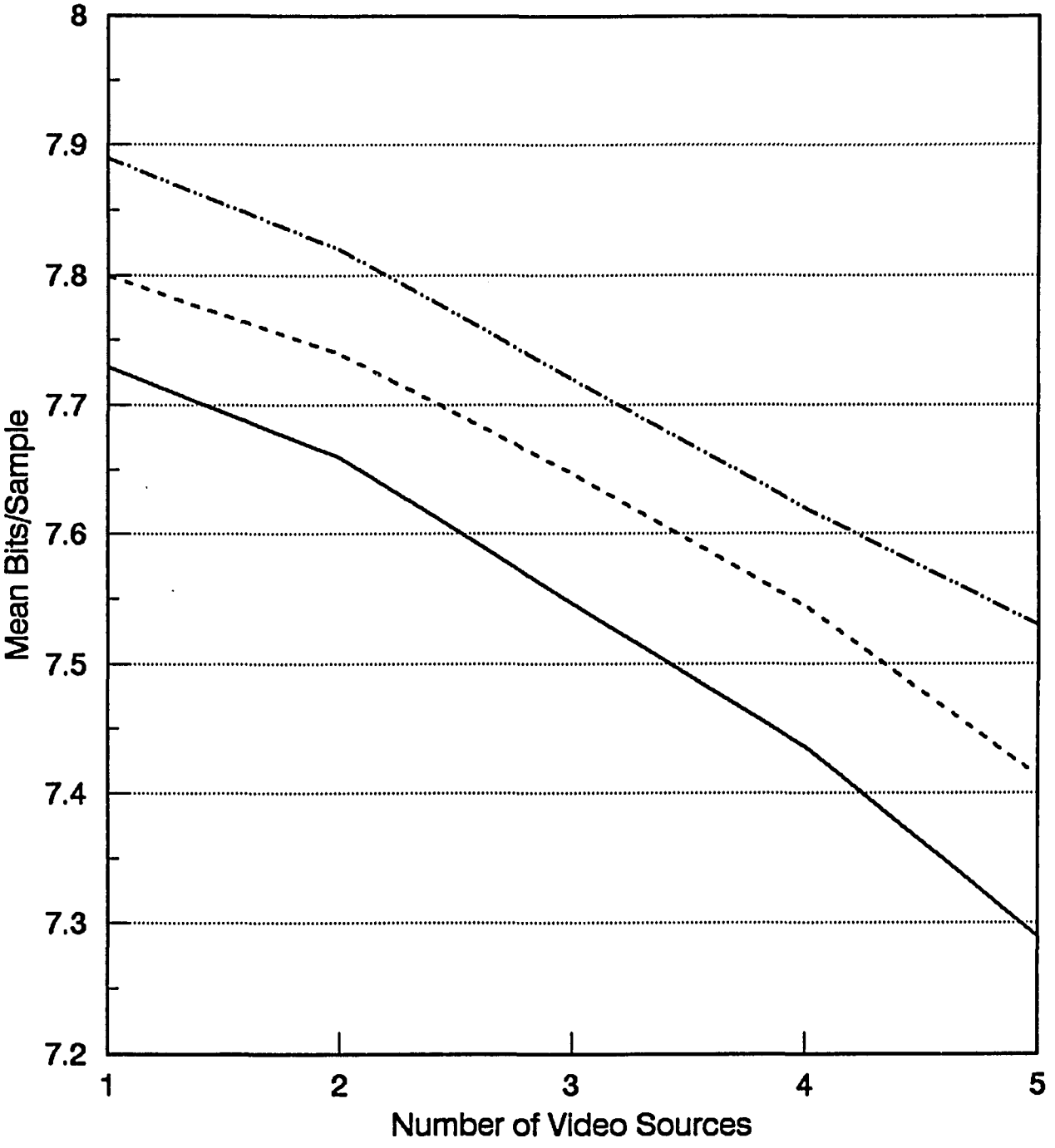


Fig.(V.12) Mean Bits/Sample Vs. Load
Video Cell Loss Rate = 10-9

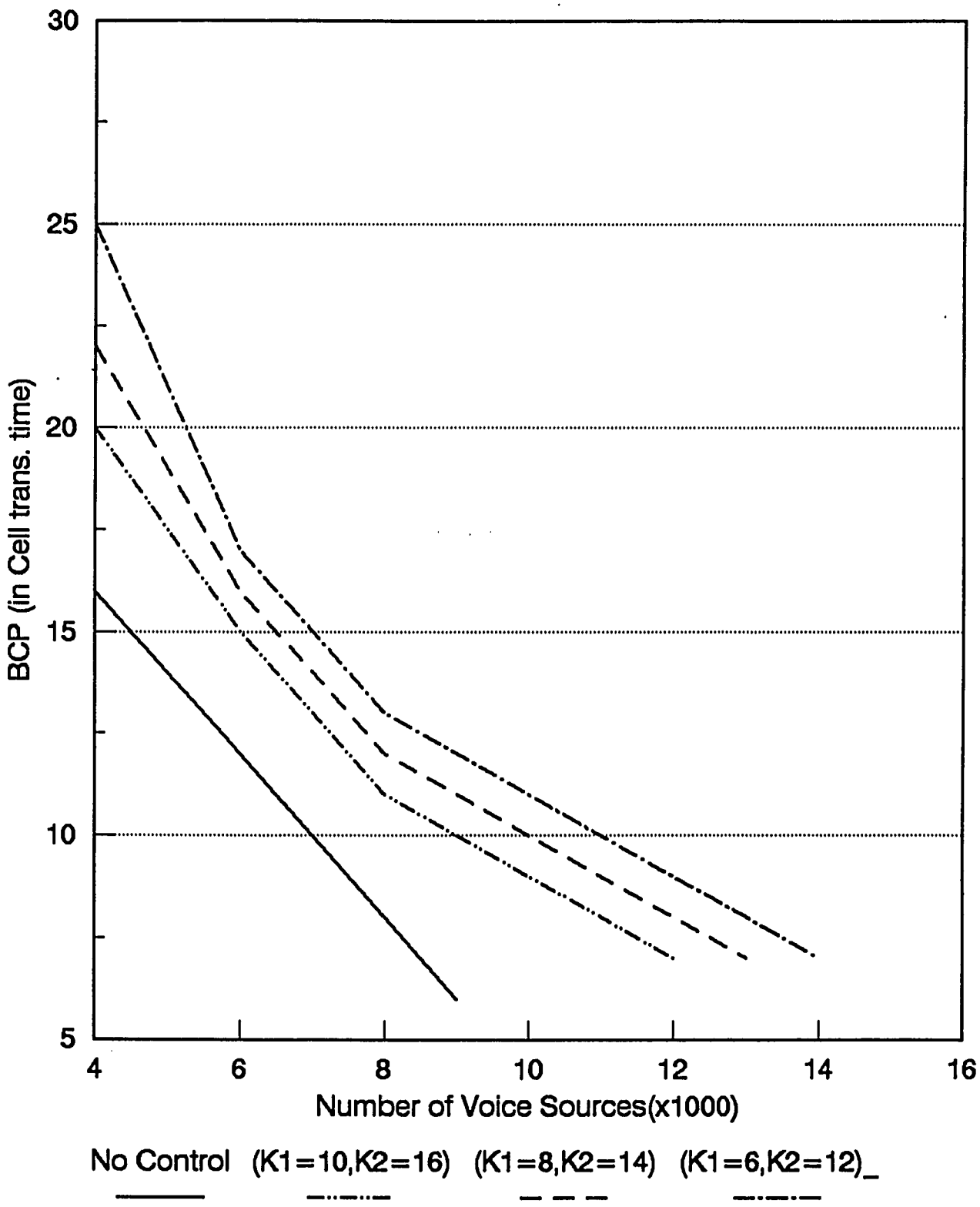


Fig.(V.13) Bandwidth Control Period Vs. Load
Single Video Source (No Video Control)

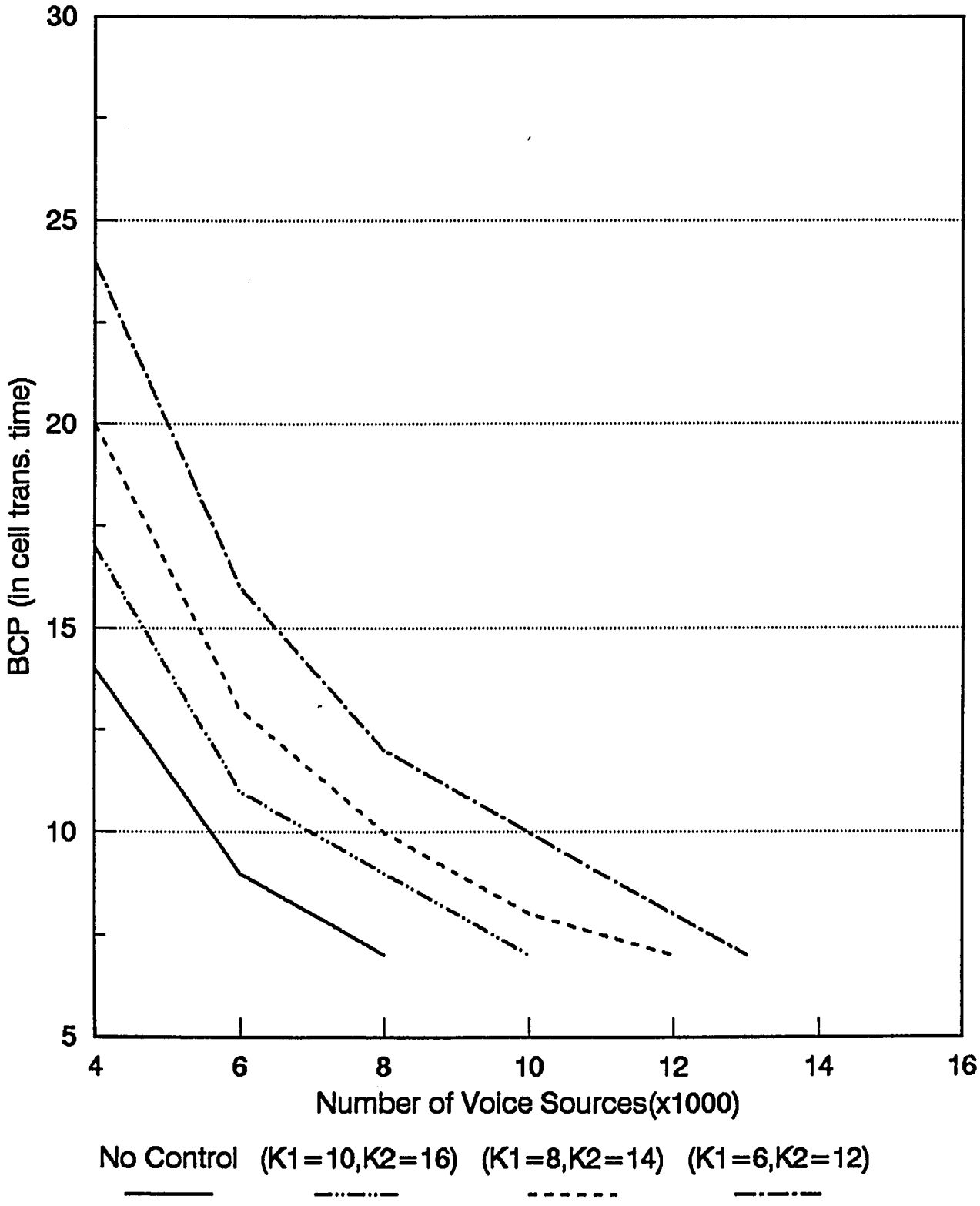


Fig.(V.14) Bandwidth Control Period Vs. Load
Three Video Sources (No Video Control)

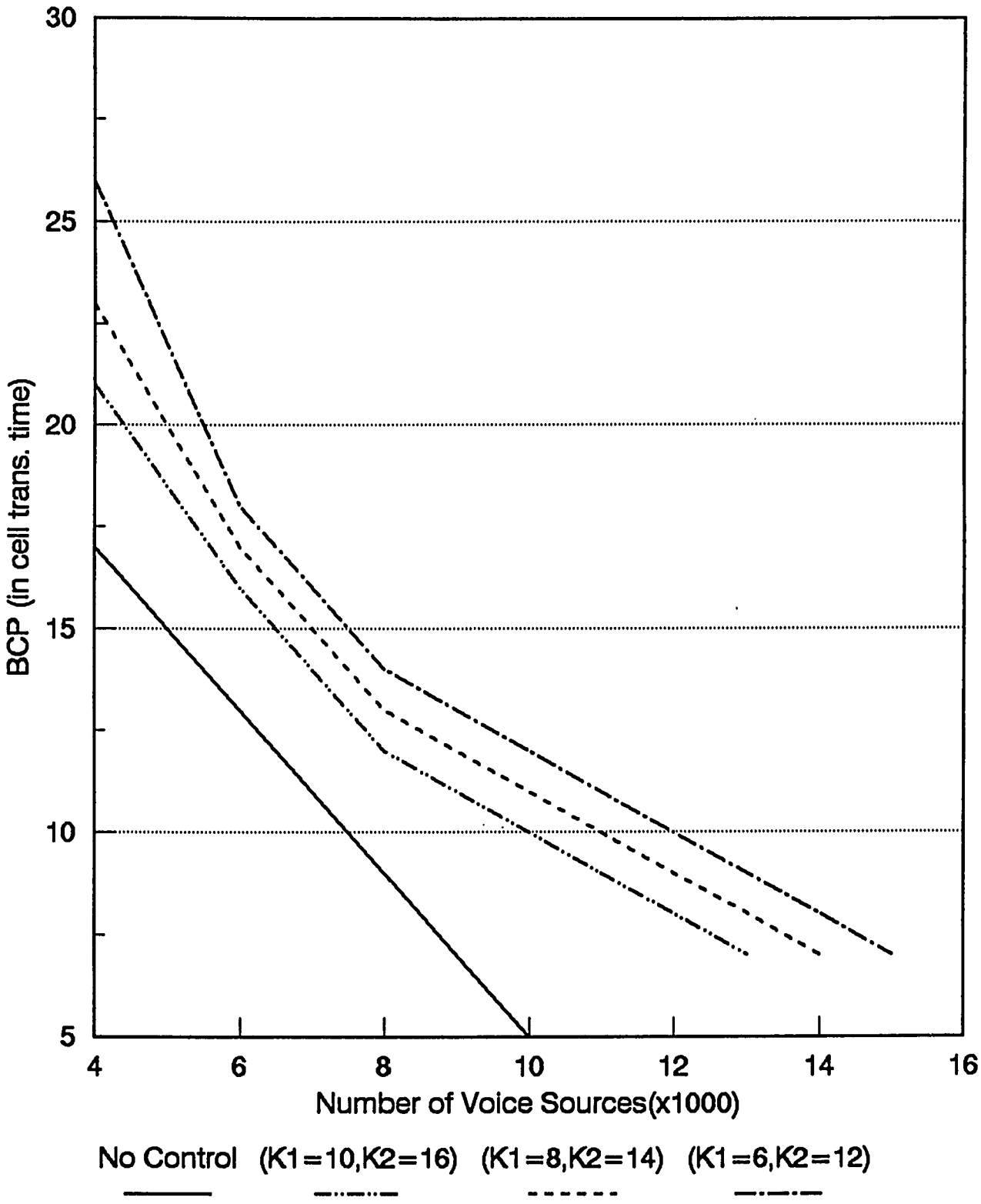


Fig.(V.15) Bandwidth Control Period Vs. Load
Single video Source (K1=6,K2=12)

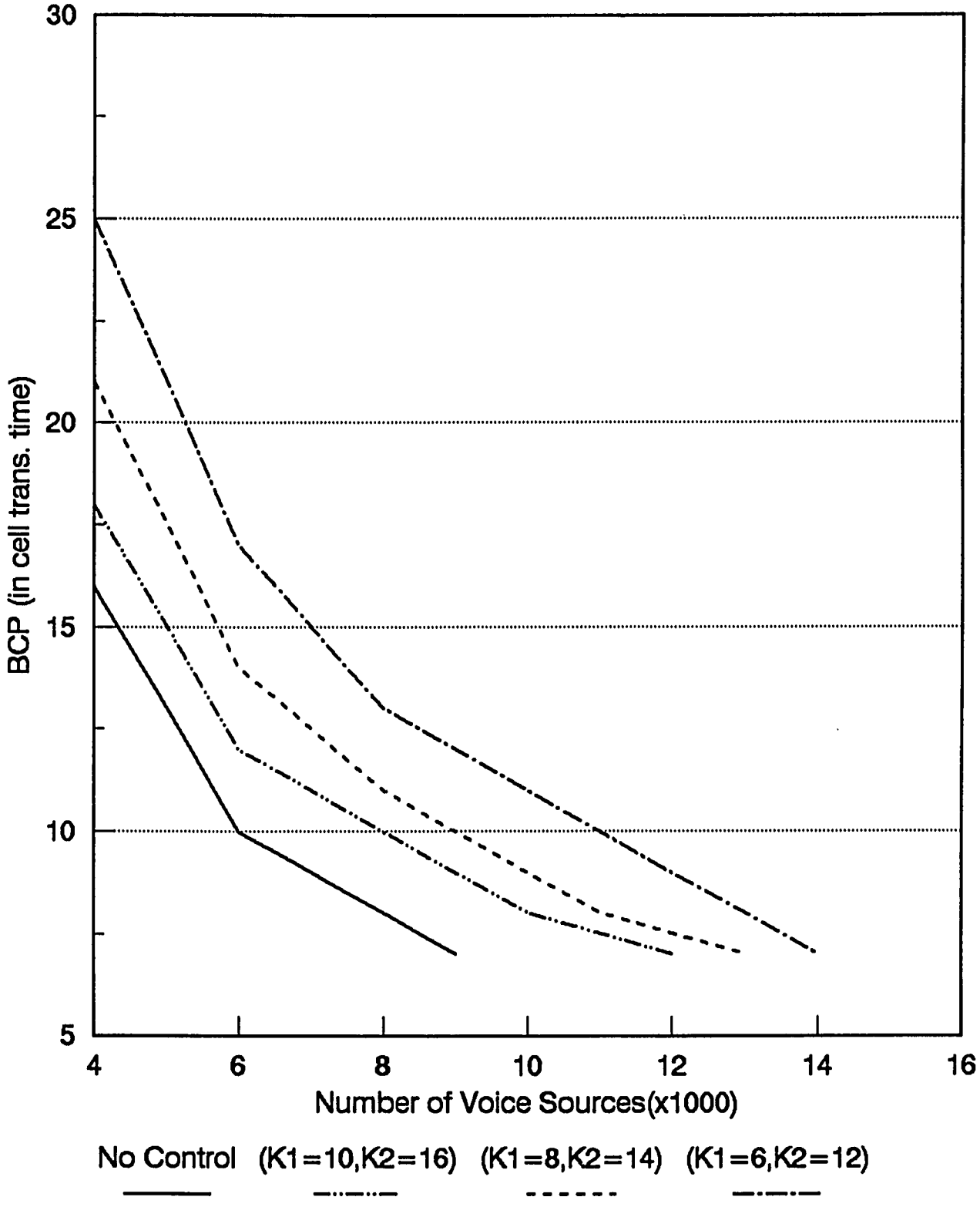


Fig.(V.16) Bandwidth Control Period Vs. Load
Three Video Sources (K1=6,K2=12)

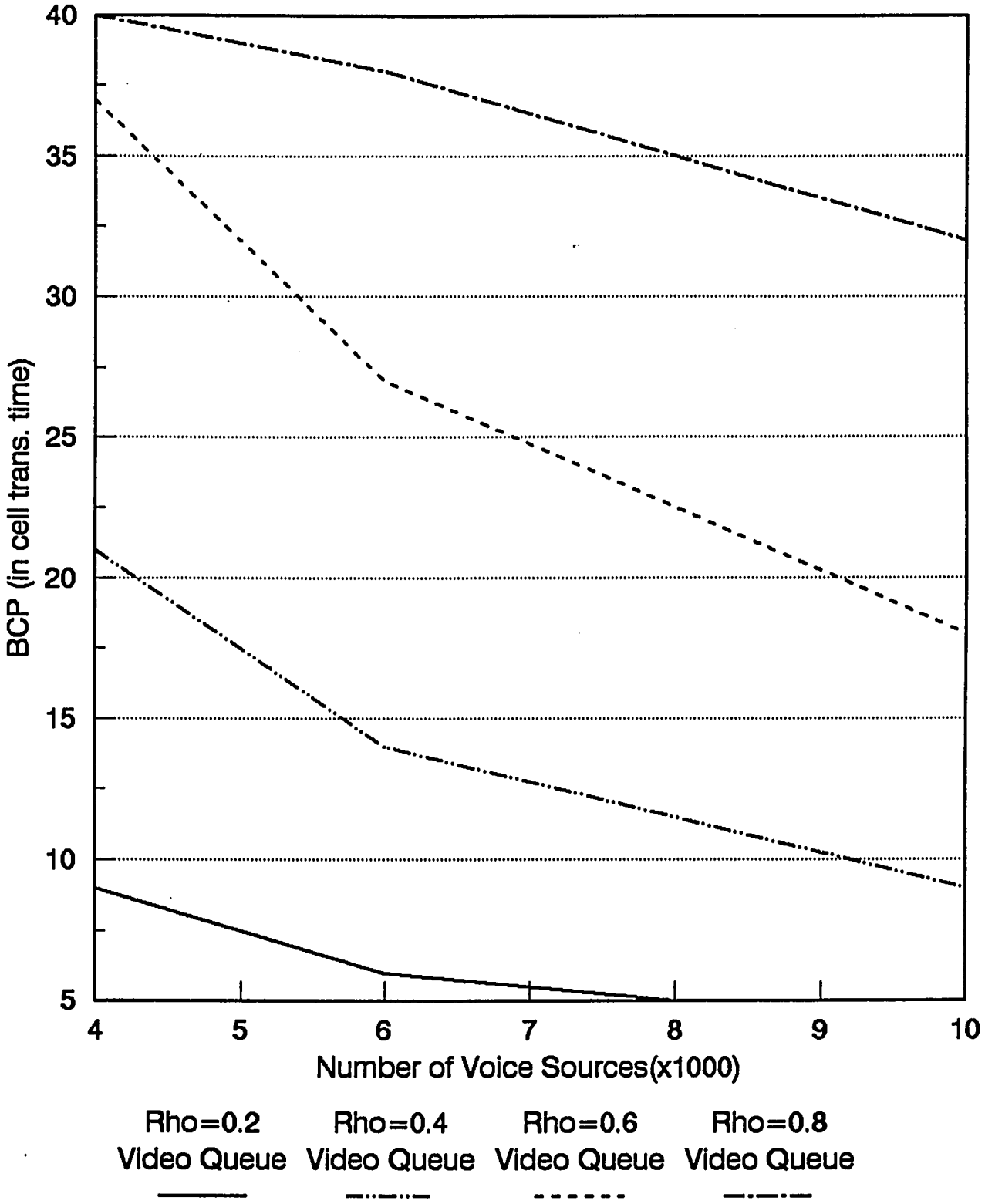


Fig.(V.17) BCP Vs. Video Utilization
Single Video Source (No Control)

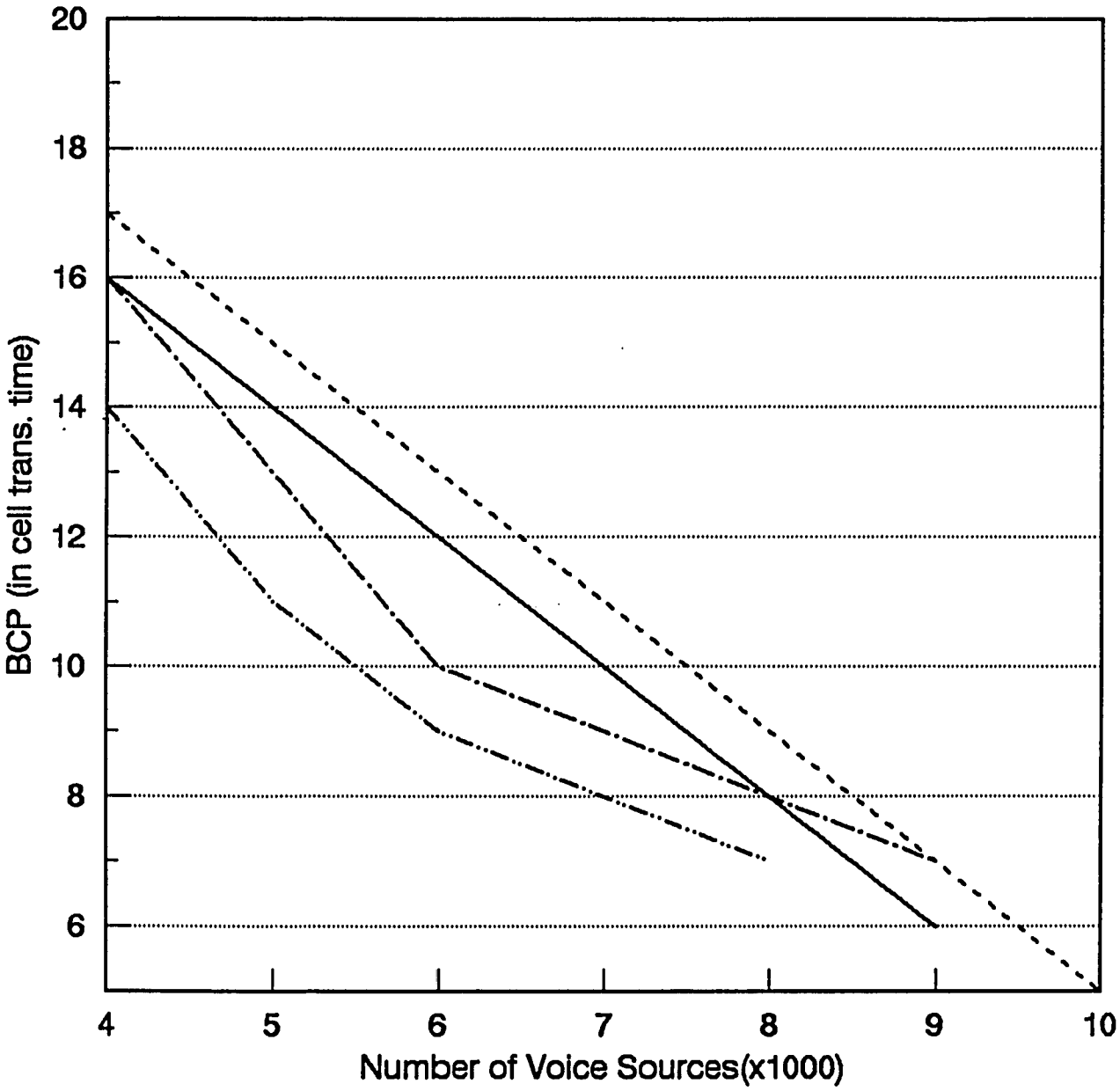


Fig.(V.18) BCP Vs. Voice Load
Voice Buffer with No Flow Control

VI. CONCLUSIONS AND FUTURE RESEARCH

VI.1. Conclusions

Congestion control in ATM networks is one of the most crucial technical issues, for the success of these networks. The complexity of the problem has several dimensions. One dimension of the problem, addresses the issue of optimal allocation of resources to a very diverse mixture of traffic, ranging from low bit rate to high peak rate video signals, each has its own class of service. On another dimension, the problem addresses the issue of optimal utilization of the network resources, and in the mean time maintain fairness to all types of traffic. Finally, any congestion control strategy must be robust, in the sense that it deals with all types of traffic in a similar manner and can thus accommodate any future traffic that is unforeseen at the present time. To design an ideal scheme, which would accomplish the above, is not an easy task, because of the transport principle of the ATM networks which is based upon the fast packet switching. ATM networks can be thought of as a virtual direct connection between two ends (similar to a high way road). Very little traffic management can be done on the transit nodes inside the network

(similar to high ways where there are no traffic lights). Access flow control must be very skilfully designed to avoid congestion inside the network.

In this dissertation, we have proposed and analyzed a hierarchical multilevel flow control model. At the cell level, we have found that an access flow control algorithm, that throttles the peak rate of the input superposition arrival process, is the one of the most efficient methods to implement a robust preventive control strategy for ATM networks. At the call and network levels, we proposed and analyzed a dynamic bandwidth control scheme which that has increased the bandwidth utilization significantly. The scheme is based upon the design of a Bandwidth Control Period (BCP) which allocates the bandwidth, to virtual paths and calls, on a statistical assignment basis. The results, reported in this dissertation, are very promising. There is a big doubt in the technical community, as to the efficacious of employing variable bit rate (VBR) coding methods to transmit video signals in ATM networks. Our reported results here, prove that the transmission of VBR video signals, using ATM, is indeed a very promising technique. However an access flow control algorithm together with a bandwidth allocation scheme, such as our algorithms reported in this dissertation, are essential in order to avoid congestion problems inside the network.

VI.2. Future Research

Our work calls for some very interesting topics for future research.

1. In our analysis, we have analyzed the queueing behavior of the multiplexers' buffers using steady state analysis. It is very interesting, to carry out a time-dependent analysis and verify the duration of a control period (t), below which the control thresholds are not activated. This analysis is necessary to differentiate the overload which can result from, the arrival of a long burst, or from normal statistical fluctuations.

2. It is very interesting to implement the controller using Neural Networks (NN). As shown, in chapters III to V, there are so many input output parameters that influence the choice of the threshold control levels and the BCP period. Look-up tables would be too complicated and may cause processing overhead. Neural networks, would be a better alternative, since they are best suited to solve pattern recognition problems with extensive processing capabilities. The congestion control problem, would then be an ideal application.

3. How would the performance of the multiplexer be affected, if the BCP rule was combined with an upper maximum limit on

the service time per queue. In other words, if we relax the assumption that the BCP values are drawn from an exponentially distributed density function, will the performance improve and what are the tradeoffs in terms of system complexity.

4. The multiplexer output stochastic process is often considered to be an input arrival process to the ATM switch. Under the bandwidth allocation strategy proposed here, what would be the stochastic nature of such process. How would the switch perform under such conditions.

5. An ATM end to end system analysis, is then required to assess the efficacious of the traffic management functions per access nodes and transit nodes under real traffic conditions (i.e. a combined video/voice/data stream).

BIBLIOGRAPHY

1. J.P. Coudreuse, A. Thomas, M. Servel "ATD techniques: An experimental packet network integrating videocommunication" ISS'84.

2. J.S. Turner, L.F. Wyatt " A Packet Network Architecture for Integrated Services ", Proc. GLOBECOMM 83.

3. M. De Prycker, P. Plehiers, M. Fastrez, J. Bauwens "Evolution towards a Belgian Broadband Experiment" ISS'87.

4. M. De Prycker "Definition of Network Options for Belgian ATM Broadband Experiment" IEEE J. Sec. Areas. Comm. Dec.88.

5. M. Wernik "Architecture and Technology Considerations for Multimedia Broadband Communications" Proc. GLOBECOMM 88.

6. M. Rider "Protocols for ATM Access Networks" Proc. GLOBECOMM 88.

7. B. Amin-Salehi, D. Spears "Support of Transport Services in BISDN" Proc. GLOBECOMM 89.

8. H. Ishii, M. Kawarasaki "BISDN Signaling Protocol Capabilities" Proc. GLOBECOMM 89.

9. B. Eklundh, I. Gard, G. Leijonhufvud "A Layered Architecture for ATM Networks" Proc. GLOBECOMM 89.

10. K. Takahashi, T. Yokoi, Y. Yamamoto "Communications Quality Analysis for ATM Networks" Proc. ICC 89.

11. J. Hui "Network, Transport, and Switching Integration for Broadband Communications" IEEE Networks Magazine March 89.

12. J.P. Coudreuse "ATM Status of Definition and Discussion of Some Open Issues" Proc. Mutimedia 89.

13. S. Yoneda "Broadband ISDN ATM Layer Management: Operations, Administration, and Maintenance Considerations" IEEE Networks Magazine May 90.

14. G. Fioretti, T. Demaria, F. Perardi, L. Piovano "ATM Based Network Transport Service" Proc. ICC 90.

15. CCITT/XVIII NTT contribution D.77 "Adaptation Layer Header Size"

16. CCITT/XVIII Italian Contribution D.353, D.354, D.355 , D.356, Geneva 89.

17. CCITT/XVIII SWP 8/1 ATM Final Report Geneva 89.

18. T1S1.1/89-063 "UNI Structure, Sonet Mapping for ATM" March 89.

19. S. Minzer "BISDN and Asynchronous Transfer Mode" IEEE Communications Magazine September 89.

20. A. Hac, H. Mutlu "Synchronous Optical Network and BISDN protocols" Computer, Vol.22, November 89.

21. R. Ballart, Y. Ching "SONET: Now It's the Standard Optical Network" IEEE Communications Magazine March 89.

22. J. Gruber, K. Nagaraj, J. Leeson, B. Fleury "Improvements in Availability and Error Performance of SONET compared to Asynchronous Transport Systems" Proc. ICC 90.

23. S. E. Minzer "Broadband User Network Interfaces to ISDN" Proc. IEEE ICC 87.

24. J.Y. Hui "Resource Allocation for Broadband Networks" IEEE

J. Sec. Areas Comm. Dec. 88

25. G. Woodruff, R. Rogers, P. Richards "Congestion Control Framework for High Speed Integrated Packetized Transport" Proc. GLOBECOMM 88.

26. T. Takahashi, A. Hiramatsu "Integrated ATM Traffic Control by Neural Networks" ISS'90.

27. A. E. Eckberg, D. Lucantoni, D. Luan "Meeting the Challenge: Congestion and Flow Control Strategies for Broadband Information Transport" Proc. GLOBECOMM 89.

28. A. Hiramatsu "ATM Communications Network Control by Neural Networks" International Joint Conference on Neural Networks June 89.

29. V. Jacobson "Congestion Avoidance and Control" Proc. ACM SIGCOMM 88.

30. C. Cooper, K. Park "Toward a Broadband Congestion Control Strategy" IEEE Networks Magazine May 90.

31. M. Gerla, L. Kleinrock "Flow Control: A Comparative Survey" IEEE Trans. Comm. April 80.

32. Raj Jain "Congestion Control in Computer Networks: Issues and Trends" IEEE Networks Magazine May 90.

33. K. Sato, S. Ohta, I. Tokizawa "Broadband ATM Network Architecture Based on Virtual Paths" IEEE Transactions on Communications August 90, also in part in GLOBECOMM 88 and ISSLS 88.

34. K. Sato, I. Tokizawa " Flexible Asynchronous Transfer Mode Networks Uitilizing Virtual Paths" Proc. ICC 90.

35. K. Noguchi, T. Okada, H. Ohnishi "Resource Management in ATM Networks" Proc. Second IEEE COMSOC International MUTIMEDIA Communications Workshop April 89.

36. T. Murase, H. Suzuki, T. Takeuchi "Continuous Bit Stream Oriented Services in ATM Networks" Proc. MUTIMEDIA Workshop April 89.

37. W. Wang, T. Saadawi, K. Aihara "Bandwidth Variation and Control for ATM Networks" Proc. ICC 90.

38. T. Kamitake, T. Suda "Evaluation of Admission Control Scheme For ATM Network Considering Fluctuations in Cell Loss Rate" Proc. GLOBECOMM 89.

39. M. Decina, T. Toniatti, P. Vaccari, L. Verri "Bandwidth Assignment and Virtual Call Blocking in ATM Networks" Proc. INFOCOMM. 90.

40. G. Gallasi, G. Rigolio, L. Fratta "ATM: Bandwidth Assignment and Enforcement Policies" Proc. GLOBECOMM 89.

41. M. Decina, T. Toniatti "On Bandwidth Allocation to Virtual Bursty Connections in ATM Networks" Proc. ICC 90.

42. H. Ohnishi, T. Okada, and K. Noguchi " Flow Control Schemes and Delay-Loss tradeoff in ATM Networks. " IEEE J. Sec Areas Comm. Dec. 88.

43. W. Verbiest, L. Pinnoo, B. Voeten "Statistical Multiplexing of Variable Bit Rate Video Sources in ATM Networks" Proc. GLOBECOMM 88

44. W. Verbiest, L. Pinnoo "The Impact of ATM Concept on Video Coding" IEEE J. Sec. Areas Commun. Dec. 88.

45. W. Verbiest, L. Pinno "A Statistical Bandwidth Allocation And Usage Monitoring Algorithm For ATM Networks" Proc. ICC 89.

46. M. Kateveneis "Fast Switching and Fair Control of Congested Flow in Broadband Networks" IEEE J. Sec. Areas Comm. Oct. 87.

47. K. Nakamaki, M. Kawakatsu, A. Notoya "Traffic Control for ATM Networks" Proc. MULTIMEDIA April 89.

48. I. Habib, T. Saadawi "Access Flow Control Algorithms in Video and Voice Packet Multiplexers in Broadband Networks" Submitted to Performance Evaluation Special Issue on High Speed Transmission Systems.

49. K. Sriram "Dynamic Bandwidth Allocation and Congestion Control Schemes for Voice and Data Multiplexing in Wideband Packet technology" Proc. ICC 90.

50. B. Kraimeche, M. Schwartz "Bandwidth Allocation Strategies in Wideband Networks" IEEE J. Sec. Areas Comm. Sept. 86.

51. K. Sriram, W. Whitt "Characterizing Superposition Arrival Processes in Packet Multiplexers for Voice and Data" IEEE J. Sec. Areas Comm. Sept.86.

52. K. Sriram, D.M. Lucantoni "Traffic Smoothing Effects of Bit Dropping in a Packet Voice Multiplexer" IEEE Transactions on Communications July 89.

53. B. Maglarias, D. Anastassiou, P. Sen, G. Karlsson, J. Robbins "Performance Models of Statistical Multiplexing in Packet Video Communications" IEEE Transactions on Communications July 88.

54. H. Saito, M. Kawarazaki, H. Yamada "Analysis of Statistical

Multiplexing in ATM Transport Networks" Proc. ICC 90.

55. H. Heffes, M. Lucantoni "A Markov Modulated Characterization of packetized Voice and Data Traffic and Related multiplexer Performance" IEEE J. Sec. Areas in Commun. Sept. 86.

56. M. Hirano, N. Watanabe "Characteristics of a Cell Multiplexer for Bursty ATM Traffic" Proc. ICC 89.

57. J. Daigle, J. Langford "Models for Analysis of Packet Voice Communications Systems" IEEE J. Sec. Areas Commun. Sept. 86.

58. P.T. Brady "A model for generating on-off speech patterns in two way conversations" Bell Syst. Tech. J. Vol. 48, Sept. 69.

59. C. J. Weinstein "Fractional Speech Loss and Talker Activity Model for TASI and for packet switched speech" IEEE Trans. Commun. Sept. 78.

60. M. Nomura, T. Fujii, N. Ohta "Basic Characteristics of Variable Rate Video Coding in ATM Enviroment" IEEE J. Sec. Areas. Comm. June 89

61. M. Nomura, T. Fujii, N. Ohta "Characteristics of Variable Rate Video Coding with Motion Compensated DCT for Burst/packetized Communications" Proc. Int. Workshop Future Prospects Burst/packetized Mutimedia Communications Nov.87.

62. F. Kishino, K. Manabe, Y. Hayashi, H. Yasuda "Variable Bit Rate Coding of Video Signals for ATM Networks" IEEE J. Sec. Areas Commun. June 89.

63. Y. Yasuda, H. Yasuda, N. Ohta, F. Kishino "Packet Video Transmission Through ATM Networks" Proc. GLOBECOMM 89.

64. T. Tsuda, S. Maki "Improvement of Picture Quality by Variable Rate Coding" Proc. Int. Workshop Packet Video Sept.88.

65. G. Karlsson, M. Vetterli, "Packet Video and Its Integration into the Network Architecture" IEEE J. Sec. Areas Commun. June 89.

66. S. Huang "Modeling and Analysis for Packet Video" Proc. GLOBECOMM 89.

67. D.J. Goodman "Embedded DPCM for Variable Bit Rate Transmission" IEEE Transactions on Communications July 80.

68. I. Habib, T. Saadawi "Flow Control Techniques in Packet Voice Multiplexers" Proc. ISMM International Conference on Parallel and Distributed Computing.

69. J. Turner "The Challenge of Multipoint Communication" Proc. ITC Sem. on Traffic Eng. for ISDN Design and Planning 87.

70. M. Sidi, W. Liu, I. Cidon, I. Gopal "Congestion Control through Input Rate Regulation" Proc. GLOBECOMM 89.

71. I. Cidon, K. Sohraby, K. Bala "Congestion Control for High Speed Packet Switched Networks" Proc. INFOCOMM 90.

72 S. Golestani "Congestion Free Transmission of Real Time Traffic in Packet Networks" Proc. INFOCOMM 90

73. I. Habib, T. Saadawi "Dynamic Bandwith Allocation and Congestion Control of Virtual Paths in ATM Networks" submitted to IEEE Trans. Comm., also to INFOCOMM'92.

74. I. Habib, T. Saadawi "Congestion Control in Video Multiplexers" Proceedings IFIP 3rd Conference on High Speed Networks, Berlin, Germany, March 91.

75. S.-Qi. Li "A Study of Information Loss in Packet Voice Systems" IEEE Transacs. On Comm. Nov.89.

76. S.-Q. Li, J. Mark "Traffic characterization for Integrated Services Networks" IEEE Transcs. On Comm. Aug.90.

77. J. Daigle, D. Lucantoni "Queueing Systems Having Phase Depandant Arrival and Service Rates" Proc. First Int. Workshop on Numerical Solution of Markov Chains Jan. 90.

78. D. Lucantoni "New Results on The Single Server Queue with Batch Markovian Arrival Process" to appear in Stochastic Models 91.

79. H. Yamada, F. Machihara "Performance Evaluation of Statistical Multiplexer with Control on Input and/or Service Process" to appear in Performance Evaluation 91.

80. F. Machihara "A New Approach to the Fundamental Period of a Queue with Phase Type Markov Renewal Arrivals" in Commun. Statist.-Stochastic Models, 6(3), pp.551-560, 1990.

81. M. Neuts "Matrix Geometric Solutions in Stochastic Models: An Algorithmic Approach" John Hopkins University Press 1981.

82. D. Gross & C. Harris "Fundementals of Queueing Theory", N.Y., J. Wiley, 85, 2nd edition

Ibrahim Wahby Habib was born in Cairo, Egypt, in 1959. He received the B.Sc. degree, in Electrical Engineering, from Ain Shams University, Cairo, in 1981. He received the M.Sc. degree, in Electrical Engineering, from Polytechnic University of New York, in 1984. He received the Ph.D. degree, in Electrical Engineering, from the City University of New York, in 1991. His research interests are in the areas of modeling and performance evaluation of high speed network, fast packet switching, neural networks and fiber-optics local area networks.