INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each original is also photographed in one exposure and is included in reduced form at the back of the book.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

UMI®

Bell & Howell Information and Learning
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA
800-521-0600

Modeling and Analysis of Self-similar Traffic in ATM Networks

Rajab Faraj

A Thesis

in

The Department

of

Electrical and Computer Engineering

Presented in Fulfillment of the Requirements

for the Degree of Doctor of Philosophy at

Concordia University

Montreal, Quebec, Canada

June 2000

National Library of Canada
Acquisitions and Bibliographic Services
395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque nationale du Canada
Acquisitions et services bibliographiques
395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence

Our file Notre référence

0-612-47710-X

Canada

ABSTRACT

Modeling and Analysis of Self-similar Traffic in ATM Networks

Rajab M. A. Faraj, Ph.D.

Concordia University, 2000.

ATM is considered by the International Consultative Committee for Telephone and Telegraph (CCITT) as the preferred transfer mode for B-ISDN. Both the need for flexible networks and the progress in technology and system concepts led to the definition of the ATM principle. ATM will provide the means to transport, at broadband rates, the traffic generated by a wide range of multimedia services. ATM is suitable for the multimedia traffic environment because it offers a great flexibility and efficiency in the use of available resources.

In this thesis we generate, model and find performance measures of self-similar traffic, which is frequently encountered in the ATM environment. We study the modeling and performance measures of Ethernet and VBR video data. However, the main emphasis in the dissertation is VBR video data. In addition, we propose a model that can be applied to this kind of correlated traffic. The model is based on multiple type ON-OFF sources. We compare the model with those that are available to correlated traffic. Finally, we apply the proposed model to congestion and admission control.

To the memory of my father,
to my mother and to my family

ACKNOWLEDGEMENTS

I would like to express my sincere thanks to my supervisor, Prof. J. F. Hayes for his support and guidance through this work. It was privilege for me to be one of his students and I am honored by his supervision and friendship. My sincere thanks to my family who stand beside me during the difficult times.

Table of Contents

CHAPTER I

Introduction

The increased demand for communications services of all kinds, has engendered the development of Broadband Integrated Services Digital Networks (B-ISDN) which is expected to provide a single and efficient transport for voice, video-conferencing, video-phone, high speed data transfer, home education, video on demand as well as a number of services which are yet to be developed. The integration of these vastly different types of traffic in a common medium with common switching and multiplexing is possible due to the development of asynchronous transfer mode (ATM). At the present time, the traffic volume for existing and new services is increasing. Figure 1.1 shows the growth of different services, voice, data, video, and multimedia [ONV94]. As can be seen, the growth is explosive in the non-voice traffic particularly video. The growth is spurred by the growth of technology. There are two aspects to this growth. First, applications, e.g. computer capability, video technology, and email have burgeoned. Second, hand in hand with the growth of applications has been the growth of telephone technology, in particular digital processing and optical fiber transmission.

This thesis examines some of the important aspects in the modelling and performance analysis of ATM systems. In particular, this thesis is concerned with the generation, modelling and performance analysis of self-similar traffic, which is frequently encountered in the ATM environment. But first, we give a brief survey on how ATM originated and on the main features which made it possible for the

ATM to be the transfer mode of choice for future telecommunications networks. In addition, we present an introduction for the characterization of broadband traffic.

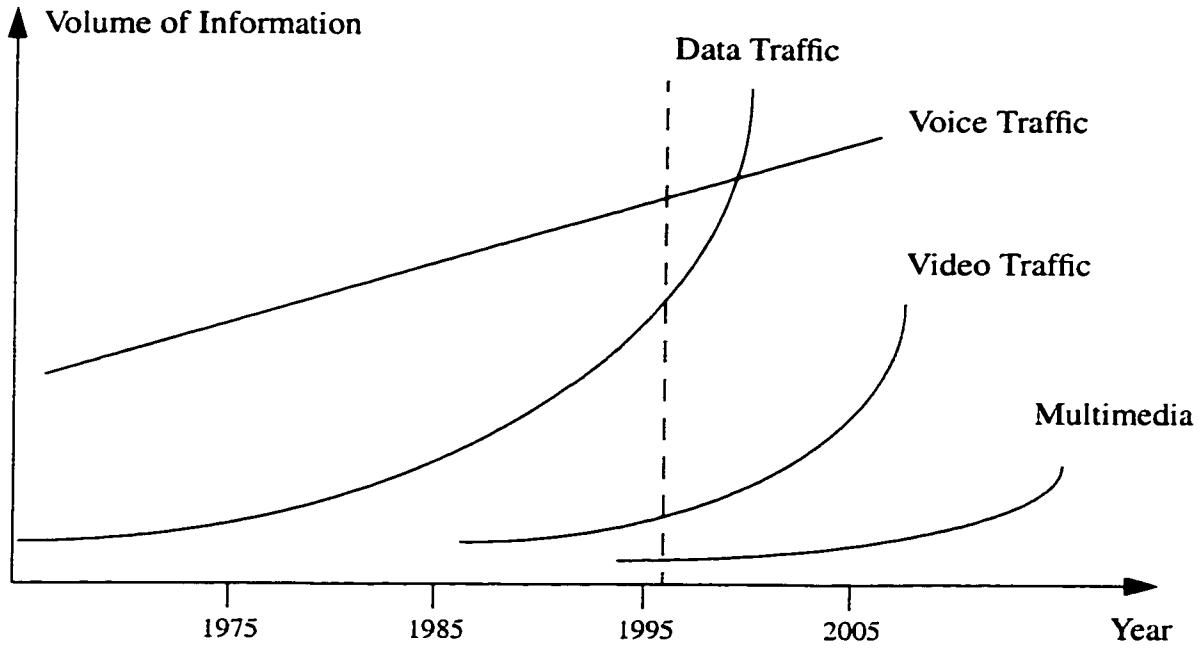

FIGURE.1.1. Growth of service [DEP95].

1.1 The evolution towards ATM

The services supported by the original ISDN concept are limited to voice and non-voice applications based on a 64 Kbit/s transmission rate. This is referred as Narrowband ISDN (N-ISDN). However, the data rates associated with N-ISDN are inadequate for many applications of interest. Accordingly, B-ISDN has been developed for higher rate services [HAN89]. B-ISDN is conceived as an all-purpose digital network. It includes 64 Kbit/s ISDN capabilities and opens the door to applications utilizing bit rates above 1.5 Mbit/s or 2 Mbit/s. The upper limit of the bit rate available to a broadband user will be somewhat above 100 Mbit/s. The services supported in a fully evolved multimedia network such as B-ISDN can be expected to produce a wide range of traffic flow characteristics, and have a wide

range of performance requirements. Figure 1.2 provides some rough ranges of the maximum bit-rate and the utilization of a channel at this rate for some general services categories [WOO90]. As shown there is a large range of services, with estimated bit rate of few Kbits/s to some hundreds of Mbits /s. Also channel utilization ranging from 0.001 (bursty traffic) to 1.0 (continuous traffic).

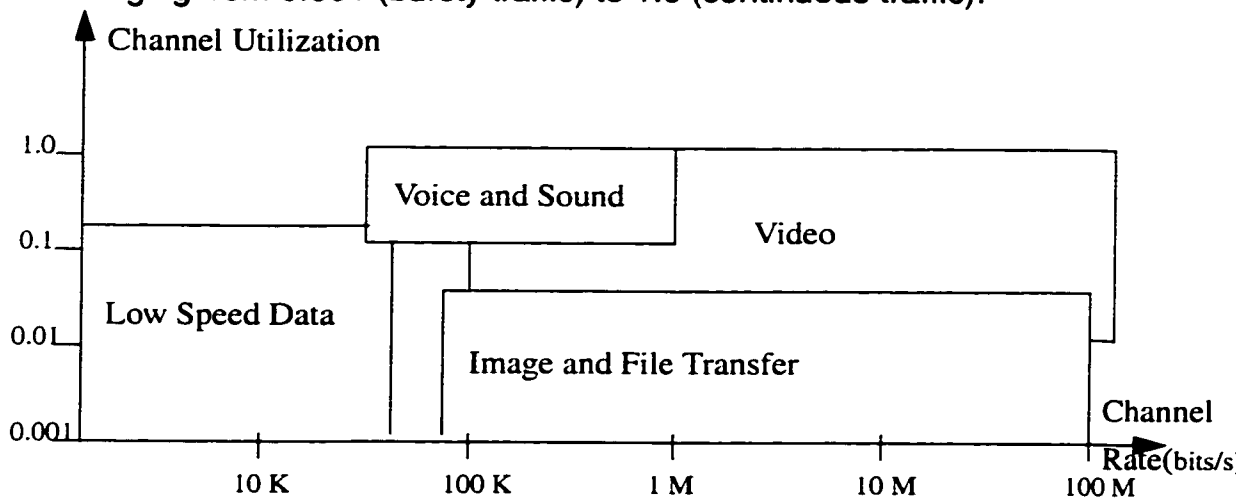

FIGURE.1.2. Multimedia traffic characteristics [WOO90].

Unlike traditional data networks, in a multimedia environment there are requirements on cell delay and cell loss performance, defined as the number of lost cells divided by the total number of cells transmitted. A wide range of applications must be supported: from delay-sensitive applications such as voice, to loss-sensitive applications such as data and image transfer. Figure 1.3 shows the traffic performance requirements for different services [WOO90].

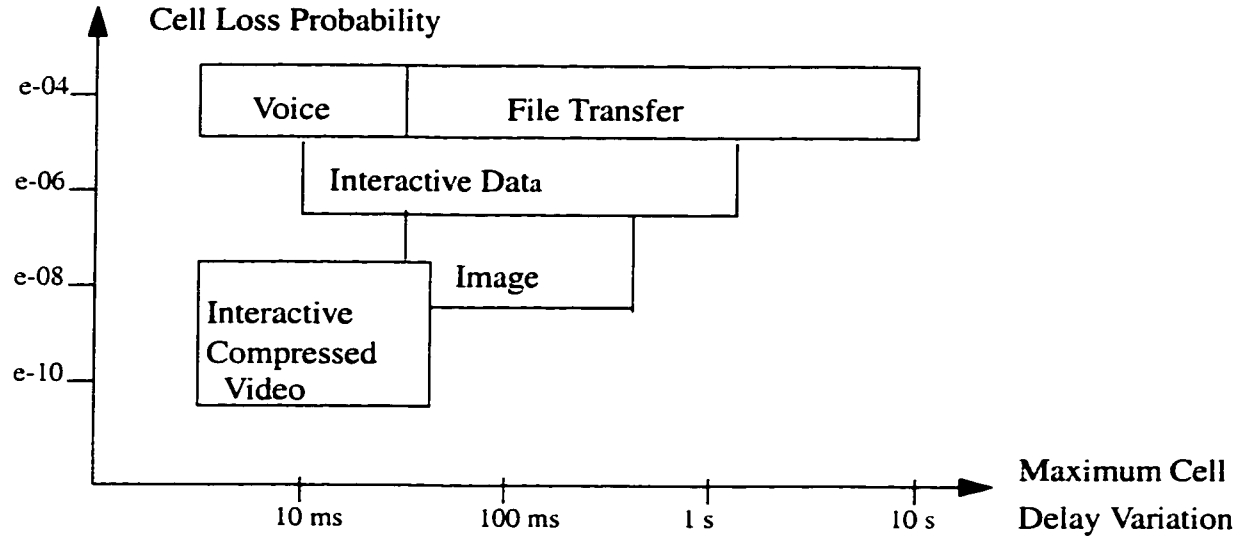

FIGURE.1.3. ATM traffic performance requirements [WOO90]

The data networks of the 2000's are inadequate to handle the applications and capabilities required by the1990's. Today's packet switched data networks have a number of problems: high cost, low-speed links, slow processors and switching delays. Tomorrow's broadband networks require new architectures to handle the changing requirements. Table 1.1 compares the present and the future of some packet network characteristics [DEP95].

TABLE 1. 1 Comparison between present and future of some packet network characteristics [DEP95]

	Present	**Future**
Bandwidth	64 Kbit/s	51.84 (OC*-1)- 9953.28 (OC-192)Mbit/s
Bandwidth allocation	Fixed	Dynamic
Services	Voice, data	Integrated voice, data, image and video.
Switch delay	50 - 100 ms	10 ms
Propagation delay	Insignificant	Dominant

An important building block for broadband technology is the Synchronous Optical Network (SONET). SONET [HOL92, BAE91], originally proposed by Bellcore (Bell Communication Research), defines optical interface, rate, and for-

*OC = Optical Carrier level

mat specification for broadband optical signal transmission. It is compatible with existing circuit-switched networks and can be used to carry ATM based payload, which will be discussed below, as well as those of the existing networks. Therefore, SONET makes the transition from existing networks to ATM networks.

Another building block for broadband technology is the Synchronous Digital Hierarchy (SDH). It should be clear that SONET and SDH are nearly synonymous. The definition of the SDH standards signals the beginning of the next stage in the evolution of the world's telecommunications network. SDH will facilitate a revolution in telecommunications services which will have far reaching effects for end users, operators, and equipment manufacturers alike. SDH has been designed to support future services such as Metropolitan Area Networks (MANs), Broadband ISDN, and personal Communications. As for SONET, SDH defines a structure which enables signals to be combined together and encapsulated within a standard SDH signal.

The SONET and SDH standards were set up for the transmission of time division multiplexing (TDM) digital signals in the 1980s. With TDM, a data stream at a higher bit rate is generated directly by multiplexing lower bit rate channels. High capacity TDM systems operate at levels up to OC-192 (10 Gbit/s), through the use of high speed optical technology.

1.2 ATM and B-ISDN

Traditional telecommunication networks are specialized; there are clear relations between services and networks. The salient example is the voice network which constituted virtually all of the common carrier services for many years.

Many of the telecommunication services have their own networks and those networks are typically not very well suited for supporting other services than those initially intended to be supported.

Our society is becoming more and more information intensive, and both the number of services and the number of users are expected to grow dramatically. The new services will require higher bit rates per user than the existing networks can offer. It would be ineffective to build a new network for every new service. Therefore, the new technology should also be able to support future services; services that we know nothing about when the technology is developed. The new system should also be able to support all the services provided by the existing specialized networks.

The B-ISDN vision is to support all kinds of services in a single network. B-ISDN needs an extremely flexible switching technology. The ATM technology has been developed to be able to fulfil the needs of the B-ISDN. While ATM considered as a transfer mode for transmission very high data rates, B-ISDN is a network specification exploiting ATM technology.

1.3 ATM architecture

ATM [BAE91, DEP95, ONV94] is considered by the International Consultative Committee for Telephone and Telegraph (CCITT) as the preferred transfer mode for B-ISDN. Both the need for flexible networks and the progress in technology and system concepts led to the definition of the ATM principle. ATM will provide the means to transport, at broadband rates, the traffic generated by a wide range of multimedia services. ATM is suitable for the multimedia traffic environ-

ment because it offers a great flexibility and efficiency in the use of available resources. All available resources in the network are shared by all services, so the statistical sharing of the resources can lead to greater efficiency. The fundamental building block for ATM is the fixed-size cell which consists of 48 octets carrying user information plus 5 octets for overhead and control. The reason for choosing a fixed-size cell is to ensure that switching and multiplexing function could be carried out quickly and easily. ATM is a connection-oriented technology (similar to the telephone networks) in the sense that before two systems on the network can communicate, they should inform all intermediate switches about their service requirements and traffic parameters.

In ATM networks, each connection is called a virtual circuit or virtual channel (VC) because it allows the capacity of each link to be shared by connections using that link on a demand basis rather than by fixed allocations. The connections allow the network to guarantee the quality of service (QoS) requirements by limiting the number of VCs. Typically, a user declares key service requirements at the time of connection setup, declares the traffic parameters and may agree to control these parameters dynamically as demanded by the network.

In figure 1.4 a layered model of ATM is shown. The physical-medium layer is responsible for the proper bit transmission. This layer is also responsible for electro-optical conversion since, in B-ISDN, the physical medium may be optical fiber. The ATM layer contains all the details of the ATM technique, and it is common to all services. The data unit of this layer is an ATM cell. This layer performs the cell header functions and cell-based multiplexing/ demultiplexing. The ATM adaptation layer (AAL) provides the higher service layers with the necessary functions which

are not provided by the ATM layer, such as, preserving timing, data frame boundaries. Four types of AALs were proposed, each supporting a different type of traffic or service expected to be used on ATM networks. The service classes and the corresponding types of AALs were as follows: Class A - constant bit rate (CBR) service such as uncompressed video, Class B - variable bit rate (VBR) services such as compressed packetized voice or data, Class C - connection-oriented data service such as connection oriented file CBR transfer, and Class D - connectionless data service such as datagram traffic and in general, data network applications. The higher layers are network layer, transport layer, session layer, presentation layer, and application layer. The higher service layers provides separate functions for the User Plane and Control Plane. It supports services such as network terminal signalling and information source coding. The Control Plane is responsible for the signalling, whereas the User Plane is responsible for the transfer of user information.

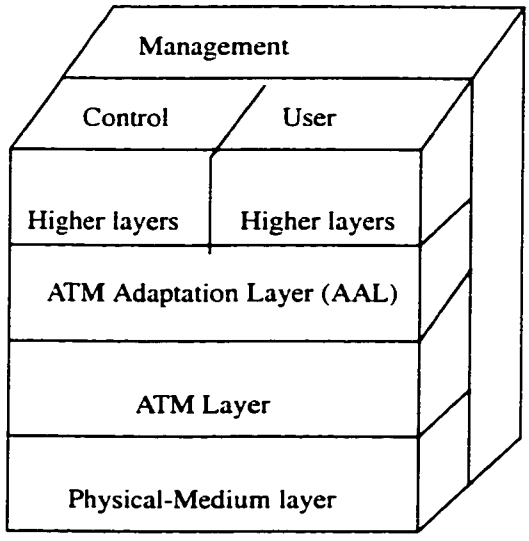

FIGURE.1.4. Layered model of ATM [ONV94]

Figure 1.5 is the cell format in ATM networks. The format of the header for ATM cells has two different forms, one for use at the user-to-network interface (UNI) and the other for use internal to the network, the network-to-node interface (NNI). The header of each cell contains, 4-bit generic flow control (GFC), a 16 bit virtual circuit identifier (VCI) and a 8 bit virtual path identifiers (VPI). The remaining fields are, a 2 bit payload type(PT) field, a 1 bit reserved field, a 1 bit priority (PR) field, and an 8 bit header error check (HEC).

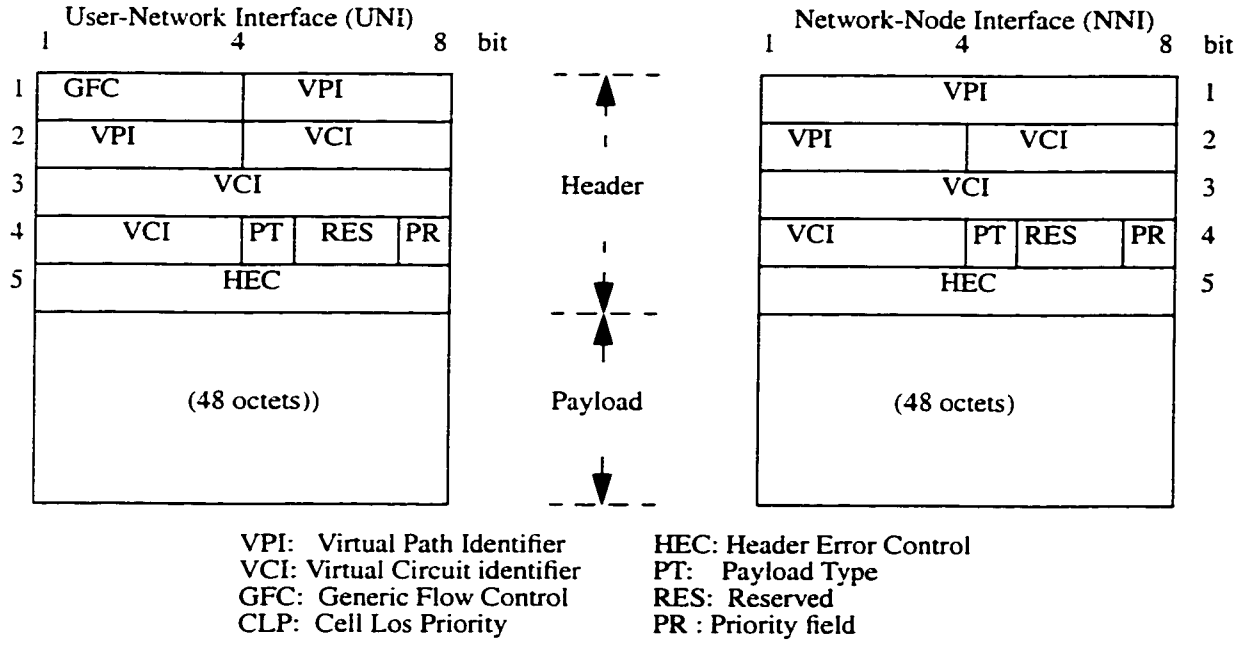

FIGURE.1.5. ATM cell format [ONV94]

The function of the GFC is to control the amount of traffic entering the network. This allows the UNI to limit the amount of data entering the network during periods of congestion. The VCI and the VPI together form the routing field, which associates each cell with a particular channel or circuit. The VCI is a single-channel identifier; the VPI allows grouping of VCs with different VCIs and allows the group to be switched together as an entity. The payload contains all the user data and ATM adaptation layer information.

1.4 How ATM works

1. ATM network uses fixed-length cells to transmit information. The cell consists of 48 bytes of payload and 5 bytes of header. The flexibility needed to support variable transmission rates is provided by transmitting the necessary number of cells per unit time.

2. ATM network is connection-oriented. It sets up a virtual channel connection (VCC) going through one or more virtual paths (VP) and virtual channels (VC) before transmitting information. The cells is switched according to the VP or VC identifier (VPI/VCI) value in the cell head, which is originally set at the connection setup and is translated into new VPI/VCI value while the cell passes each switch.

3. ATM resources such as bandwidth and buffers are shared among users, they are allocated to the user only when they have something to transmit. So the network uses statistical multiplexing to improve the effective throughput.

After this introductory background material on ATM networks and their traffic characterization, we are ready now to focus more on some of the main ATM performance analysis issues which will be dealt with in this dissertation.

1.5 Congestion control in ATM networks

The concept of congestion in the network layer is a very simple one. The performance of any system will degrade if the amount of work that the system is forced to do is more than it can cope with. In this context, if there are too many packets present in a given part of the subnet, we say that the subnet is congested. This situation is shown graphically in figure 1.6 [HAY84]. It is clear from the diagram that performance degrades very sharply when congestion occurs.

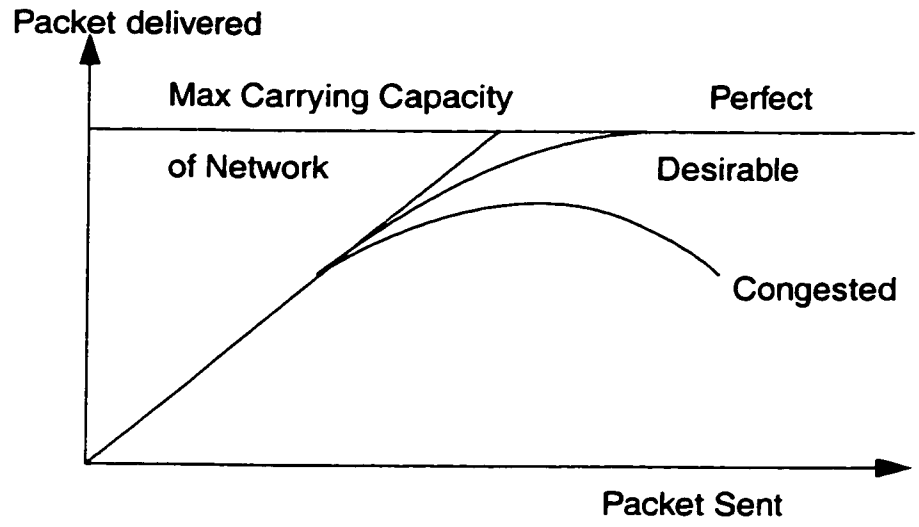

FIGURE.1.6. Illustration of the effect of flow control [HAY84]

Congestion control is a particular challenge in the ATM environment [BAI91] because of its unique traffic characteristics, high link speed, diverse service requirement and the diverse characteristics of the traffic ATM is expected to support. Most traffic sources in ATM networks are bursty. A bursty source generates a large amount of traffic (cells) at some high peak rate for a short period of "active" time and generates little or no traffic for some "idle" time. If bandwidth were allocated based on peak rate (deterministic multiplexing), network resources would be wasted when the source is ideal. Since an ATM network supports a large number of such bursty traffic sources, statistical multiplexing is more efficient, allowing more traffic sources to share the bandwidth. However, severe network congestion may occur if a large number of traffic sources become active at the same time.

A layered and distributed congestion-control framework has been proposed to apply to the design of congestion control in ATM networks, which is portioned into three control domains, the call layer, the burst layer and the cell layer. In the

call layer, connection admission control (CAC) should be used, by examining the load, service requirements and traffic characteristics. Since the time scale is larger than the propagation delay, link by link closed loop negotiation can be implemented. In the cell layer, bandwidth enforcement preventive control must be used, since the time scale is small. In section 1.5.2 and 1.5.3 we will discuss the two main levels of congestion, call level and cell level, in ATM networks. In section 1.5.4, we present a mechanism used to reduce the level of traffic burstiness. In [HON91, HUA95, PER97] the following congestion control schemes at the various levels have been advocated.

1.5.1 Preventive control, reactive (conventional) control and proactive control.

There are three types of congestion control schemes, which have been developed for ATM networks, namely preventive congestion control, reactive congestion control and proactive congestion control. As its name indicates preventive congestion control takes any action necessary to prevent congestion (i.e., before congestion occurs). Reactive congestion control is responsible for any necessary action to recover from a congested situation. Proactive congestion control on the other hand is based on the prediction of congestion (overload) before it occurs at a node. A forecast on the future of the overload is generated for one round trip propagation delay ahead and feedback signals are send back to the source. Traffic adjustment at the network input take place in response to these feed back signals [AME91, PER97].

The congestion control schemes used for existing networks react to the congestion after it happens and tries to bring the degree of network congestion to an

acceptable level. In many of the conventional packet switching networks, for example, the applied congestion control schemes are rate-based or credit-based mechanisms, which will be discussed in section 1.6, are under the category of reactive control. A major problem with reactive control in high-speed networks is the propagation delay. Because of this delay, there can be a significant amount of traffic in transient in the links. Since the speed of ATM is very high, any action the sources taken may be too late to resolve the buffering and switching congestion. Hence, most reactive congestion control schemes are effective over short distances only. Therefore, some of the congestion control schemes available for existing networks are no longer applicable.

Congestion control in ATM networks is based on preventing congestion rather than reacting to it. Preventive congestion control does not wait until congestion actually occurs, but rather tries to prevent the network from reaching an unacceptable level of congestion. Most often, preventive congestion control is implemented at the access nodes of the ATM network. There are also two ways to implement preventive control, namely admission control and bandwidth enforcement which is part of traffic policing function that will be discussed shortly.

Proactive congestion control scheme is an alternative to reactive control algorithms because of the high transmission speed and the high bandwidth-delay product of multimedia networks. Proactive congestion control is based on the prediction of overload, given a current underload network state and the subsequent transmission of feedback signals to the network input in the case of anticipated congestion; traffic adjustment at the network input take place in response to these feed back signals [PER97, HU95, AME91].

1.5.2 Admission control (call-level congestion control)

Congestion can occur when the network accepts too many calls, which causes the QoS to deteriorate. However, reducing the number of calls is not always the solution, since there is also a need to maintain the network utilization [SAI91]. Call-level is the first of two main levels of congestion prevention in ATM networks. At this level, it is required to avoid long-term congestion and maintain the traffic load at a manageable level.

Admission control determines whether to accept or reject a new connection at the time of call setup. When a new call requests connection, the network should first decide whether it can admit the call or not. The factors that affect the admission of a call include the availability of network resources (network bandwidth), the traffic characteristics of the new call and of existing calls sharing the same resources, and the QoS requirements (such as the probability of loss and delay) of the new and existing calls. If the network does not have enough bandwidth available along the path connecting this call to support the QoS requirements of the new call and the existing calls, the service request should be rejected.

1.5.3 Traffic policing (cell-level congestion control)

Cell-level is the second main level of congestion prevention in ATM networks. At this level, the objective is to avoid short-term congestion. After a connection is set up, some flow control is still required to provide good performance and guarantee fairness among the users and eliminate the possibility of congestion. This kind of congestion is based on the declared parameters; consequently a policing procedure is needed to ensure that any change in the user's traffic characteristics will not affect the overall performance of the network. The main empha-

sis in policing function is bandwidth enforcement. Bandwidth enforcement is implemented at the edge of the network. Once a violation is detected, the traffic flow is enforced by discarding or buffering violating cells. A good survey on these policing mechanisms is available in [GAL90]

A well-known example of policing mechanism is the so-called leaky bucket [HON91]. Figure 1.7 shows the leaky bucket policing scheme. The leaky bucket scheme uses tokens to enforce the authorized source traffic pattern. Token are generated at a fixed rate determined by the bandwidth granted to the call. An ATM cell must capture a token before it is transmitted; otherwise it must wait in the queue until a token is generated. A cell is discarded if it encounters a full buffer on arrival. If the number of tokens in the token pool exceeds some predefined threshold value, the process of token generation stops. This threshold value corresponds to the burstiness of the transmission; the larger the threshold value, the higher the burstiness. This method enforces the average input rate while allowing for a certain degree of burstiness.

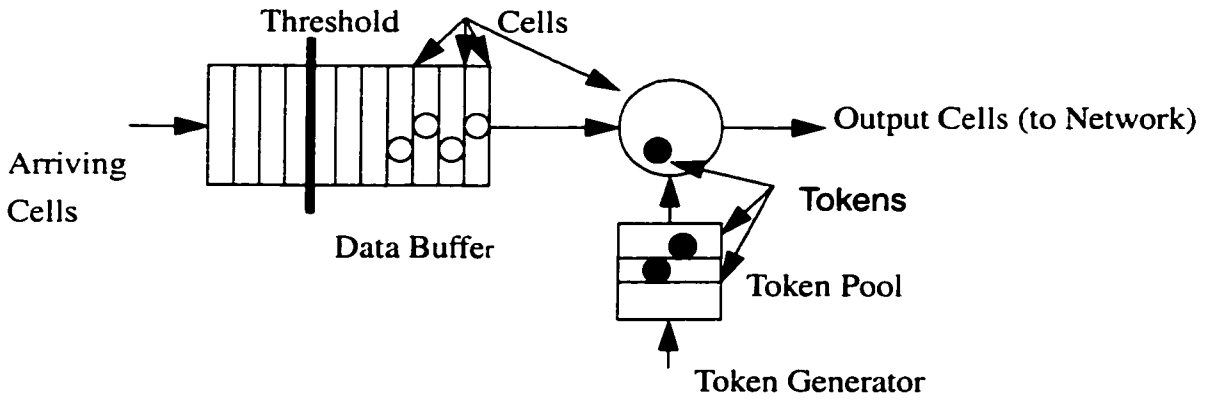

FIGURE.1.7. Leaky bucket combined with cell buffering for both policing and smoothing [HON91]

The leaky Bucket method can also enforce the peak bandwidth by generating tokens at the rate corresponding to the peak rate. One disadvantage of this

policing mechanism is that, it is not adaptive to the network load, where violating cells are discarded even when the network load is light, and thus network resources are wasted.

In order to adapt to network congestion, a virtual leaky bucket is proposed [GAL89], where instead of discarding excessive cells, they are simply marked with tags in their cell header. Using this method, if the network is not overloaded, none of the cells will be dropped and the network throughput can be improved. In the case of congestion, however, the marked cells will be dropped. One possible disadvantage of this marking scheme is that processing time in each node is increased slightly because each node has to distinguish tagged cells from nonviolating cells when the node is in congested state. Also, network resources are used on cells that are eventually dropped.

1.5.4 Traffic smoothing

While traffic policing emphasizes bandwidth enforcement, traffic smoothing emphasizes traffic shaping to reduce burstiness and improve network throughput [HON91]. Traffic smoothing as traffic policing is also performed at the cell level. However, unlike policing functions which work on the network side, this smoothing function works on the user side. The basic premise of traffic smoothing is the use of buffering to achieve a certain transmission bits per smoothing interval. The smoothing function allows the users to control their traffic parameters, typically the cell's minimum interarrival time (maximum bit rate) and the maximum source activity (fraction of time during which the source transmits) allowed in a given time period. Policing can be combined with smoothing in a system in which cells queue

instead of being discarded when the token pool is empty (see figure 1.7). It is noted that smoothing the traffic entering the network is achieved at the cost of increased delay and occasional cell losses to maintain overall performance throughout the network [CID88].

1.6 Rate based control and credit based control

In most data networks, such as the typical Ethernet LAN or X.25 WAN, there is no explicit contract between the network and the user specifying the traffic profile and QoS expected. Rather, the network is expected to provide each user with a fair share of the available bandwidth. However, in an ATM network, fair allocation of bandwidth requires users to adjust their transmission rates according to the feedback from the network. Unlike other packet networks, ATM networks also carry fixed bandwidth services required for multimedia applications constant bit rate (CBR) traffic and guaranteed bandwidth services for high-priority data applications-variable bit rate (VBR) traffic. The remaining bandwidth, not used by guaranteed bandwidth services, must be shared fairly across all users. Typically, the CBR and VBR classes are assigned higher priority by the network switches and get a share of the link bandwidth first. The ATM Forum refers to services that make use of the left over bandwidth as available bit rate (ABR) services. The ATM traffic tracking is shown in figure 1.8. Table 1.2 shows network traffic types and their requirements.

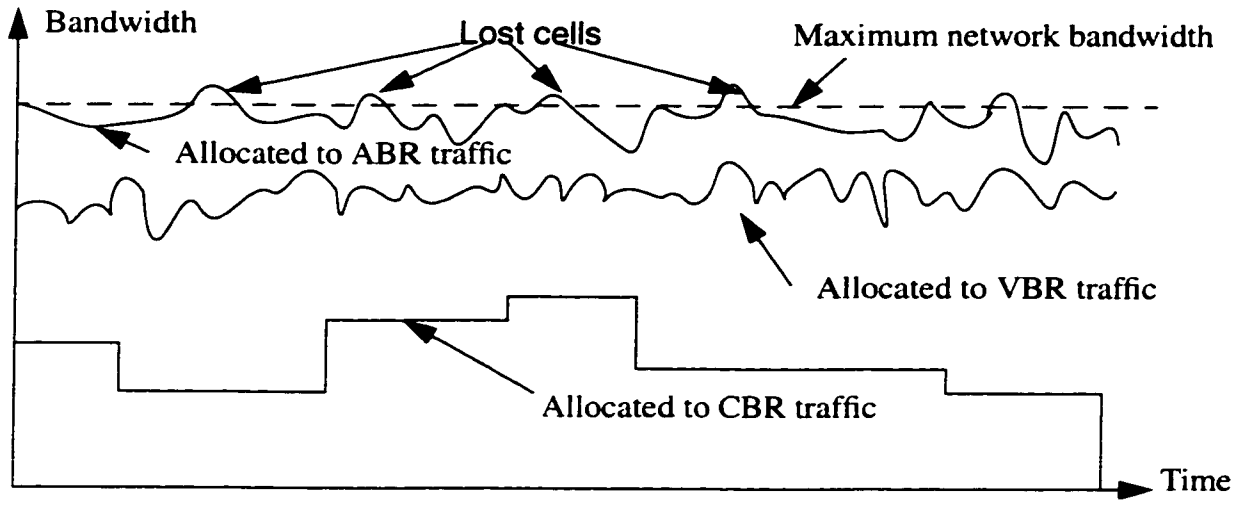

FIGURE.1.8. Constant bit-rate (CBR) traffic is guaranteed a fixed amount of bandwidth. Bandwidth is also guaranteed for variable bit-rate (VBR) traffic, although in this case the amount varies. Whatever bandwidth remains is dedicated to available bit-rate (ABR) traffic with no guarantees.

TABLE 1. 2 Network Traffic Types and their requirements

Traffic Type	Example	Bandwidth required
Constant	Voice	Guaranteed
Variable	Compressed Video	Guaranteed
Available	Data	Not Guaranteed

The CBR service is aimed at supporting voice and other asynchronous applications. The VBR service is designed to support video and audio applications while ABR service is designed to primarily support data applications.

ABR is to support best effort applications by dynamically sharing the available network resources (left over by other sources) among all ABR users. It relies on the feed back control mechanism to throttle the source rate according to the current load of the network. As we have discussed for reactive control in section 1.5.1, the distance between the source and the destination should not be large in order to avoid having a significant amount of traffic in transit. To implement the

flow-control for ABR services, two approaches has been developed: One is "rate-based" scheme, the other a "credit-based" approach.

In the rate-based approach, the ATM network sends information to the user specifying the bit rate at which the user should be transmitting. When the network becomes congested, the end-stations sending ABR traffic are told to slow down. In the credit-based approach, switches and end-stations exchange information about the available buffer space on each link of the network. End-stations sending ABR traffic would send only when sufficient buffer space was available.

Rate-based flow-control schemes are end-to-end feedback mechanisms. That is, they have one source and one destination station for each feedback loop. Within the feedback loop, the destination end alerts the source end to slow transmission when congestion begins to occur. If there are ATM switches between the loop's source and destination, these devices simply forward and augment the flow-control information moving between the destination and the source.

Credit-based flow-control schemes also make use of a feedback loop, but they use hop-by-hop loops rather than end-to-end loops. Each link maintains its own independent control loop; when traffic moves across a network, it moves through a series of hop-by-hop feedback loops. The receiving end of each link issues "credits" to the transmitting end indicating the number of cells the transmitting station is allowed to send. Source end-stations transmit only when they have permission to do so from the network.

Under the credit-based approach, each link in the network runs the flow-control mechanism independently for each virtual circuit. A certain number of cell buffers are reserved for each virtual circuit at the receiving end of each link. One

round-trip's worth of cell buffers must be reserved for each connection, so the amount of buffering required per connection depends on the propagation delay of the link and the required transmission rate of the virtual connection.

The main differences between rate-based and credit-based approaches are shown in the Table 1.3.

TABLE 1. 3 Comparison of rate-based and credit-based schemes

Rate based control	Credit based control
Does not need per VC queueing	Require per VC queueing
Does not guarantee zero cell loss	Guarantees zero cell loss
Can not easily handle loss of RM cells	Loss of credit control cells does not create a problem, the previous value of credit is used

1.7 Characterization of broadband traffic

1.7.1 Burstiness

Variability in the rate of arrival processes has been a characteristic of data communication between or among computers from the earliest days of networking. The primary cause of variability was human: for example, users paused for different lengths of time to think before typing intervals. The new range of traffic types have a wide range of characteristics as well as a range of performance requirements (see figures 1.2 and 1.3). The introduction of networking technology has directly led to an increase in the variability of arrival processes. The appearance of real computer networks, file transfer protocols, and computer mail resulted in more variability, as users could now instruct a system to send a message to a remote location, or copy data from one location to another. Measurement studies showed that local-area traffic is bursty [LEL91],variable bit rate (VBR) video traffic is also bursty [BER95]. The advent of high-speed local-area

networks, and, the development of the computer workstations, opened new doors for data communication; remote file systems, more sophisticated protocols, e-mail, world wide web, and others. All of these advances contribute to increasing variability since faster protocols allowed for transmission of more data in a shorter time.

While the networking research community began to experience congestion and to think about ways to cope with it, telecommunication researchers, who had long been studying telephone voice channels, had already understood the effects of bursty arrival processes on queues. Briefly, the packet arrival process is highly correlated and that Poisson approximation for the arrival process gives in erroneous results since it fails to account for the burstiness. That is to say, the Poisson model does not fit a wide range of traffic types because of burstiness and is no longer useful to model such a new class of traffic. Beyond this, traffic may be short-range dependent (SRD) or long-range dependent (LRD).

1.7.2 Short and long-range dependence

Stochastic traffic models of packet networks currently considered in the literature are almost exclusively Markovian in nature, or more generally, result in short-range dependent (SRD) traffic processes for which the correlation falls exponentially. Long-range dependence (LRD) is represented by a single parameter H, after H.E. Hurst [HUR51] who studied the long term storage in water reservoirs. The simplest models with LRD are self-similar processes. Self-similar processes (or fractal processes) means that any portion of the curve, if blown up in scale, would appear identical to the whole curve. The Hurst parameter H

implies a certain relationship of autocorrelations over all time scales. Thus, the LRD process is a second order self-similar process. A Wide Sense Stationary (WSS) process* X is said to be exactly second-order self-similar if the corresponding aggregated process $X^{(n)}$ (obtained by averaging the process X over successive non-overlapping blocks of size n) have the same autocorrelation function as X, for all $n \geq 1$ [LEL93, BER93]. Moreover, a process is called asymptotically second-order self-similar if the covariance function of the aggregated process $X^{(n)}$ for large k and n is given by $\rho_k^{(n)} \rightarrow \rho_k$, i.e, the covariance function of the aggregated process does not depend on the block size n, but depend on the lag k.

The presence of LRD in a time series indicates that while long-term correlations (large lags) are individually small, their cumulative effect is non-negligible and produces scenarios which are drastically different from those experienced with traditional SRD models such as Markovian processes. While the commonly made assumptions require that observations separated by a large time span are roughly independent, in practice, it is not the case and long time series measurements violate this independence assumption and exhibit long-range dependence instead (for example, variable-bit-rate video, and Ethernet local area network traffic).

The idea of SRD and LRD can be made explicit by examining the correlation of the processes. Let $X = (X_t : t=1,2,3,.....)$ be a WSS process with mean $\mu = E(X_t)$, variance $\sigma^2 = E\{[X_t - \mu]^2\}$ and autocorrelation coefficient at lag k, $\rho_k = Cov(X_t, X_{t+k})/\sigma^2 = E\{(X_t - \mu)(X_{t+k} - \mu)\}/\sigma^2 = (\gamma_k/\sigma^2)$, $k = 0, 1, ..$ that depends only on k and not on t.

The literature does contain studies which show stationerity, e.g. [GUS91]

We can think of a packet traffic process X consisting of a set $\{X_t\}$, where X_t is the number of packets that arrive in the *t-th* time unit. The variance of the sum of n identically distributed random variables X_1, X_2, ..., X_{n-1}, X_n is given by:

$$Var(X_1 + ...+X_n) = n\sigma^2 + \sum_{\substack{j=1 \\ j \neq i}}^{n} \sum_{i=1}^{n} Cov(X_j, X_i) \qquad (1.1).$$

But,

$$\sum_{\substack{j=1 \\ j \neq i}}^{n} \sum_{i=1}^{n} Cov(X_j, X_i) = 2(n-1) \cdot \gamma_1 + 2(n-2) \cdot \gamma_2 + ...+2 \cdot \gamma_{n-1} \qquad (1.2).$$

Using the relation $\rho_k = \gamma_k/\sigma^2$, (1.2) can be written in the following form,

$$\sum_{\substack{j=1 \\ j \neq i}}^{n} \sum_{i=1}^{n} Cov(X_j, X_i) = 2 \cdot \sigma^2 \sum_{k=1}^{n-1} (n-k)\rho_k \qquad (1.3).$$

Substitute (1.3) in (1.1),

$$Var(X_1 + ...+X_n) = n\sigma^2 + 2 \cdot \sigma^2 \sum_{k=1}^{n-1} (n-k)\rho_k \qquad (1.4).$$

Let $S_n = X_1 + X_2 + ... + X_n = \sum_{i=1}^{n} X_i$, and $v_n = Var(S_n/n)$

As indicated above we may classify processes into two categories short- and long-range dependence [COX84]:

1) short-range dependent processes such as, Markov chains and auto-regression moving average of finite order. These processes satisfy the following conditions:

i) $\rho_k \sim r^k$, $\sum_{k=1}^{\infty} \rho_k < \infty$, $0 < r < 1$.

ii) for $n \to \infty$ $v_n \sim n^{-1}$.

iii) $f(\lambda) = \sum_{k=1}^{\infty} \rho_k e^{ik\lambda} < \infty$, $f(0) = \sum_{k=1}^{\infty} \rho_k$=constant (positive and finite).

Condition *i*, indicates that the autocorrelation coefficient ρ_k is summable.

Condition *ii*, indicates that the variance v_n decays like the reciprocal of the sample mean.

Condition *iii*, indicates that the spectral density $f(\lambda)$ converges at the origin.

2) long-range dependent processes such as, self-similar traffic. These processes satisfy the following conditions:

iv) $\rho_k \sim k^{2H-2}$, $\sum_{k=1}^{\infty} \rho_k = \infty$, $0.5 < H < 1$.

v) for $n \to \infty$ $v_n \sim n^{2H-2}$.

vi) spectral density $f(\lambda) \sim \lambda^{1-2H}$, as $\lambda \to 0$.

Condition *iv*, indicates that the autocorrelation coefficient ρ_k is non-summable.

Condition *v*, indicates that the variance v_n decays more slowly than the recporical of the sample size.

Condition *vi*, indicates that the spectral density $f(\lambda)$ diverges at the origin $(1/\lambda^{2H-1})$.

For SRD, the Hurst parameter $H = 0.5$. For LRD traffic $0.5 < H < 1$.

1.7.3 Measurements of burstiness and dependencies

Knowledge of information source characteristics is important in ATM networks because of traffic control. Therefore, some measures are necessary for characterizing the burstiness [GUS91].

1.7.3.1 Burstiness measurements

There are three commonly used definitions to measure burstiness, the peak to mean ratio (PMR), the coefficient of variations and the index of dispersion for counts. The PMR for the observed traffic depends critically on the time interval over which the bandwidth (or bit rate) is determined. As the interval over which traffic is observed is decreased, the PMR increases. Such behavior differs extremely from the "burstiness" of simple arrival models such as that for Poisson arrival process.

Other measure of bursty traffic is the coefficient of variation (COV). It is defined as the ratio of standard deviation to the mean. This measure gives more information than the PMR about the trends in the traffic, since it represents the deviation from the mean. The COV for arrivals for the Ethernet traffic decreases as the interval length over which arrivals occurs increases. However, the COV of Poisson (also batch Poisson, hyperexponential, or Markov modulated Poisson processes (MMPP)) varies very little as the interval length over which arrivals occurs increases.

The index of dispersion for counts (IDC) is used as a measure for capturing the variability of traffic over different time scales. For a given time interval of length k, the IDC, which will be derived in section 1.7.3.3, is defined as the variance of the number of arrivals during this interval divided by the mean value of the same quantity. The IDC for the self-similar traffic increases monotonically as the interval length over which arrival occurs increases. Conventional traffic models, such as MMPP, hyperexponential and batch Poisson process distributions, have IDC that

converge to a constant over a time scale on the order of the time constant of the model. The large range of monotonically increasing dispersion observed for actual traffic indices indicates that even fairly sophisticated MMPP models may not characterize the actual traffic behavior well over a large range of time scales.

The LAN traffic measurements show a high level of variability on every time scale. For this kind of traffic, both the PMR and the COV are unsatisfactory measures. Although indexes of dispersion can capture more information than other measures, the complexity of the analysis and the arbitrary selection of vital parameters, such as the sequence length, makes it too complex for the real-time traffic characterization of systems. The Hurst parameter H (slope of $IDC = 2H - 1$) provides a more satisfactory measure of "burstiness" for self-similar traffic than the above commonly used measures. It implies a certain correlation of the process over all time scales. In the following two sections, we discuss the indices of dispersion for interval and for count.

1.7.3.2 Index of dispersion for intervals (IDI)

Let us define for packet-arrival processes, the length of intervals between successive arrivals as the length of time between the beginning of the transmission of a given packet and the beginning of the transmission of the previous packet and denote it by X_i. Under this definition, the transmission time of the previous packet is included in the interarrival time. The variance of the sum of n identically distributed random variables X_i is given by (1.4), and it is again given by,

$$Var(X_1 + ...+X_n) = n\sigma^2 + 2 \cdot \sigma^2 \sum_{k=1}^{n-1} (n-k)\, \rho_k \qquad (1.5).$$

where $\sigma^2 = Var(X)$, and $\rho_k = Cov(X_i, X_{i+k})/Var(X)$ is the autocorrelation coefficient at lag k.

The variance of the sum of intervals is useful in describing the arrival process because of the dependency on the autocorrelation coefficient as shown below. The IDI at lag n is the variance of the sum of n successive interarrival times normalized by the factor $nE^2(X)$.

$$J_n = \frac{Var(X_1 + ...+X_n)}{nE^2(X)}, \qquad n = 1, 2, ... \tag{1.6}$$

Where $E(X)$ is the common mean.

The IDI can be expressed in terms of the squared coefficient of variation (SCOV) of intervals $J_1 = Var(X)/E^2(X)$ and the autocorrelation coefficient at lag n, i.e. $\rho_n = Cov(X_i, X_{i+n})/Var(X)$, as

$$J_n = J_1\left[1 + 2\sum_{j=1}^{n-1}\left(1 - \frac{j}{n}\right)\rho_j\right] \tag{1.7}$$

The IDI given by (1.7) is used as a measure of the variability of packet arrival processes. Packet-arrival processes normally have positive autocovariance (or positive autcorrelations coefficients) since interarrival shorter than the mean interarrival time and those longer than the mean interarrival time tend to occur in separate bursts. Therefore, in packet arrival processes, the IDI of a sequence increases with increasing n. Further, the limit of equation (1.7) is:

$$\lim_{n\to\infty} J_n = J_1\left[1 + 2\sum_{j=1}^{\infty}\rho_j\right] \tag{1.8}$$

equation (1.8) indicates that, the value of the IDI when $n \to \infty$ is proportional to the sum of the autocorrelation coefficients.

1.7.3.3 Index of dispersion for counts (IDC)

The index of dispersion for counts (IDC) at time t is the variance of the number of arrivals in an interval of length t divided by the mean number of arrivals in t.

$$I_t = \frac{var(number\ of\ arrivals\ in\ an\ interval\ of\ length\ t)}{E(number\ of\ arrivals\ in\ an\ interval\ of\ length\ t)} = \frac{var(N_t)}{E(N_t)} \quad (1.9).$$

where N_t indicates the number of arrivals in an interval of length t.

In estimating the IDC of the real data, we will only consider the time at discrete and equally spaced instants. Let c_i denote the number of arrivals arrived within an interval of length $\tau_i - \tau_{i-1}$ (τ_i's are equally spaced instants), we define the IDC as:

$$I_t = \frac{var\left(\sum_{i=1}^{n} c_i\right)}{E\left(\sum_{i=1}^{n} c_i\right)} = \frac{var(c_\tau)}{E(c_\tau)}\left[1 + 2\sum_{j=1}^{n-1}\left(1 - \frac{j}{n}\right)\xi_j\right] \quad (1.10).$$

Where $var(c_\tau)$ and $E(c_\tau)$ are the common variance and mean of the c_i's and ξ_j is the autocorrelation coefficient of the c_i's at lag j. The limit of equation (1.10) is

$$\lim_{t\to\infty} I_t = \frac{var(c_\tau)}{E(c_\tau)}\left[1 + 2\sum_{j=1}^{\infty}\xi_j\right] \quad (1.11).$$

As for the IDI case of equation (1.8), equation (1.11) indicates that the IDC is proportional to the sum of autocorrelation coefficients.

IDC for Poisson process has a constant value equal to one. The IDC given by equation (1.10) suggests that packet count process can never regarded as a Poisson process. For some other processes, such as the batch Poisson process, it is also constant but has a value greater than one. For correlated processes such

as MMPP, the IDC is an increasing function until it reaches a constant value after some lag n. For self-similar processes, the IDC is monotonically increasing function as the lag n increases. This monotonic increase in the self-similar IDC can be observed from the summation term of equation (1.10), where the autocorrelation coefficients are not summable. However, for Markov processes such as MMPP, the autocorrelation coefficients are summable.

Although indexes of dispersion can capture more information than other measures, the complexity of the analysis and the arbitrary selection of vital parameters, such as the sequence length, makes it too complex for the real-time traffic characterization of systems. The Hurst parameter, H (slope of $IDC = 2H - 1$) provides a more satisfactory measure of "burstiness" for self-similar traffic than the above commonly used measures. It implies a certain correlation of the process over all time scales.

1.8 Outline and organization of the dissertation

The intent of this dissertation is to generate, model and find performance measures of self-similar traffic. We study the modeling and performance measures of Ethernet and VBR video data. However, the main emphases in our dissertation is the VBR video data. In addition, as we will show, we propose a model that can be applied to this kind of correlated traffic. The model is based on multiple type ON-OFF sources. We compare this model with the models that are available to correlated traffic. Finally, we apply congestion and admission control to the proposed model.

In the first part of this thesis, we consider traffic measurements of Ethernet and video data that are available to us. We calculate some statistical characteris-

tics for both kinds of data. We also investigate some methods of generating self-similar traffic. From these, autocorrelation (or covariance since the two are related), index of dispersion for counts, probability of loss and mean queue length are calculated. In addition, we propose a method for matching the index of dispersion for counts of some models with that of the real data. Given the estimated traffic parameters, we generate synthetic traffic using OPNET and Matlab software and compare the autocorrelation (or covariance), index of dispersion for counts, probability of loss and mean queue length with that of the real traffic.

In the second part of this thesis, we use PMPP, MMPP and Markov chain models to model and investigate their accuracy to the real Ethernet and video data. We match some statistical indices, such as, covariance, IDC to the real data and use the generated traffic based on the matching to find performance measures such as proability of loss and mean queue length. The results obtained in the first two parts of this dissertation can be used to answer some significant questions which arise in the design and performance analysis of ATM systems, such as:

-How is bursty traffic characterized?.

-Which of the techniques for generating self-similar traffic in a simulation give the most accurate results over the wide range of traffic types?.

-Which analytical models are effective?.

-What statistical models characterize the data accurately?

-How well do synthetic traffic models preform as predictors of cell loss rates or delays?.

In the third part of this thesis, we propose a model for characterizing correlated cell arrival of real self-similar video data. Based on a second order statistical

analysis, we have used heterogeneous ON-OFF source model to characterize the traffic. The model consists of m class ON-OFF sources. We find traffic indices such as, covariance and IDC, and calculate performance measures such as probability of loss and mean queue length of the data and the model. We also compare the results of the traffic characteristic indices and performance measures of the data and the model with that of the Maglaris model [MAG88]. We apply our model to congestion control. The probability distributions for the time to overload in a round trip delay for different VBR video data using a heterogeneous ON-OFF source model are found. Admission control is also considered. The number of admitted sources is adjusted in order to have a certain performance criterion.

We have organized this thesis as follows:

In the next chapter we present traffic measurements of Ethernet and video data. We calculate some statistical characteristics of the Ethernet and video data that will be used in the subsequent chapters, to handle the correlation in the arrival process. In chapter 3, we present an overview of some analytical and simulation models that can be used in generating self-similar traffic. We also, generate self-similar traffic based on some of these simulation models. We calculate IDC, autocorrelation (or covariance), probability of loss and mean queue length and discuss the effect of self-similarity on them. In chapter 4 and chapter 5, we investigate the effectiveness of modeling self-similar traffic using conventional models such as FGN, FBM, F-ARIMA, PMPP, MMPP and Markov chains. In chapter 6, we present our proposed model and apply the model to the real data. We calculate the parameters that characterize the model from matching the data and find the most important traffic indices and performance measures. We compare the mod-

els investigated in chapter 4, and 5 with our model in terms of covariance, IDC, probability of loss and mean queue length. In chapter 7, we apply congestion and admission control to the proposed model. Finally, in chapter 8, we give a conclusion, followed by a summary of the main contributions of the thesis and some suggestions for future research.

CHAPTER II

Traffic Measurements

2.1 Introduction

Before modern broadband networks can be a reality, various issues need to be resolved. In order to design and develop network functions, it is necessary to comprehend the characteristics and requirements of the traffic to be carried. In order to study the traffic characteristics of broadband network effectively, it is necessary to obtain measurements which facilitate the identification of traffic descriptors. Leland, et. al. [LEL91] performed traffic measurements and identified important characteristics associated with LANs.

Traffic measurements taken from real data are highly desirable for validating traffic models, designing congestion control algorithms, and developing switch architectures. In addition network QoS control algorithms such as call admission mechanisms, routing algorithms, and bandwidth allocation policies depend heavily on the characteristics of ATM traffic. In this chapter we consider traces for the real Ethernet data and VBR video. There have been several traffic surveys which exhibit LRD, such as Ethernet LAN's, and VBR video [LE94,BE94]. We have calculated various statistics for the LAN data and VBR video. Packet length distributions, interarrival time distributions, autocorrelation functions, IDC and variance-time analyses for LAN traces are presented. The distribution of the num-

ber of cells per frame, autocorrelation function, IDC and variance-time analysis for VBR sequences are also presented.

2.2 LAN traces

The traces of actual Bellcore Ethernet traffic data are available with *ftp* from *Bellcore.com*, directory *pub/lan_traffic*. They consist of the time stamp (representing the time in seconds since the start of a trace) and an integer length (representing the Ethernet packet length in bytes). Four Ethernet traces that are widely referenced in the literature will be considered. These traces are part of traffic measurements that was done by Lelend, et. al. The traces are referenced as pOct.TL, pAug.TL, OctExt.TL and OctExt4.TL [LEL91].

In the following, we consider packet length, interarrival distributions, autocorrelation, IDC and variance-time analysis of the four traces. Traces pOct.TL and pAug.TL are internal traffic and consists of all packets on a LAN. Traces OctExt.TL and OctExt4.TL are remote or external Ethernet traffic, consisting of all those Ethernet packets that originate on one LAN but are routed to another LAN, that is, WAN. The maximum packet size in Ethernet is 1518 bytes, used mostly during file transfer applications. Ethernet peak rate is 10 Mbps. Table 2.1 shows the Ethernet traces, their packet length and type.

TABLE 2. 1 Traces of Ethernet Traffic Measurements [LEL91]

Ethernet trace	Length in packets in [sec]	Trace type
pOct.TL	100,000 [210]	LAN
pAug.TL	100,000 [252]	LAN
OctExt.TL	100,000 [28148]	WAN
OctExt4.TL	100,000 [4935.8]	WAN

2.2.1 Packet length

Statistics of our estimates for the four Ethernet traces are shown in Table 2.2. Figures 2.1 - 2.4 display, respectively, the histograms of the packet length distribution as a function of the packet data field length for the above given traces. The Ethernet protocol forces all packets to have at least the minimum size of 64 bytes and at most the maximum size of 1518 bytes. The change in the packet size scale of figure 2.3 is because the maximum packet length in bytes for the OctExt.TL trace is 594 while for the other traces it is 1518.

TABLE 2. 2 Statistics of the Ethernet packet length for pOct.TL, pAug.TL, OctExt4.TL, and OctExt.TL.

Parameter	Maximum Packet length in Bytes	Mean Packet Length in Bytes	Median Packet Length in Bytes	Minimum Packet Length in Bytes	Mean Arrival rate in Bytes/sec	Peak / Mean Ratio	Hurst Parameter
pOct.TL	1518	537.22	174	64	254920.0	2.826	0.78
pAug.TL	1518	464.17	162	64	184140.0	3.270	0.80
OctExt.TL	594	160.15	64	64	568.9533	3.71	0.88
OctExt4.TL	1518	191.59	64	64	2116.60	7.92	0.90

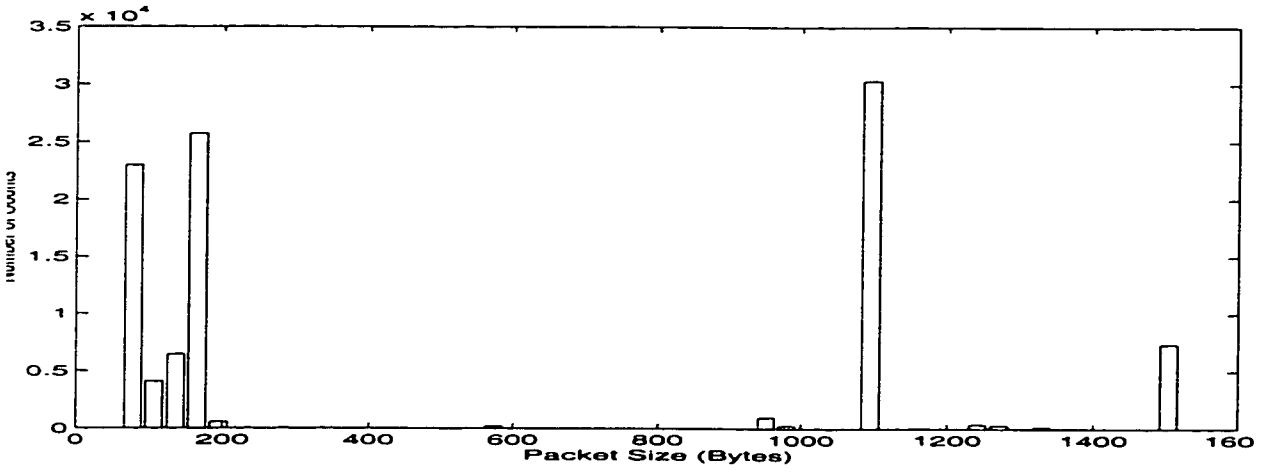

FIGURE.2.1. Histogram of packet length distribution for pOct.TL

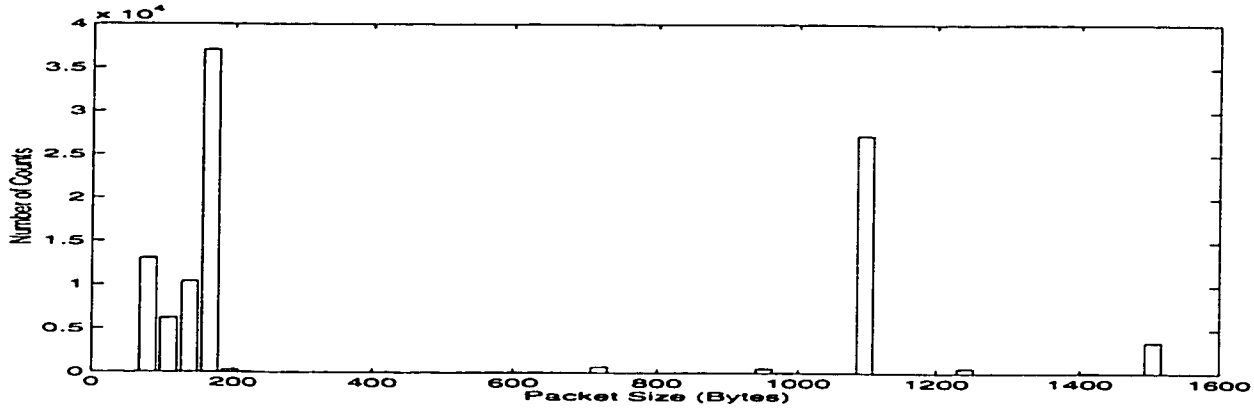

FIGURE.2.2. Histogram of packet length distribution for pAugt.TL

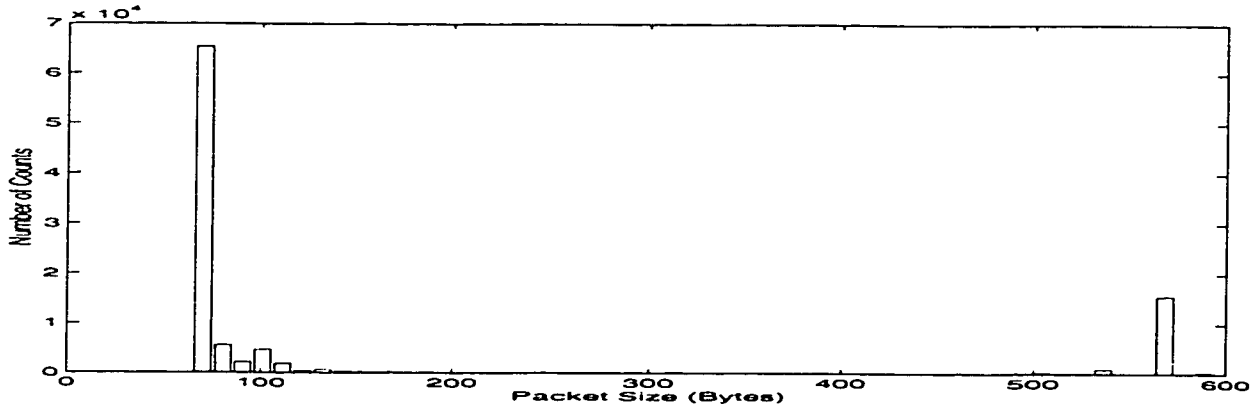

FIGURE.2.3. Histogram of packet length distribution for OctExt.TL.

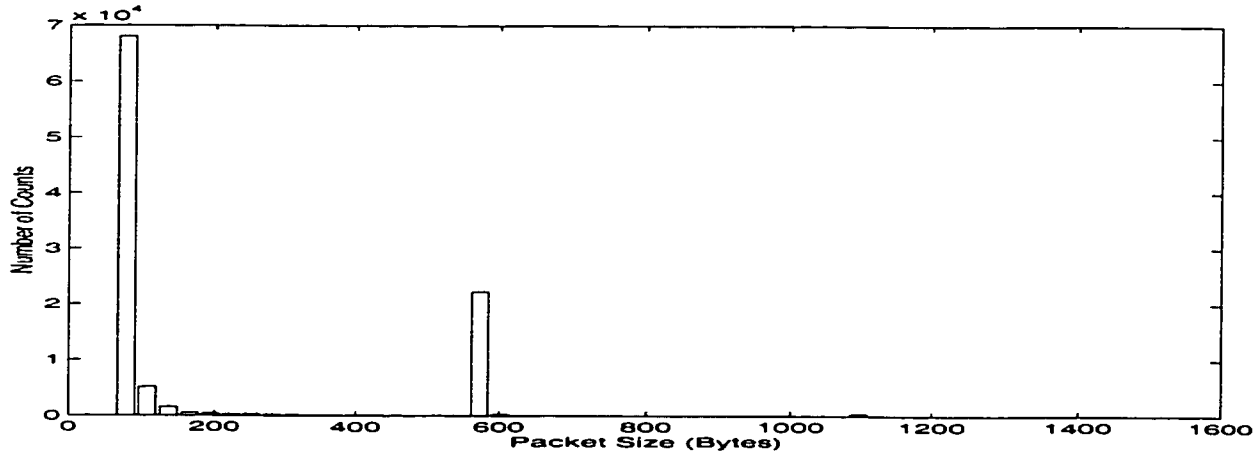

FIGURE.2.4. Histogram of packet length distribution for OctExt4.TL.

2.2.2 Interarrival time

Interarrival time is computed as the difference between the times when the transmission of two packets in a sequence began. Figure 2.5 shows our estimate of the histogram of packet interarrival time distribution for LAN internal traces pOct.TL and pAug.TL, while figure 2.6 shows the histogram of packet interarrival time distribution for WAN external traces OctExt.TL and OctExt4.TL. Notice that, the time scale in figure 2.6 is approximately two orders of magnitude larger than that of figure 2.5. This is due to that the external traces OctExt.TL and OctExt4.TL have interarrival times correspondingly larger than internal traces pOct.TL and pAug.TL. The mean interarrival time for the internal traffic pOct.TL and pAug.TL is smaller than that for the external traffic OctExt.TL and OctExt4.TL, which results in higher mean arrival rate for the internal traffic (see Table 2.1). Moreover, histograms of the packet interarrival time for both internal and external traces have large mass at the beginning. This is because of the large number of packets arrive with small interarrival times.

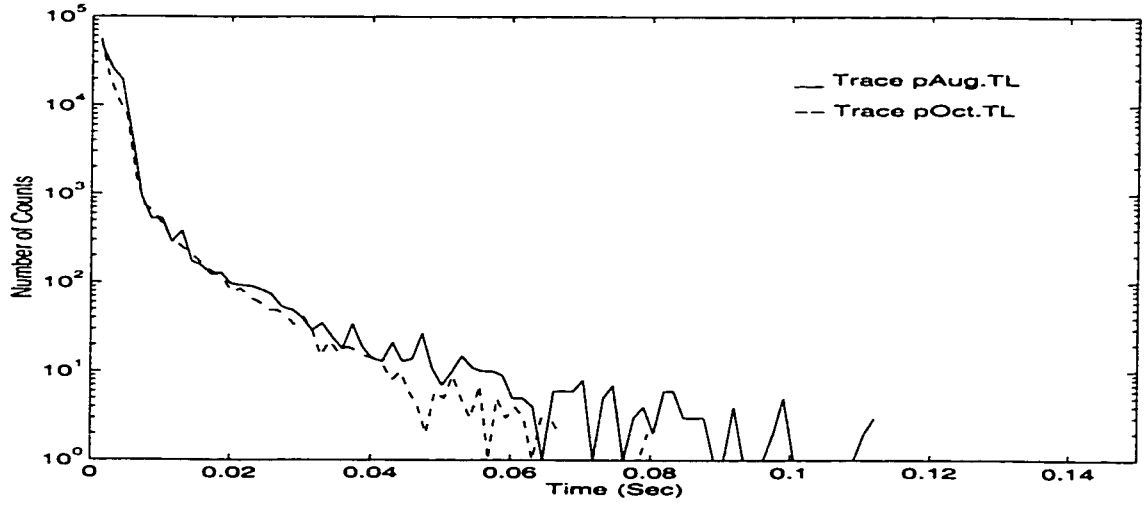

FIGURE.2.5. Histogram of packet interarrival time distribution for pOct.TL, and pAug.TL internal traces

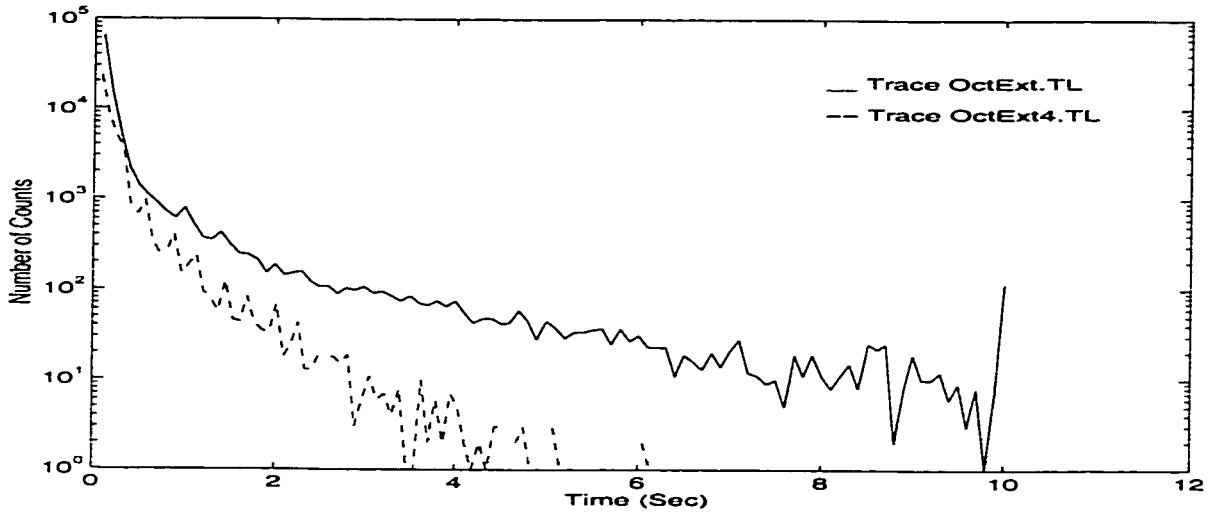

FIGURE.2.6. Histogram of packet interarrival time distribution for OctExt4.TL, and OctExt.TL external traces

2.2.3 Autocorrelation, IDC and variance-time analysis

In figure 2.7 we show our calculation of the autocorrelation function of the four traces. The autocorrelation function of OctExt4.TL is larger than that of OctExt.TL, pAug.TL and pOct.TL, respectively. As can be seen, the autocorrelation functions decays at less than exponential rate and exhibits a heavy tail property (LRD), however, they have a large correlation at low lags (SRD). OctExt.TL and OctEx4.TL traces have an autocorrelation functions that decays more slowly than the other two traces pOct.TL and pAug.TL which indicates that the former traces are more bursty than the later ones

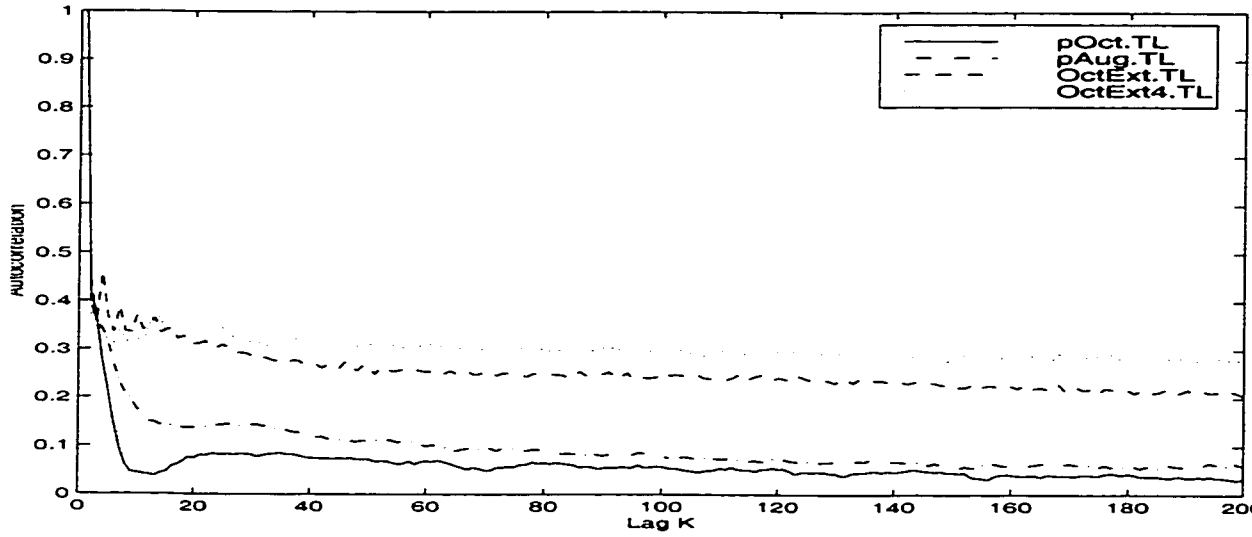

FIGURE.2.7. Estimated autocorrelation function for the four traces, pOct.TL, pAug.TL, OctExt.TL, and OctExt4.TL

Our estimation for the IDC normalized by its value at lag 1 is shown in figure 2.8. Leland, et. al. [LEL94] estimate the Hurst parameter H of pOct.T, pAug.TL and OctExt.TL to be 0.78, 0.80 and 0.86 respectively and are approximately the same as our estimate shown in the most right column of Table 2.2. Our estimation of the H parameter for OctExt4.TL is 0.904. As can be seen in figure 2.8, the slope of OctExt4.TL is larger (which means larger Hurst parameter H and more bursty trace as shown in Table 2.2) than that of the other traces OctExt.TL, pAug.TL and pOct.TL, respectively.

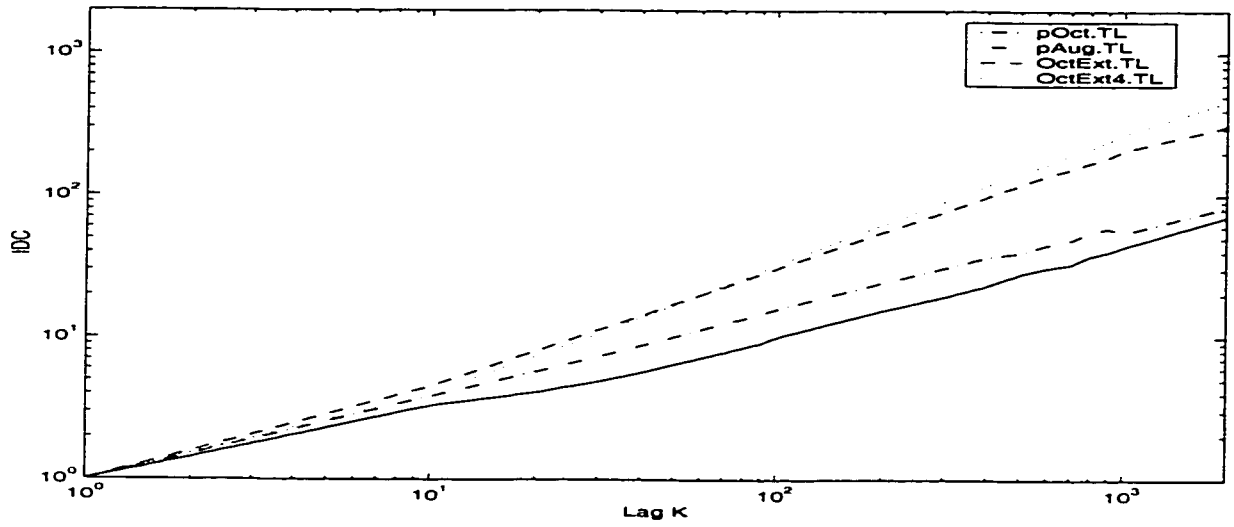

FIGURE.2.8. Estimated IDC for the four traces, pOct.TL, pAug.TL, OctExt.TL, and OctExt4.TL.

The so-called variance-time plots are obtained by plotting $\log 10(var(X^{(k)}))$ against $\log 10(k)$. See section 1.7.2 for the definition of $X^{(k)}$. The relationship between the slope of the variance β and H parameter is $\beta = 2 - 2H$ [COX84]. The variance time analysis for the four sets of Ethernet data are shown in figure 2.9. The variance time curves are normalized by the sample variance at lag 1. As shown, we have a slowly decaying variance for the four traces. That is, the variance of the sample mean decreases more slowly than the reciprocal of the sample size. Refer to condition *v* of section 1.72. The results are in agreement with that of the autocorrelation and the IDC where WAN OctExt4.TL and OctExt.TL are more correlated than LAN pAug.TL and pOct.TL

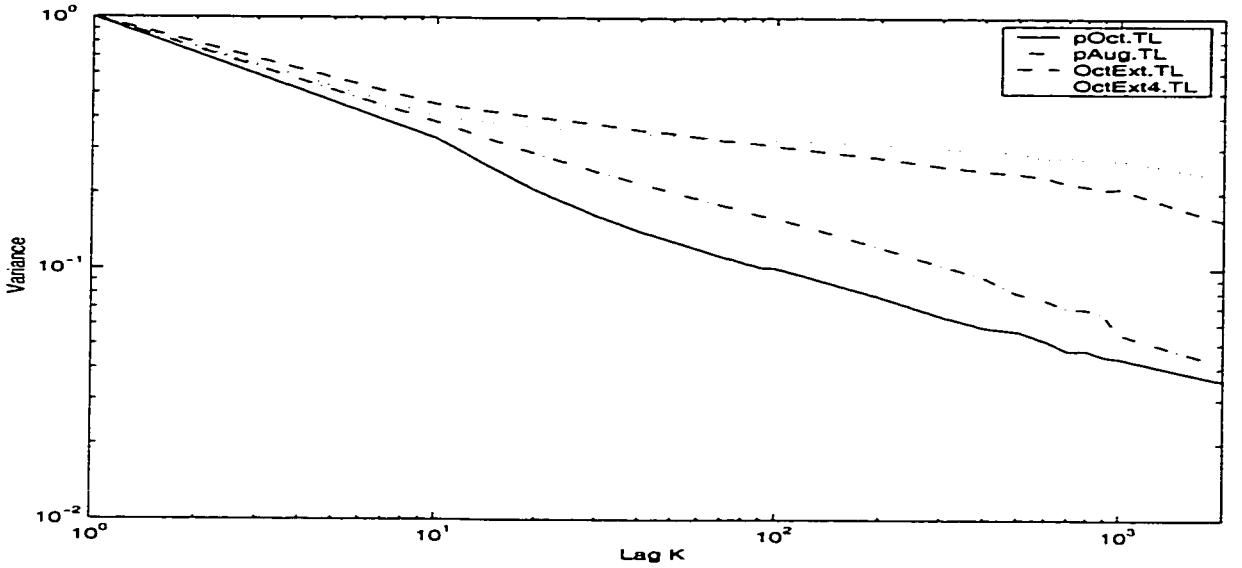

FIGURE.2.9. Estimated variance-time analysis for the four traces, pOct.TL, pAug.TL, OctExt.TL, and OctExt4.TL.

2.3 Video traces

Video is expected to be among the dominant services of future broadband networks*, with respect to both the number of users and the resulting traffic volume as we have shown in figure 1.1. Video data is a hot area of research and still its behavior is not well known as we will discuss in the next chapter. The statistics of the traffic generated by a video source are greatly influenced by the nature of the pictures transmitted, by the QoS provided, and by the coding technique adopted. The video traces that available to us are generated at a rate of 25 frames/sec. Four different video sequences worth of actual video, representing different kinds of scenes, and recorded using different types of VBR video codes are considered. The sequences consist of frame number and the number of cellsper frame. For a summary description of the different sequences, including

*For these reasons, most of our modeling and performance studies will be on video

length of the sequence, number of bytes/cell, provider and scene type, see Table 2.3.

TABLE 2. 3 Qualitative description of available VBR video sequences.

VBR Video sequence	Length in Frames [in Minutes]	bytes / cell	Provider	Scene Type
confcam	48497 [32.4]	64	Siemens	video-conferencing
eva	48600 [32.4]	64	Siemens	Video-phone
issaural	50151 [33.4]	14	Alcatel	Popular TV series
film	51364 [34.2]	14	Alcatel	Movie

2.3.1 The distribution of the number of cells per frame

Some statistics of our estimation for the four video traces are shown in Table 2.4. Our estimation of the number of cells per frame histograms for video sequences video-conferencing, video-phone, TV series and Movie are shown in figures 2.10 - 2.13, respectively. In these figures the distribution of the number of cells per frame follows similar histograms and resembles a Gamma function [HEY92]. As can be seen from column 6 of Table 2.4 and from these figures, as the median to mean ratio increases the histogram of the number of cells per frame has a smaller mass at the beginning. TV series sequence has smaller mass at the beginning than other sequences because it has the largest median to mean ratio, however, sequence video-phone has the largest mass at the beginning because of its smaller median to mean ratio. Furthermore, all sequences have number of cells per frame distributing that is skewed to the right (tail on the right) and the mean lies to the right of the median. For symmetrical distribution, the mean and median are equal. Column 7 of Table 2.4 shows the peak-to-mean ratio of the four sequences. Figure1 of [HEY96] shows the peak-to-mean ratio for different VBR video sequences including Movie and TV series of values about 2.24 and 2.21

respectively. Our estimation for video-conferencing and video-phone are shown in Table 2.4. The two sets of results are certainly well within statistical variation.

TABLE 2. 4 Estimated statistics of the number of cells per frame for traces video-conferencing, video-phone, TV series and Movie.

Parameter	Maximum number of cells per frame	Mean number of cells per frame	Median number of cells per frame	Minimum number of cells per frame	Median to Mean ratio	Peak / Mean ratio	Hurst Parameter
video-conferencing	629.0	130.2967	113.0	23.0	0.867	4.827	0.72
video-phone	897.0	170.62	146.0	21.0	0.856	5.257	0.74
TV series	11801.0	5336.4	5108.0	2523	0.957	2.211	0.90
Movie	13325.0	5948.4	5690.0	3649	0.957	2.240	0.96

It is clear from the last two columns of Table 2.4 that it is not appropriate to compare teleconferencing data with entertainment data, because of the wide differences in their statistical characteristics. Also as shown in the table, entertainment data such as TV series and Movie have higher; maximum number of cells per frame, minimum number of cells per frame, median number of cells per frame and Hurst parameter than teleconferencing data, video-conferencing and video-phone. However, teleconferencing data have higher peak to mean ratio than entertainment data does. That is teleconferencing data have larger peak to mean ratio but they are less correlated than entertainment data. This is a clear indication that it is not appropriate to compare the two kinds of traffic in terms of their QoS and other congestion control criteria which we will discuss in due course.

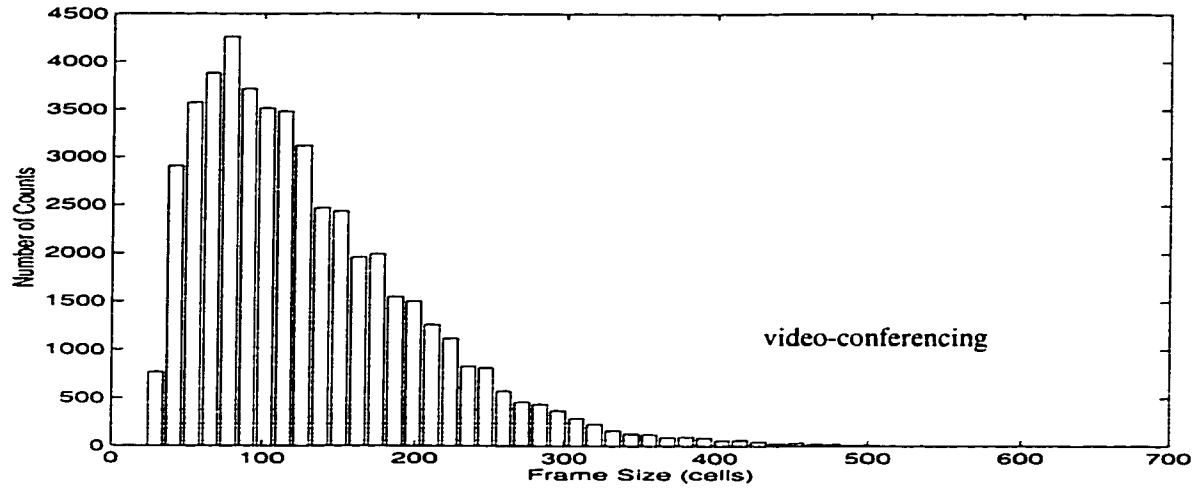

FIGURE.2.10. Histogram of the number of cells per frame for video-conferencing sequence.

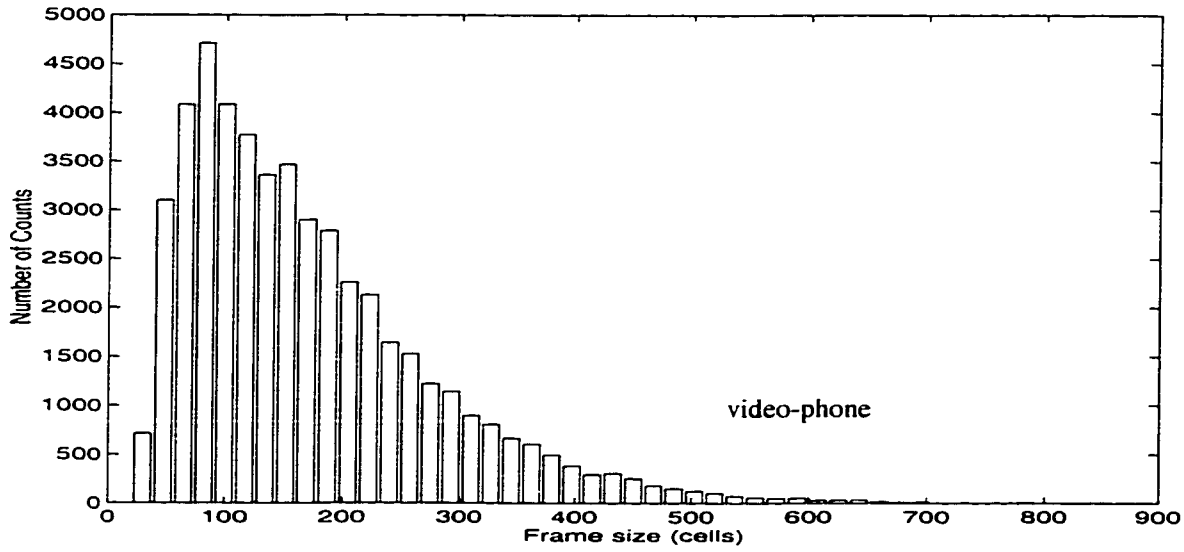

FIGURE.2.11. Histogram of the number of cells per frame for video-phone sequence.

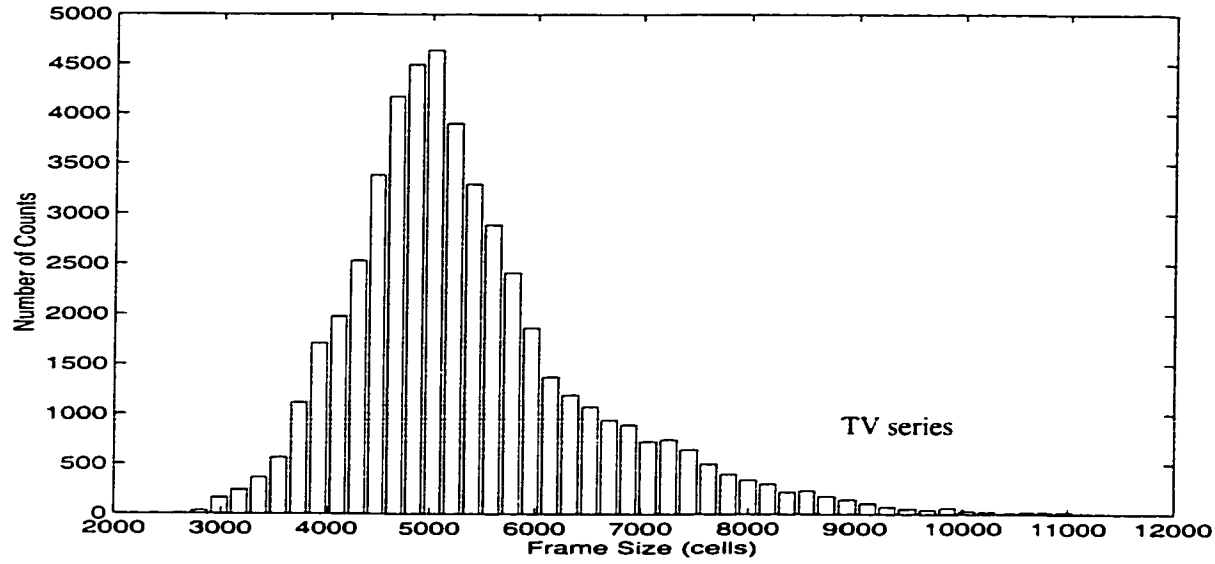

FIGURE.2.12. Histogram of the number of cells per frame for TV series sequence.

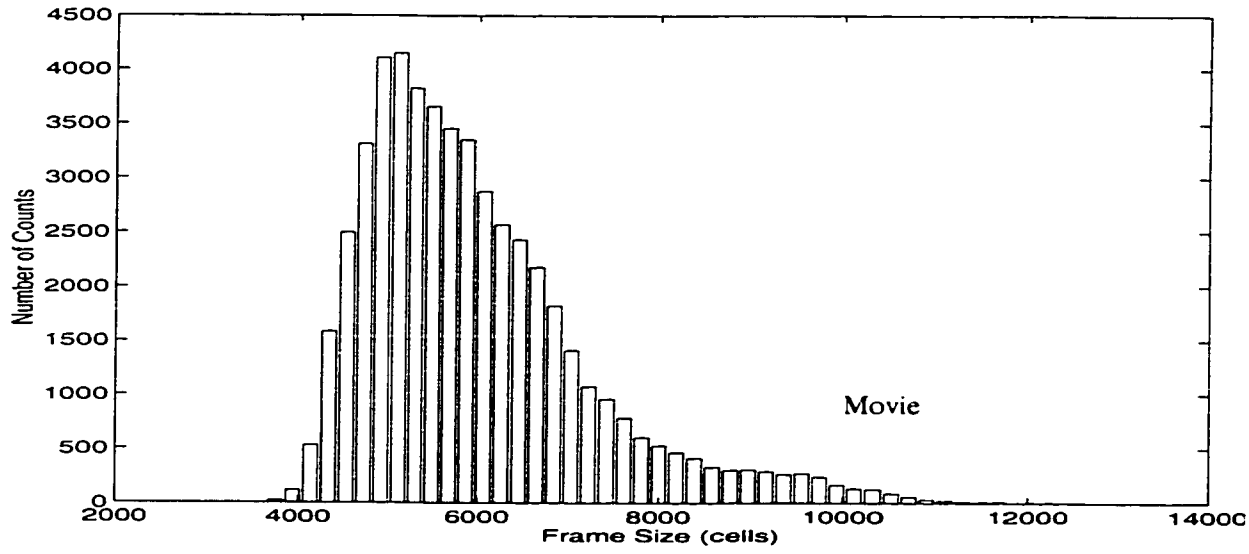

FIGURE.2.13. Histogram of the number of cells per frame for Movie sequence.

2.3.2 Autocorrelation, IDC and variance-time analysis

The importance of long-range dependence is determined by the importance of autocorrelation, IDC and variance of the traffic for very large values of lag. In

figure 2.14 the autocorrelation functions that we have calculated for the four sequences are shown. Figure 2a - b, figure 4 and figure 5 of [HEYM96] shows the autocorrelation functions of video-conferencing, video-phone, Movie and TV series sequences, respectively, and are exactly the same as that based on our calculations which shown in figure 2.14. The Movie sequence has a larger auto-correlation function than that of TV series, video-phone, and video-conferencing, respectively. As can be seen, all sequences have a large autocorrelation at low lags (strong SRD at low lags). As the lag increases, the correlation decreases in a non-exponential decay. The rate of decay has to do with the type of the video data and therefore with its burstiness.

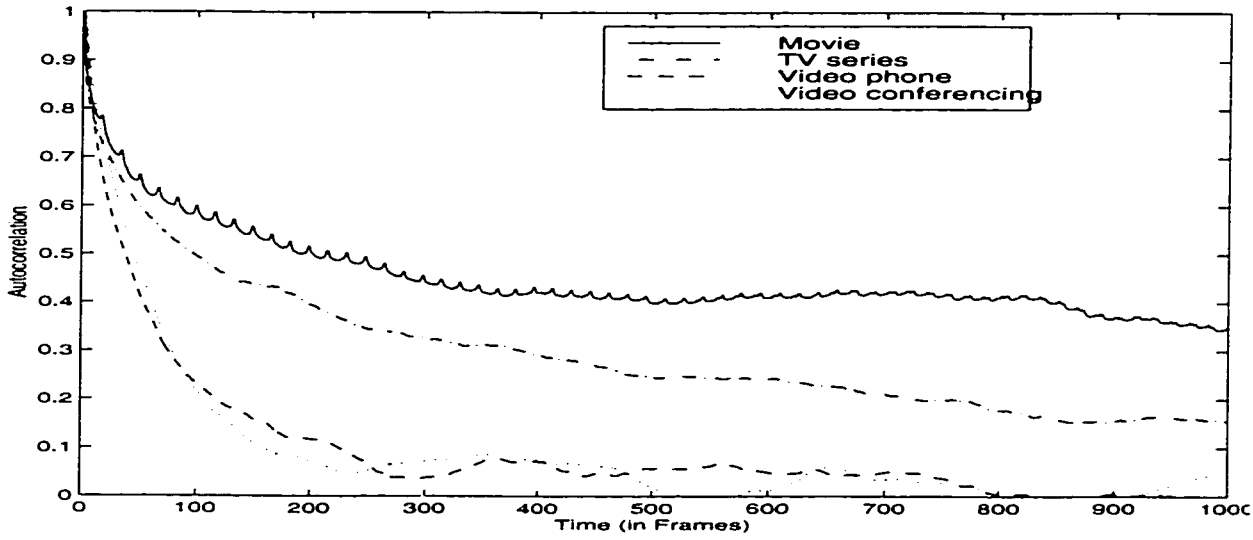

FIGURE.2.14. Estimated autocorrelation function plots for the four video sequences, video-conferencing, video-phone, TV series and Movie.

In figure 2.15, we plot the IDC normalized to its value at lag 1 for the VBR video data. Beran, et. al. [BER95] (see also [HEYM96]) estimate the Hurst parameter H for video-conferencing and video-phone to be about 0.7, and for TV series to be about 0.9; their estimate for the Hurst parameter H for Movie is greater than one. As shown in Table 2.4, our estimates for video-conferencing, video-phone and TV series are 0.72, 0.74, and 0.9 respectively. For Movie it is 0.96 which is different from the value estimated by Beran, et. al. of value greater than 1.0. As shown in figure 2.15, the slope of the IDC for Movie is larger than that of TV series, video-phone and video-conferencing respectively. The most right column

of Table 2.4 shows the estimated H parameter for the four sequences. These results show that Movie sequence is more bursty than the TV series, video-phone and video-conferencing, respectively, which results in higher Hurst parameter H.

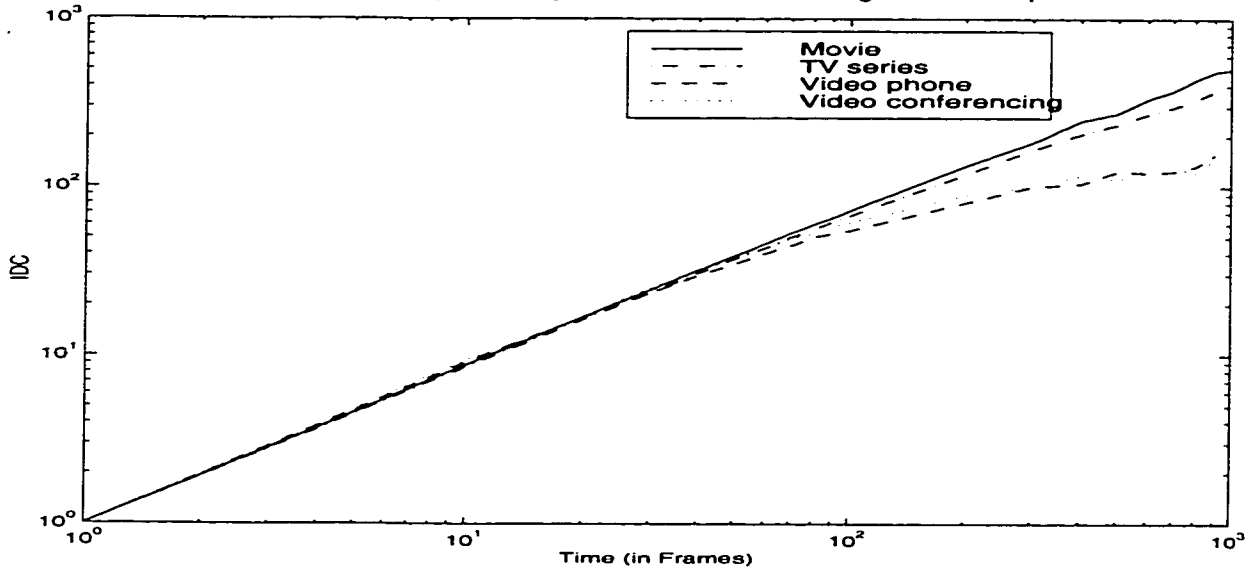

FIGURE.2.15. Estimated IDC plots for the four video sequences video-conferencing, video-phone, TV series and Movie.

Figure 2.16 depicts the variance-time plots corresponding to the four VBR video data. The variance time curves are normalized by the sample variance at lag 1. As shown, highly bursty traffic such as Movie and TV series decrease more slowly than the sample size as compared with the medium correlated traffic such as video-phone and video-conferencing (see condition *v* of section 1.7.2).

From the above results, it is clear that self-similar traffic mathematically manifests itself in a number of equivalent ways; autocorrelations decay hyperbolically rather than exponentially and variance of the sample mean decrease more slowly than the reciprocal of the sample size. This difference in the autocorrelations rate of decay, slope of IDC, and rate of decrease of the variances, suggest that long-range dependence parameters such as H-value might be useful for differentiating between types of VBR scenes

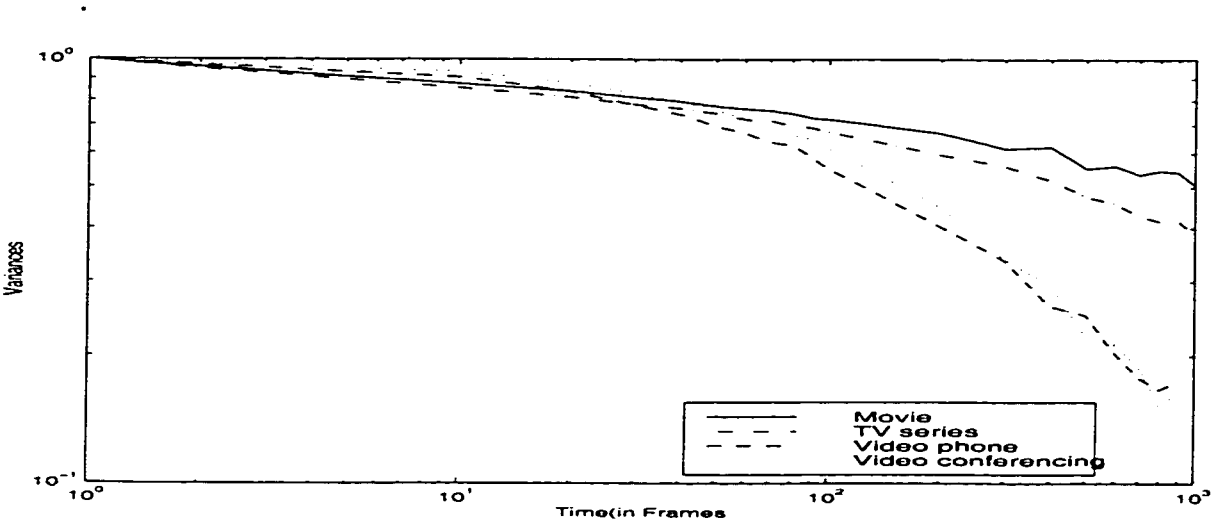

FIGURE.2.16. Estimated variances-time plots for the four video sequences, video-conferencing, video-phone, TV series and Movie.

2.4 Discussion

We investigated different Ethernet LAN, WAN and video data traces. We calculated different traffic statistics for both kinds of data. As the interarrival time distribution gets longer for the Ethernet packets, the traffic becomes more bursty. This can be seen from the autocorrelation, IDC and variance-time plots, where traffic with long interarrival times such as OctExt.TL and OctExt4.TL have autocorrelations that decay more slowly than the traffic with short interarrival times such as pOct.TL and pAug.TL. Moreover, the slope of the IDC of the long interarrival times traffic is larger than the traffic with short interarrival times traffic. The distribution of the number of cells per frame for the video data has the form of the Gamma distribution. We calculated the autocorrelation, IDC and the variance time plots. From the observation, we found that entertainment traffic is more bursty than teleconferencing traffic. Entertainment traces exhibits a larger H-value. The autocorrelation and variance decay rate of the entertainments traces are less than that of the teleconferencing traces. This difference in the H-value, the rate at which autocorrelation and variance decay suggest that long-range

dependence parameters such as H might be useful for differentiating between types of VBR scenes. As shown in Table 2.2 for Ethernet data and Table 2.4 for video data, the Hurst parameter can be used as a rough indication of scene activity. For video telconferencing, for example, H tends to be smaller, typically between 0.7 and 0.75. Entertainment data have higher H-values often greater than 0.9. LAN have H-values around 0.8 while WAN has H-values around 0.9.

CHAPTER III

Characterization and Modeling of Self-Similar Traffic in ATM Networks

3.1 Introduction

As we have discussed in chapter 1, ATM networks are expected to support a diverse set of applications with a wide range of characteristics. See figure 1.2. Our interest in source characterization and modelling is focused on models that are good predictors of cell losses and queue lengths in ATM networks transporting many traffic connections. Figure 3.1 shows the current state of understanding of source characterization in ATM networks [ONV94].

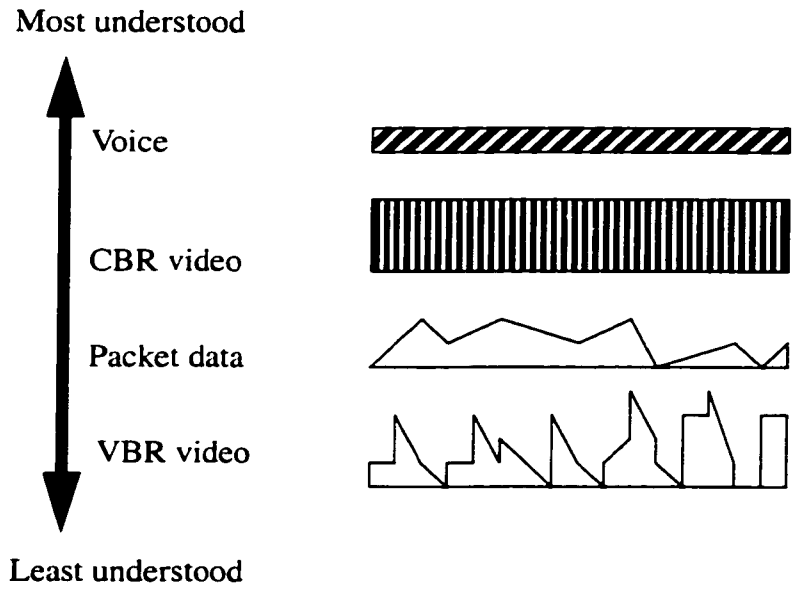

FIGURE.3.1. Current state of information on source characterization in ATM networks [ONV94].

Voice source characterization have been studied for several decades and are well understood [BRA69,DAI86,SRI86]. A voice source alternates between talk spurts (active) and silent period (idle). In the normal conversation the active

period fits the exponential distribution reasonably well while the duration of the silent periods is less well approximated by the exponential distribution. CBR video is a continuous bit stream of constant rate and can be simply described by their peak rate. The burstiness of a CBR source is equal to one, and the source is active during the duration of the connection (or the silent periods are also transmitted at the peak rate). The CBR video QoS and the amount of bandwidth required in the network increases as the rate that the data flowing into the network increases. The difficulty for CBR video is that in estimating the constant bit rate to provide the desired QoS. Typical examples of CBR services includes voice, video and audio.

Conventionally, the term data is used for any applications that is not voice, audio or video. Data networks have been operational for decades; however, traffic characteristics of data sources are not well understood. The main difficulty arises due to the fact that there is no typical data connection. Large amounts of data are transmitted in a file transfer on a continuous basis during the duration of the connection, whereas only a few hundreds bytes are generated by an e-mail. The problem of characterizing the source characteristics of data applications is further complicated by the fact that it is difficult to predict in advance the traffic characteristics of a connection, even if the particular application type is known. A common data application is LAN interconnections. As we have seen in our discussion of Ethernet traffic in chapter 2, LAN sources are very bursty with very long bursts. These characteristics emphasize correlations in the arrival streams, causing a deterioration of QoS. The serious implications for ATM networks design is that

conclusions based on traditional models may not be applicable because of the self-similar nature of the traffic models.

VBR video and image are new research area. The current knowledge on their source behavior is limited and based on different system implementations. In general, the traffic generated by a typical source either alternates between the active and silent periods or has a varying bit rate generated continuously. The peak to average bit rate of a VBR is often greater than one. Video applications generate traffic in a continuous manner at varying rate. As mentioned above, video is a relatively a new service in communication networks and its traffic characteristics are not well understood. It is also quite different from voice or data in that its bit streams exhibit various types of correlations between consecutive frames. As we have seen in the discussion of section 2.3 of chapter 2, VBR video traffic exhibits the self-similarity properties and the traffic is bursty over long time intervals. Recent extensive measurements of real traffic data [BER95, LEL93], have led to the conclusion that VBR video traffic cannot be sufficiently represented by traditional models, but instead can be more accurately matched by self-similar models.

Conventional characterizations assumes that packet traffic consists of alternating active and silent periods with well defined statistics. In contrast, studies have noted that there is no natural burst length, and bursts occur over many time scales [LEL94]. At every scale, bursts resolve into bursts over smaller time scales, and so on, over many time scales. It is this "burst within burst" structure that underlies the self-similarity properties observed in actual traffic data. The chal-

lenge is to capture this complexity in a way that can be analyzed and applied in practice.

Fowler and Leland [FOW91] examine the burstiness of data traffic over a wide range of time scales, and discusses the impact of burstiness on network congestion. Norros [NOR94] proposed a stochastic process as a model for the content of storage having self-similar input and being emptied at a constant rate. The tail behavior of the steady-state queue-length distribution, for large buffer, in a single server, infinite-capacity queue is derived. Huang, et. al. [HUA95] presented a unified approach which, in addition to accurately modeling the marginal distribution of empirical records, also models directly both the short-range dependence and long-range dependence of empirical autocorrelation structures. Veitch [VEI93] provided models which are capable of describing the long-term correlations and self-similar burstiness found in packet networks and in VBR video. Two processes designed to capture the self-similarity are presented; the fractal arrival processes (continuous and discrete), and the WSS processes. Gusella [GUS91] propose a fitting procedure based on measurements of both intervals and counts for the MMPP. His approach is simpler than that of Heffes [HEF86]. Gusella did not find the performance measures of the traffic, while in [HEF86] their models predicts the mean and the variance of queueing delays; however, the QoS parameter that is of greatest interest is cell loss.

3.2 Traffic modeling in ATM networks

Modeling of the arrival process is an essential part of ATM network design and performance evaluation. Three major categories of traffic models have been considered; input traffic models for voice sources, input traffic models for data sources, and input traffic models for video sources ([HEF86, SRI86, MAG88, BAI91, BER95]). For voice sources, delay is the most critical QoS requirement; while moderate cell loss can be tolerated (see figure 1.3). Data sources have QoS requirements different from voice. These services are usually very sensitive to cell loss (thus, demanding a very small cell loss rate), while their delay requirements are not so strict. Video is sensitive to both cell loss and delay. Small cell loss and delay are required such that required QoS and small delay jitter, (the variance of the delays), are achieved.

In this section we survey commonly used traffic models. Such models are employed in two fundamental ways: either as part of an analytical model, or to drive a discrete-event simulation. We treat each in turn. The most common modeling context is queueing, where traffic is offered to a queue or a network of queues and various performance measures are calculated.

3.2.1 Analytical models

The most general model of an ATM traffic source would be one taking into count the entire past history of cell generation so as to determine the time of the next cell to be generated. Such a model would be very complicated to describe, as well as very hard to fit to actual sources. It would also be analytically intractable. It is required to have models that have a small number of fitting parameters to make them analytically tractable and at the same time have the same properties as the actual sources. However, such models cannot possibly fit satisfactorily to all kind of sources. Nevertheless, general models have attracted significant attention, and have been applied successfully in several contexts. In the following we

discuss some general models, Poisson and its discrete-time analog Bernoulli models, Markov chain model, Markov-modulated Poisson process models, and the fluid-flow model.

3.2.1.1 Poisson processes

Poisson models are the oldest traffic models, dating back to the advent of telephony [PAP84]. A Poisson process can be characterized as a renewal process whose interarrival times $\{A_n\}$ are independent and exponentially distributed with rate parameter λ.

$$P\{A_n \leq t\} = 1 - \exp(-\lambda t) \qquad (3.1).$$

equivalently, the arrival process, satisfies the Poisson distribution,

$$P\{N(t) = n\} = \exp(-\lambda t)((\lambda t)^n / n!) \qquad (3.2).$$

where $N(t)$ is the number of arrivals in $(0, t)$ which in disjoint intervals are statistically independent (a property known as independent increments).

Poisson processes have some elegant analytical properties. First, the superposition of independent Poisson process results in a new Poisson process whose average rate is the sum of the component average rates. Second, the independent increment property renders Poisson a memoryless process. This, in turn, greatly simplifies queueing problems involving Poisson arrivals.

Generation of data from a single data source is frequently characterized by a Poisson arrival process (continuous time case) or by a geometric arrival process (discrete time case). An extension to this is the compound Poisson model, where arrivals are generated in batches; the batch generation times still form a Poisson process. The batch size may assume a general distribution; however, when restricted to the geometric distribution, analytical expressions can be obtained more easily.

3.2.1.2 Bernoulli processes

Bernoulli processes are the discrete-time analog of Possion processes. Here the probability of an arrival in any time slot is p, independent of any other one. It follows that for k slots, the probability distribution for the number of arrivals is binomial,

$$p\{N_k = n\} = \binom{k}{n} p^n (1-p)^{k-n}, \quad n = 0, \ldots\ldots k. \tag{3.3}$$

On taking the limit $k \to \infty$ and $p \to 0$ while keeping $kp \to \lambda$, the number of arrivals is Poisson. The time between arrivals is geometric with parameter p:

$$P\{A_n = j\} = p(1-p)^j, \quad j \geq 0 \tag{3.4}$$

The Bernoulli model, very often used in analytical and simulation studies, is inappropriate for modeling bursty traffic data because it assumes that there is no correlation between arrivals [HEF86,SRI86]. There are alternative approaches to describe bursty traffic which will be discussed in the sequel.

3.2.1.3 The Markov chain model

The Markov chain models can be used to approximate the self-similar traffic, even if they do not have the LRD that self-similar traffic exhibits. In [HEY96] it is shown that LRD is not a crucial property in determining acceptable QoS when the load is not heavy as would be the case when delay is important. We assume that the state transitions occur only at the end of a time slot. A complete description of the system is given by the state transition probability matrices. The conditional probabilities $P_r[X_n = j/X_{n-1} = i]]$ are called the single-step transition probabilities or just the transition probabilities. If these probabilities are independent of n, then the chain is said to be homogenous and the probabilities $P_r[X_n = j/X_{n-1} = i]]$ can be written as p_{ij}. The matrix formed by placing p_{ij} in the (i, j) location is known as the transition probabilities matrix or chain matrix (call it P). Therefore, for the homogenous chains, the 1-step transition probabilities are given by,

$$P = p_{ij} = P_r[X_n = j / X_{n-1} = i] / P_r[X_{n-1} = i] \quad (3.5).$$

In the limit; we have

p_{ij} =number of transition from i to j / number of transitions out of i

when the denominator is greater than zero.

These transition probabilities are estimated from the data which allow the calculation of the steady-state probability vector $\pi = [\pi_1, \pi_2, ...,\pi_N]$ by solving the following:

$$\pi P = \pi \quad (3.6).$$

$$\pi_1 + \pi_2 + ... + \pi_N = 1 \quad (3.7).$$

Once the steady-state probability vector π has been obtained, the computation of most of the interesting performance measures, such as the cell loss probability and the average delay, is straightforward. The drawbacks to the Markov chain model are that it has too many parameters and there is no apparent connection between the parameters and some easily measured statistics of the data [HEY92]. This limits the use of Markov chain in analysis, however, it can also be used in simulation.

3.2.1.4 Markov-modulated Poisson processes

The Markov-modulated Poisson process (MMPP) is a model of bursty traffic that has received much attention in recent years [HEF86,SRI86]. It is a powerful, analytically tractable model that can represent aggregate traffic generated by the superposition of several point processes. The MMPP process is a doubly stochastic Poisson process where the rate process is determined by the state of a continuous-time Markov chain.

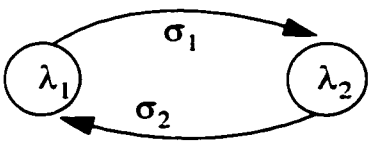

FIGURE.3.2. Two-state MMPP model.

An MMPP with higher number of states can represent traffic with higher burstiness. However, for analytical tractability, we are restricted to a small number of states. In the two-state MMPP model of figure 3.2, an aggregate arrival process is characterized by two alternating states. When the Markov chain is in state i, $(i = 1, 2)$ the arrival process is Poisson with rate λ_i, and the transition rate of going out of state i is given by σ_i. The two states of the switched Poisson process correspond to the long and short burst rates. It is usually assumed that the duration of each state follows exponential (continuous time case) or a geometric (discrete time case) distribution with different rates. Therefore, four parameters are necessary to describe an MMPP: the mean duration of each state, and the Poisson rate in each state. The infinitesimal generator Q associated with the Markov chain and the rate matrix Λ are given by:

$$Q = \begin{bmatrix} -\sigma_1 & \sigma_1 \\ \sigma_2 & -\sigma_2 \end{bmatrix}, \qquad \Lambda = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}$$

In voice traffic correlations exist only over short intervals. A common approach for modeling aggregate arrivals from a number of ON-OFF sources is to use a two-state MMPP [SRI86,HEF86]. Moreover, as shown in [MAG88] video traffic may be characterized by a MMPP model based on many independent identical ON-OFF minisources.

The Matrix geometric technique is used to deal with models such as MMPP. It proported to be a very powerful technique compared to the classical probability generating function approach in solving queueing systems problems [NEU79]. In [HUA95] heterogeneous multimedia traffic sources are modeled as an MMPP. Analytical model based on matrix analytical techniques is developed to evaluate the performance of a broadband satellite communication system for multimedia services. One advantage of characterizing the superposition of different sources

streams as an MMPP is that, once we obtain the parameters of the process, we can feed it into any system. The main objective behind the selection of certain fitting parameters is to capture the correlation effects in the arrival process. MMPP has the property of capturing the time-varying arrival rates and correlation between interarrival times. However, this is only possible over a limited time scale.

3.2.1.5 Fluid-flow processes

Another technique that models bursty traffic is stochastic fluid models [ANI82,MIT88, MAG88] where for general stochastic fluid model, the arrival rate, service rate are state dependent and the buffer content is a non-negative continuous random variable. The bursty traffic sources are modeled as Markov modulated fluid sources in which the state of the controlling continuous time Markov chain determines the rate of fluid generation. In this discussion, we will give the fluid flow more emphasis as compared with the MMPP and Markov chain processes in as much as these studies will be used later in our work.

The fluid flow approximation method has been applied successfully in various ATM traffic studies. Anick, Mitra and Sondhi [ANI82] made one of the earliest contributions, obtaining analytical models for homogenous traffic by treating the flow of cells from each active bursty source as a continuous fluid flow. Others have extended this work to multiple source types and different bursty source models. In [SEN89] this model used to capture the long term behavior of a video source corresponding to scene changes in video signals.

Typical fluid models assume that sources are bursty of the ON-OFF type shown in figure 3.3. While in the OFF state, traffic is switched OFF, whereas in the ON state traffic arrives deterministically at a constant rate R. For analytical tractability, the duration of ON and OFF periods are assumed to be exponentially distrib-

uted and mutually independent with means β^{-1} and α^{-1} respectively. N such ON-OFF sources are used and this will lead to figure 3.4 naturally.

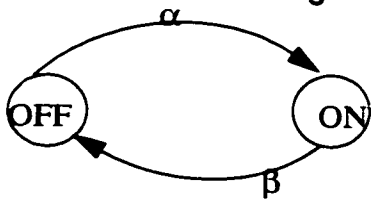

FIGURE.3.3. ON-OFF source model.

There are two fundamental reasons why fluid flow models are appropriate in an ATM environment. The small and uniform cell size (53 bytes), the constant interarrival time between cells for several contiguous cells at the time of generation fits easily in the fluid framework. The numerical complexity of solving fluid models with finite buffers does not depend on buffer size.

The equilibrium queue distribution is described by a set of differential equations together with a set of boundary equations describing the queue behavior at its limit. These equations can be solved to yield equilibrium distributions of delay and packet loss. A differential equation can be derived which governs the birth-death process shown in figure 3.4 [ANI82]. Let $p_i(t, u)$ be the probability that, the queue length does not exceed u and i sources are active at time t, c is the channel capacity and N is the number of ON-OFF sources. Then at time t, two elemental events can take place during the next interval Δt, i.e., a new source can start or a source turn off with probabilities $(N-i)\alpha\Delta t$ and $i\beta\Delta t$ respectively.

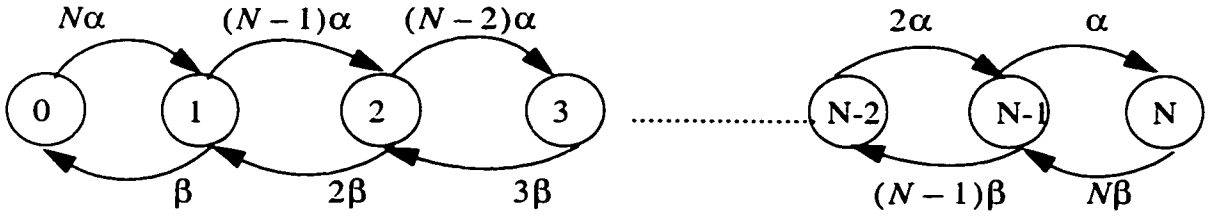

FIGURE.3.4. Birth-Death model for the superposition of N ON-OFF sources

The probability of no change is:

$$1 - \{(N-i)\alpha + i\beta\}\Delta t + O(\Delta t^2).$$

Now,

$$p_i(t+\Delta t, u) = (N-i+1)\alpha\Delta t\, p_{i-1}(t, u-(i-c)\Delta t) + (i+1).$$
$$\beta\, p_{i+1}(t, u-(i-c)\Delta t) + [1-\{(N-i)\alpha + i\beta\}\Delta t]. \qquad (3.8).$$
$$p_i(t, u-(i-c)\Delta t) + O((\Delta t)), \qquad i = 0, 1, ...,N$$

Moving $p_i(t, u-(i-c)\Delta t)$ term to the L.H.S., dividing both sides by Δt, using Taylor expansion around the point u and finally letting $\Delta t \to 0$ gives,

$$\frac{\partial}{\partial t}p_i(t, u) + (i-c)\frac{\partial}{\partial x}p_i(t, u) = (N-i+1)\alpha p_{i-1}(t, u) - \qquad (3.9).$$
$$[(N-i)\ \alpha + i\beta]p_i(t, u) + (i+1)\beta p_{i+1}(t, u)$$

As we are interested only in time-independent, equilibrium probabilities, define $F_i(u) = \lim_{t\to\infty} p_i(t, u)$, $0 \le i \le N, u \ge 0$ as the stationary probability that i sources are active and the buffer content is not greater than u. Setting $F_i(u) = 0$ for $i \notin [0, 1, ...N]$, the following linear differential equation is obtained

$$(i-c)\frac{d}{dx}F_i(u) = (N-i+1)\alpha\, F_{i-1}(u) - [(N-i)\ \alpha + i\beta]F_i(u) + \qquad (3.10).$$
$$(i+1)\beta F_{i+1}(u)$$

Substituting for the values of $i = 0, 1, 2, ...N$ in equation (3.10), we have the following in matrix notation

$$\frac{d\boldsymbol{F}}{du}(u) = \boldsymbol{A}\boldsymbol{F}(u) \qquad u \ge 0 \qquad (3.11).$$

which has a solution of the form:

$$\boldsymbol{F}(u) = e^{\boldsymbol{A}u}\boldsymbol{b} \qquad (3.12).$$

where $\boldsymbol{F}(u) = (F_0(u), F_1(u), ...F_N(u))^T$; $\boldsymbol{A} = \boldsymbol{D}^{-1}\boldsymbol{M}$,

$\boldsymbol{D} = dg(-c, 1-c, ...,N-c)$, $\boldsymbol{b}$ is a constant column vector and $\boldsymbol{M}$, an $N \times N$ matrix, is the transpose of the infinitesimal generator matrix for the underlying Markov process and is given by:

$$M=\begin{bmatrix} -N\alpha & \beta & & & \\ N\alpha & -[(N-1)\alpha+\beta] & 2\beta & & 0 \\ & (N-1)\alpha & -[(N-2)\alpha+2\beta] & & \\ & & \cdot & & \\ & & \cdot & & \\ & 0 & & -[\alpha+(N-1)\beta] & N\beta \\ & & & \alpha & -N\beta \end{bmatrix}$$

The joint source and queue length distribution $F(u)$ has a spectral expansion solution,

$$F(u) = \sum_i exp(z_i \cdot u) \cdot a_i \varphi_i \qquad (3.13).$$

a_i are the coefficients in the spectral expansion solution. $(z_i, \varphi_i)$are the eigenvalues and the corresponding eigenvectors of a generalized eigensystem representing the differential equations about $F(u)$, which satisfies the following eigenvalue problem:

$$z\varphi D = \varphi M \qquad (3.14).$$

The coefficients a_i are obtained from the boundary conditions. Decomposition results for the model allow the eigenvalue and eigenvector pairs (z_i, φ_i) to be accurately computed [KOS86].

From $F(u)$, various performance measures can be obtained. The probability of overflow beyond u (probability an arrival that is blocked by a finite buffer), which is a measure of cell loss probability in ATM networks, can be found from the queue length distribution:

$$G(u) = Pr(\text{buffer content} > u) = 1 - \mathbf{1}^T F(u) = - \sum_{i=0}^{N-c-1} exp(z_i \cdot u)\, a_i \left(\mathbf{1}^T \varphi_i\right), \qquad (x \geq 0) \qquad (3.15).$$

where $\mathbf{1}$ denotes the vector with unity for all its components and T denotes transposing.

The *nth* moment of the queue length can be expressed as:

$$E(u^n) = \int_0^\infty u^n \, d\{\mathbf{1}^T \boldsymbol{F}(u)\} = n \int_0^\infty u^{n-1} \, G(u) \, du \qquad (3.16).$$

The equilibrium probability distribution for the queue length is given by the marginal probability

$$F(u) = \sum_i F_i(u) \qquad u \geq 0 \qquad (3.17).$$

Quantities of interest for QoS evaluation are expected values of queue length and average queueing time.The expected value of the queueing length is given by,

$$\bar{Q} = \int_0^\infty u dF(u) \qquad (3.18).$$

Knowing the average queue length $\bar{Q} = E(u)$, the average queueing time can be found using Little's formula.

3.2.2 Simulation models

In the previous section, we have discussed analytical models some of which approximately characterize the LRD. Analytically tractable models for LRD traffic are an area of active research and many of the results are from simulations. While there are numerous stochastic simulation models which exhibit the LRD property, they are useful in simulation but not analysis. Among these are the autoregressive processes, the Pareto modulated Poisson processes (PMPP), the exactly self-similar fractional Gaussian noise (FGN) and its continuous version known as Fractional Brownian Motion (FBM) and the asymptotically self-similar fractional autoregressive integrated-moving average (F-ARIMA). One of the objectives of our work is to compare these techniques for generating LRD traffic and to find out which techniques gives the most accurate measures of QoS. This will be discussed in chapter 4 and chapter 5. Moreover, the synthetic traffic generated using these models is used in simulation studies for gaining a better understanding of queueing and network-related QoS issues.

3.2.2.1 Autoregressive traffic process

The class of linear autoregressive models is the most common example of autoregressive processes. It has the form:

$$X_n = \sum_{r=1}^{p} a_r X_{n-r} + b\varepsilon_n, \quad n > 0 \qquad (3.19).$$

where $X_0, X_1, \ldots X_{p-1}$ are random variables, the a_r, b are real constants, and the ε_n are zero mean, identically independent distributed (IID) random variables, called residuals, which are independent of the X_n. Equation (3.19) describes the simplest form of a linear autoregression scheme called AR (p), where p is the order of the autoregression.

The recursive form of the equation makes it clear how to randomly generate the next random element in the sequence X_n from a previous one. For a first order model AR(1), equation (3.19) has the following form:

$$X_n = a_1 X_{n-1} + b\varepsilon_n \qquad (3.20).$$

Taking the expectation of both sides of (3.20) and denoting the mean of ε_n by η, the steady-state average of the AR(1) is given by

$$E(X) = b\eta/(1-a_1) \qquad (3.21).$$

The discrete autocovariance $C(n) = E\{(X_n - E(X))(X_{n+k} - E(X))\}$ is given by

$$C(n) = \left(b^2 a_1^n\right)/\left(1-a_1^2\right) \qquad ,n \geq 0 \qquad (3.22).$$

Matching equations (3.21) and (3.22) to the real data, the AR(1) parameters a_1, b and η are easily obtained [MAG88].

This model provides a rather accurate approximation of VBR video sources [HEY92]. Going for higher order AR models will give more accurate results but with complexity trade-off. However, analysis of a queueing model with the above arrival process can be very complex and may not be tractable; therefore, this model is suitable only for use in simulations.

3.2.2.2 Pareto-modulated Poisson process

The Pareto-modulated Poisson process (PMPP) can also be used to characterize self-similar traffic [SUB95]. Similar to the MMPP case, the two state PMPP model is shown in figure 3.5. In each of the two states, the traffic is modeled as a Poisson process with rates λ_i, $i = 1\ 2$. However, the duration of each state is independent and identically distributed with Pareto distribution of parameter α. The duration of the states are chosen to have a Pareto distribution of probability density function $f_i(t) = \alpha a^{\alpha}/t^{\alpha+1}$, $a > 0, t \geq a$ which have the thick-tailed property in order to capture the long term dependency in the arrival process. Three parameters are necessary to describe the two-state PMPP: the mean duration of the state α and the Poisson rate in each state λ_1 and λ_2.

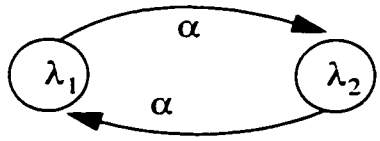

FIGURE.3.5. Two state PMPP model

No analytical results exist for finding the performance measures such as the loss probabilities and the queue length distribution of the Pareto model. All the work done is based on simulation by matching the IDC of the symmetric two-state PMPP to the real data to find the parameters and then generating synthetic traffic. In the following we consider matching IDC to the real data:

The relationship between α and H is given by [COX84]:

$$H = (3-\alpha)/2. \tag{3.23}$$

For the symmetric case of PMPP, it is shown in [SUB95] that the IDC is given by:

$$I_t = 1 + \frac{(\lambda_1 - \lambda_2)^2}{\lambda_1 + \lambda_2} \frac{(\alpha - 1)}{\alpha} \cdot t^{2-\alpha} \tag{3.24}$$

The mean arrival rate λ is given by:

$$\lambda = \frac{\lambda_1 + \lambda_2}{2} \tag{3.25}$$

The procedure that we developed of finding the PMPP parameters λ_1, λ_2, and α will be discussed in section 4.2.2 when we fit the Pareto model to the Ethernet data. Given α, λ_1 and λ_2 synthetic traffic is generated, using OPNET (a short review of OPNET will be presented in section 3.6), and then the performance measures of the generated and real data are compared.

As we have presented in chapter 2 and as shown in [LEL91], data traffic is highly bursty. A traffic model which takes into account the LRD characteristic is needed. It is shown in [SUB95] that the PMPP model is well suited to the characterization of the data. Application of this model to the real Ethernet data will be considered in the next chapter.

3.2.2.3 Fractional Gaussian noise (FGN), Fractional Brownian Motion (FBM) and Fractional Autoregressive Moving average (F-ARIMA) processes

As we have discussed in section 1.7 the crucial feature of self-similar processes is that they exhibit LRD, i.e.; their autocorrelation function decays less than an exponential rate. Fractional Gaussian Noise (FGN) is a stationary Gaussian process. The autocorrelation function of the exactly second-order self-similar FGN with $0.5 < H < 1$ is given by:

$$r(k) = \frac{1}{2}(|k+1|^{2H} - 2|k|^{2H} + |k-1|^{2H}) \qquad , \; k = 1, 2, ... \tag{3.26}$$

where k is the lag and H is the Hurst parameter.

In [MAN68] Fractional Brownian Motion (FBM) was introduced. It is a continuous time analogue of FGN. It has a correlation function of the form

$$r(t_1, t_2) = \frac{1}{2}\left\{ t_1^{2H} + t_2^{2H} - |t_1 - t_2|^{2H} \right\}, \qquad t_1, t_2 > 0 \tag{3.27}$$

where $0 < H < 1$

Fractional Autoregressive Moving Average (F-ARIMA) was introduced in [HOS84]. It is a function of three parameters; p, d, q, where p and q are non-negative integers and d is real. These processes are examples of independent asymptotically second-order self-similar processes with self-similarity parameter

$H = d + 0.5$, where the parameter d measures the strength of long-range dependence. The properties of an F-ARIMA(p, d, q) process at high lags or low frequencies are similar to those of an F-ARIMA$(0, d, 0)$ process with the same value of d. F-ARIMA$(0, d, 0)$ processes are the simplest of the fractionally differenced ARIMA processes. The F-ARIMA$(0, d, 0)$ autocorrelation function is given by

$$r(k) = \frac{(d)(1+d).....(k-1+d)}{(1-d)(2-d).....(k-d)}, \qquad k = 1, 2, ... \tag{3.28}$$

where $d = H - 0.5$, $0.5 < H < 1$

In the following, we are going to present an algorithm for generating three classes of traffic FGN, FBM and F-ARIMA processes [HOS84]. The differences among them their autocorrelation functions as given by (3.26), (3.27) and (3.28). The long synthetic FGN trace $X = \{X(k) : k = 1,2,.....\}$ is generated according to Hosking's procedure [HOS84]. This method is applicable to any Gaussian process as long as the autocorrelation function $r(k)$ is known. The initial point in the generated sequence, $X(0)$ is a sample from the Normal distribution with zero mean and variance v(0) denoted by $N(0, v(0))$. Set $M(0) = 0$ and $D(0) = 1$; the mean and the variance of n points are generated by the following iteration for $k = 1,2,....n$:

$$M(k) = r(k) - \sum_{j=1}^{k-1} \phi(k-1, j) r(k-j) \tag{3.29}$$

$$D(k) = D(k-1) - M^2(k-1)/D(k-1) \tag{3.30}$$

$$\phi(k, k) = M(k)/D(k) \tag{3.31}$$

$$\phi(k, j) = \phi(k-1, j) - \phi(k, k)\phi(k-1, k-j), \quad k = 1, ...k-1 \tag{3.32}$$

The mean and the variance are given by,

$$m(k) = \sum_{j=1}^{k} \phi(k, j) X(k-j) \tag{3.33}$$

$$v(k) = (1 - \phi^2(k, k)) v(k-1) \tag{3.34}$$

where $\phi(k, k)$ is the *kth* partial correlation coefficient of X_k and $\phi(k, j)$ are partial linear regression coefficients. The points are generated by choosing each $X(k)$ independently from $N(m(k), v(k))$.

3.3 Generation and the simulation study of self-similar traffic

In this section we use the methods discussed in the previous section to generate self-similar traffic, in particular FGN, FBM and F-ARIMA. We consider in our simulation experiments three cases, one with $H = 0.5$, which represent short range dependence (Poisson process) and one with $H = 0.7$, which represents medium bursty traffic and finally $H = 0.9$ representing highly bursty traffic.

We show in figure 3.6, figure 3.7 and figure 3.8 FGN, FBM and F-ARIMA traces with input $H = 0.5$, 0.7 and 0.9. We can see that, as the H value increases, the FGN, FBM and F-ARIMA traces indeed become more and more correlated.

We investigate the traffic characteristics of the models with the goal of identifying traffic descriptors and performance measures and formulating a traffic model designed to generate traffic for simulation. Several basic statistics for the FGN, FBM and F-ARIMA are examined. An important traffic descriptor is the burstiness, expressed here as covariance, IDC, and the variances-time analysis. To decide whether our model is acceptable or not is to find how its performance measures behave. The most important of these performance measures are, the probability of loss and the mean queue length from which the average queueing delay can be found using basic queueing theory.

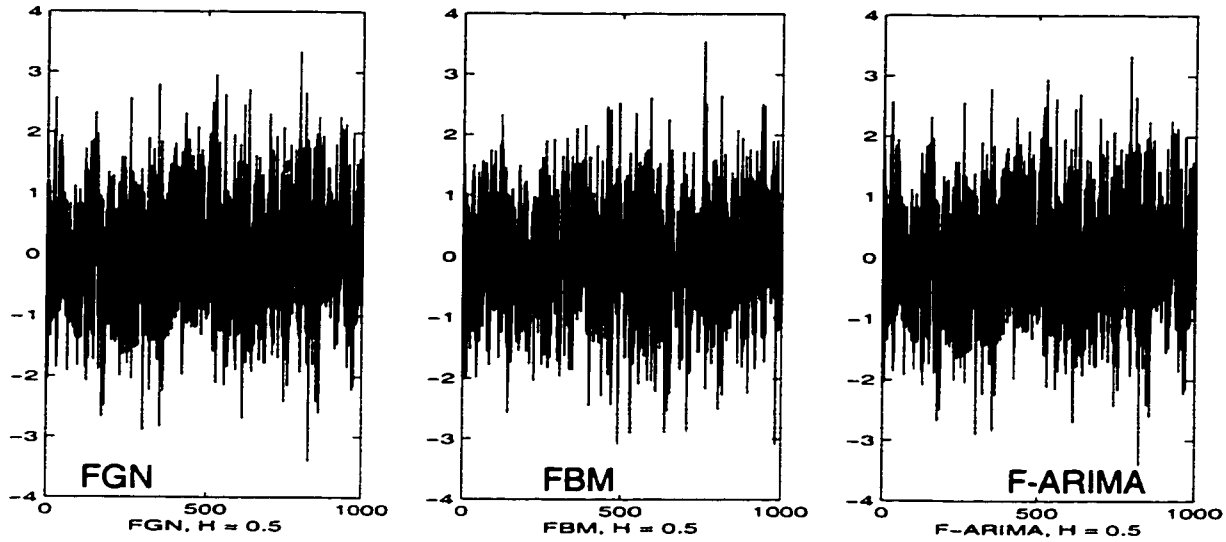

FIGURE.3.6. FGN and FBM, F-ARIMA traces for $H = 0.5$.

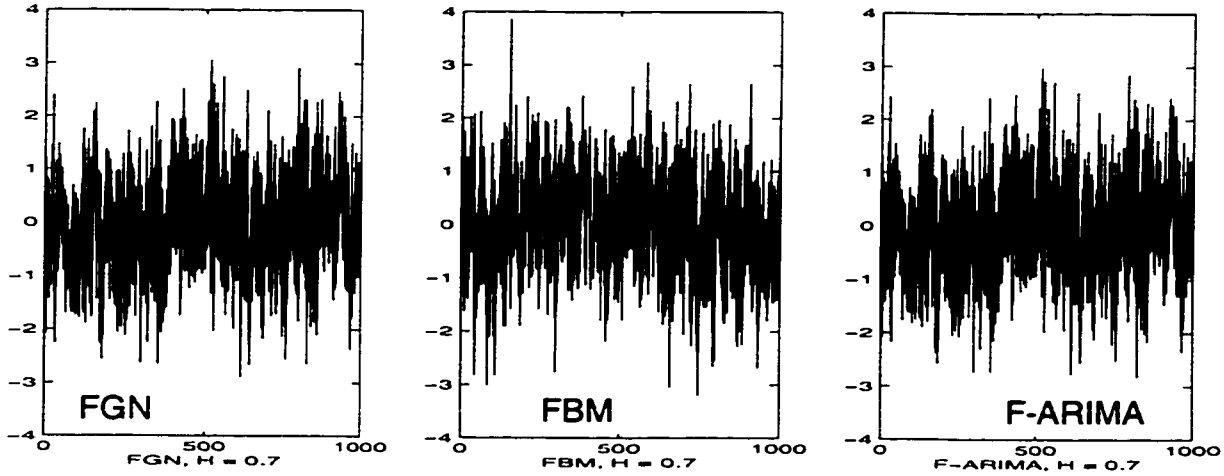

FIGURE.3.7. FGN, FBM and F-ARIMA traces for H=0.7.

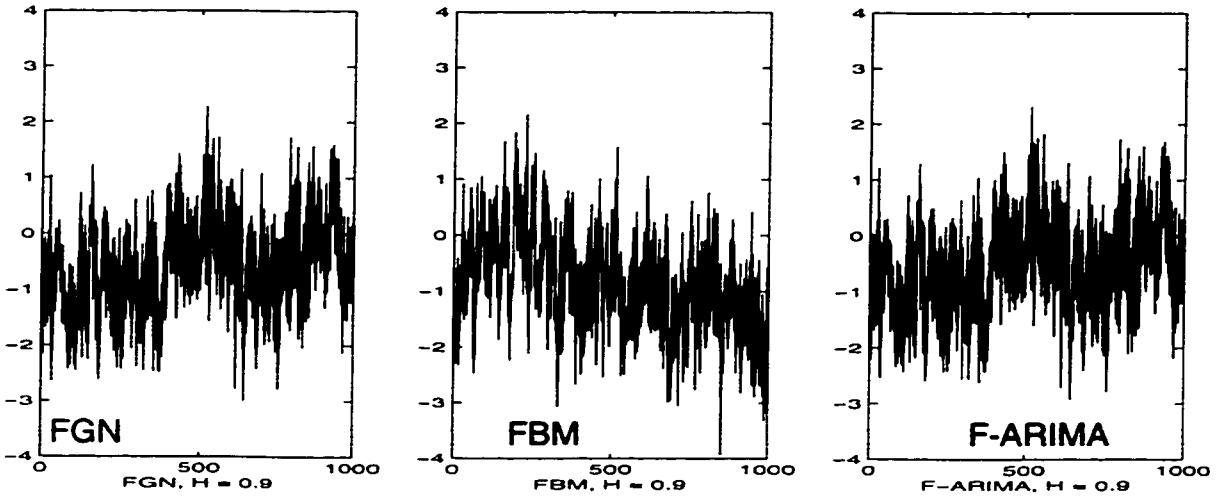

FIGURE.3.8. FGN, FBM and F-ARIMA traces for H=0.9.

3.3.1 Covariance, IDC and variance-time analysis

Based on the values of the Hurst parameter H, using Matlab software, simulated traffic is generated and measured for consistency. Figure 3.9 shows the given and the estimated values of the covariance functions of FGN, F-ARIMA and FBM. The given values are the input to model and the estimated values are those based on the data generated from the model. As can be seen, for low lag, all models have a very strong SRD. As the lag increases, and for H-values of 0.6 the covariance functions decay more quickly as compared with high H-values of 0.9. That is, highly correlated traffic has larger time constant than uncorrelated traffic.

Figure 3.10 shows the given and the estimated values of the IDC's normalized by its value at lag 1 for FBN, FGN, and F-ARIMA. The IDC is directly proportional to the Hurst parameter H through the relation (slope of $IDC=2H-1$). As can be seen, as the H-value increases, the slope of the IDC increases. For H=0.5, the slope of the IDC is 0, which means non-correlated traffic. The accu-

racy of the traffic generators is measured in terms of the difference between the given and the estimated value of H. Good results are obtained and the error, the difference between the given and estimated value of H, is very small and in the range of 5%.

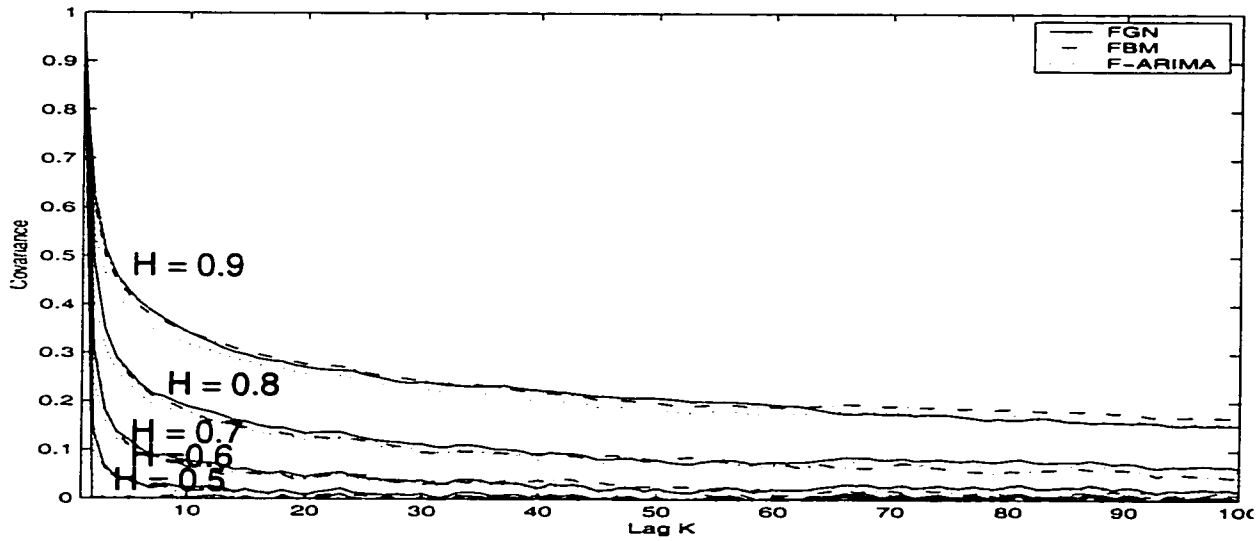

FIGURE.3.9. Simulation results for covariance function of FGN, F-ARIMA and FBM for different given values of H.

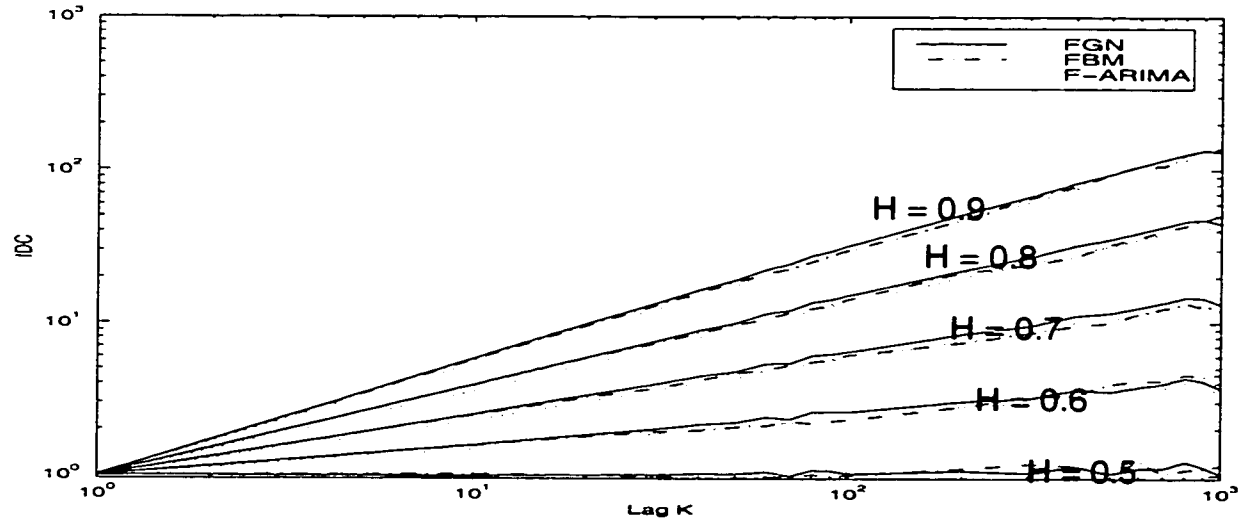

FIGURE.3.10. IDC's for FGN, FBM, and F-ARIMA for different given values of H.

Figure 3.11 shows the given and the estimated values of the variances normalized by the sample variance at lag 1 for FBN, FGN, and F-ARIMA. As the Hurst parameter H increases, the variance of the sample size decreases more slowly than the reciprocal of the sample size, which is the case when H = 0.5 as shown in figure 3.11.

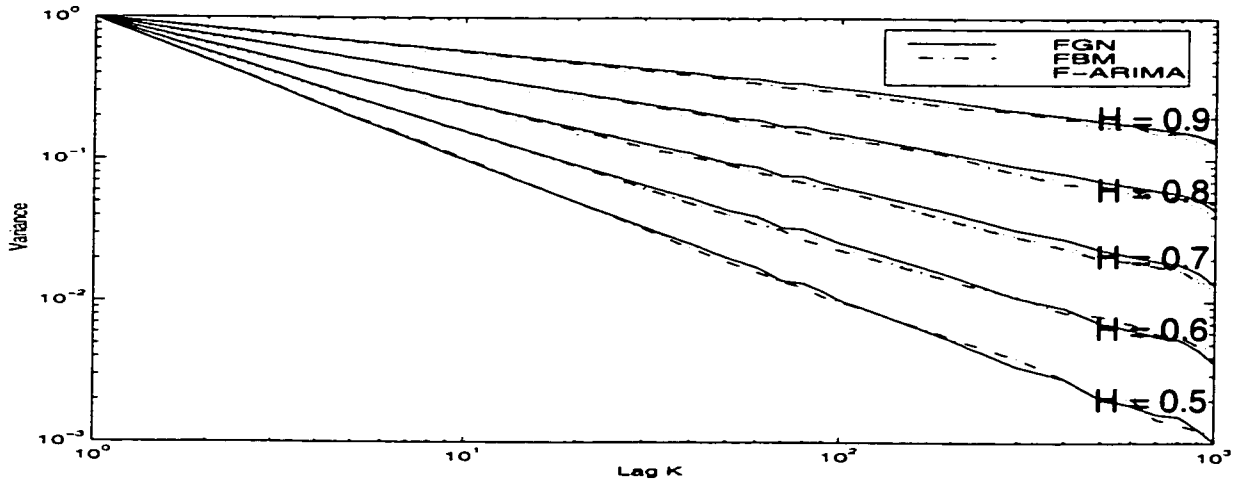

FIGURE.3.11. Variances-Time plots for FGN, FBM, and F-ARIMA for different given values of H.

3.4 Performance analysis

In an ATM network, the sources access the buffer through statistical multiplexing and when the buffer size is finite, cells can be discarded if the buffer becomes full. The real test of the utility of a model is its ability to predict probability of loss in which the size of the buffer is finite. Therefore the probability of cell loss due to buffer overflow is among the most important performance measures in an ATM multiplexer, especially when dealing with loss-sensitive traffic such as data and video [TSY97].

3.4.1 Probability of loss

Using Matlab software, we generated different traces of 100000 samples long, almost the same number as that for the real Ethernet data discussed in chapter 2 and shown in Table 2.1, for FGN, FBM and F-ARIMA models. The input to the models are, mean value, variance and the Hurst parameter. The traces have different Hurst parameter. The idea here is to see how the loss in ATM multiplexer is affected by the level of the correlation as a function of the load and constant buffer size or vice versa. We compare the probabilities of loss for the traces as a function of the load. The comparison is based on the same buffer size but different Hurst parameter. The traffic that we simulated has a mean value of 5 packets/sec. However, it is possible to fix the mean value to any desired value; it depends mainly on the input parameters to the simulator that generate the traffic. We choose to have a buffer size of 2 packets, which is reasonable compared to the mean value of the simulated trace of 5 packets/sec.

The probabilities of losses for simulated FGN, FBM and F-ARIMA traffics for different Hurst parameter H as a function of the traffic intensity with finite buffer of size 2 packets are shown in figure 3.12. As can be seen, as the Hurst parameter increases, the probability of loss increases. That is, more loss occurs for highly correlated traffic assuming the same buffer size. Moreover, as the load increases the probability of loss increases. The three models perform almost the same when the traffic intensity and the correlation index are not large. However, as the traffic intensity and the correlation index of the traffic increases, the three models become slightly different.

This is an important observation; as the correlation effect between the samples increases, the probability of loss of traffic flowing through a buffer of fixed size

increases. This observation of the effect of the Hurst parameter values on the traffic can be used later in our work when we deal with real data.

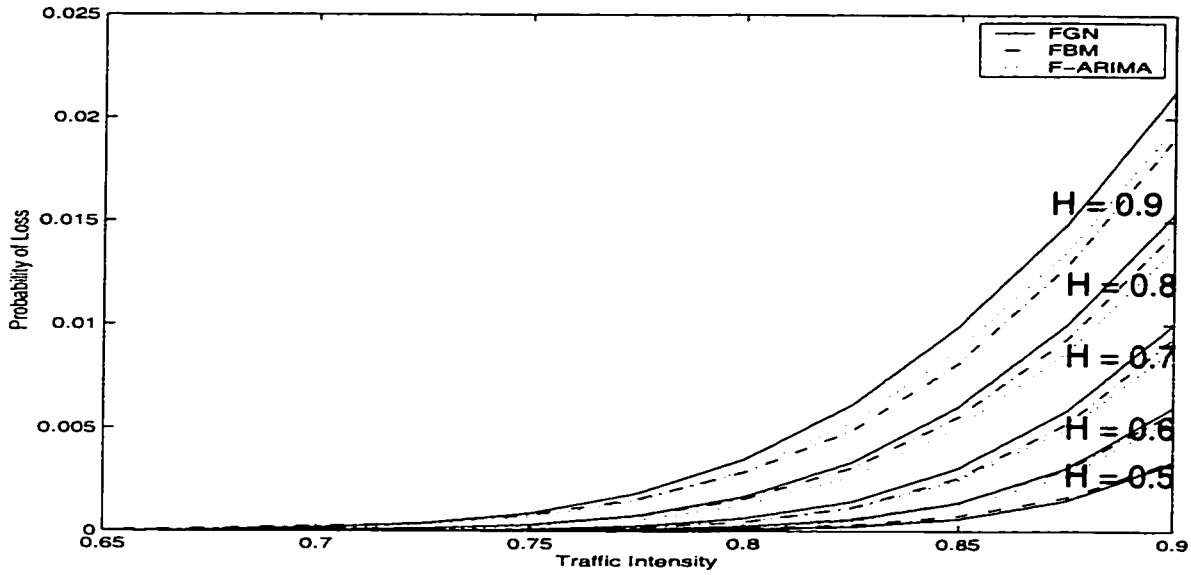

FIGURE.3.12. Simulation results for probability of loss of FGN, FBM and F-ARIMA for buffer size = 2 packets for different given values of H.

3.4.2 Mean queue length

In figure 3.13 we show the mean queue length for the FGN, FBM and F-ARIMA traffic that we generated. The generated traces have the same mean value of 5 packets/sec but different Hurst parameter H. The buffer size is fixed and has a value of 2 packets. As for the proability of loss, the mean queue length as a function of the traffic intensity increases as the traffic becomes more bursty. As the traffic intensity and the index of the correlation increases, the three simulation models, FGN, FBM and F-ARIMA become slightly different in their prediction of the mean queue length. For low traffic intensity and low correlation index, the three models behave almost the same.

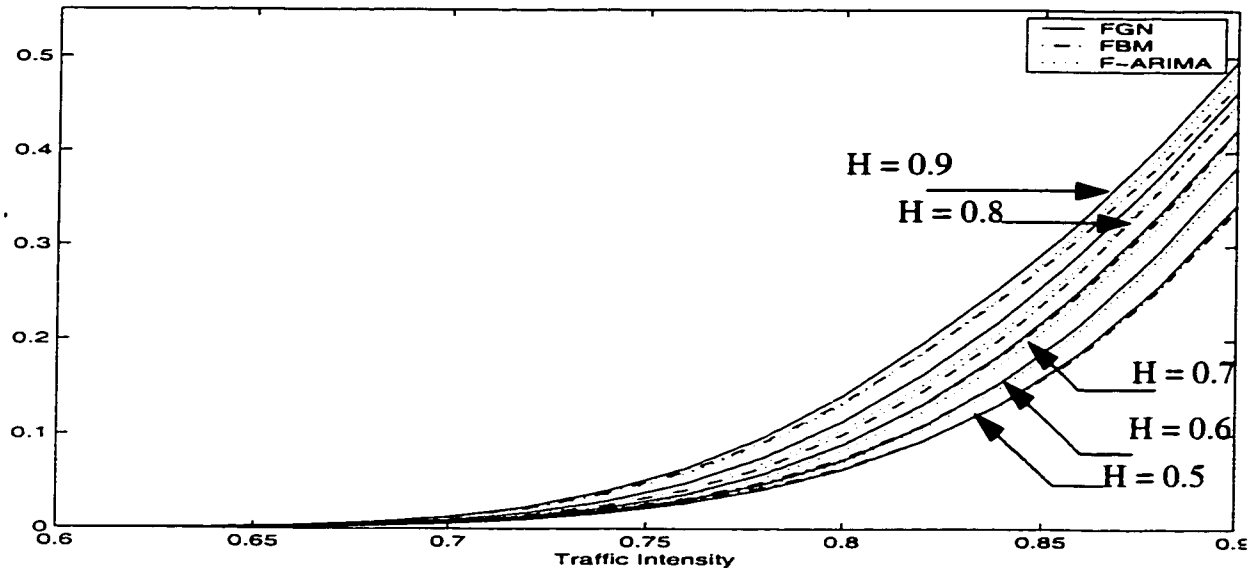

FIGURE.3.13. Simulation results for mean queue length of FGN, FBM and F-ARIMA for buffer size = 2 packets for different given values of H.

3.5 Modeling multiplexed sources

In the previous section we discussed how the probability of loss and mean queue length for self-similar processes behave as the index of correlation increases. In this section, we show that by multiplexing several statistically independent and identical highly correlated sources will result in a reduction of the probability of loss and mean queue length. With this model, the advantage of multiplexing is due to smoothing the peaks and valleys of traffic as a result of averaging. We present multiplexing of the FGN, FBM and F-ARIMA source models.

The objective here is to show how the probability of loss and mean queue length of a self-similar process behave as the number of multiplexed sources increases. The values of the probability of loss and mean queue length is not an issue here. Smaller values of probability of loss can be simulated only by using very long traces and this will take a long simulation time. The number of samples of the self-similar processes that we generated are about the same as the number

of packets for Ethernet traces and number of frames for video sequences that are available to us which we discussed in the previous chapter. That is our work is based on generating number of self-similar process samples that is comparable to the length of the data traces.

It is clear from figure 3.14 and figure 3.15 that, as the number of multiplexed sources N increases, the probability of loss and the mean queue length of the process is reduced. This is an interesting result, since increasing the number of multiplexed source will reduce the correlation between the samples of the process and in turn the traffic will be smoother. Smoother traffic obviously will be much easier to deal with when it is fed to an ATM multiplexer of finite buffer.

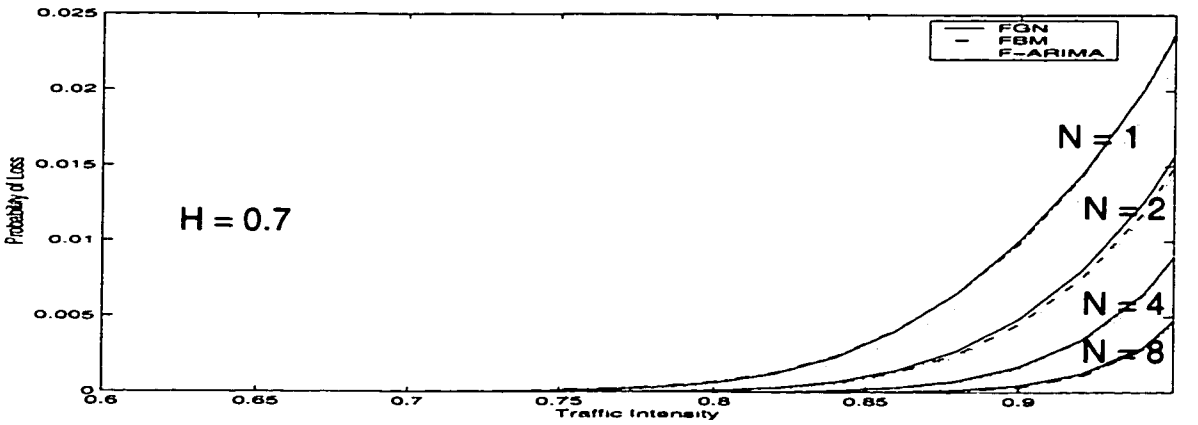

FIGURE.3.14. Probability of loss for multiplexed FGN, FBM and F-ARIMA sources, N= 1, 2, 4 and 8, for H = 0.7.

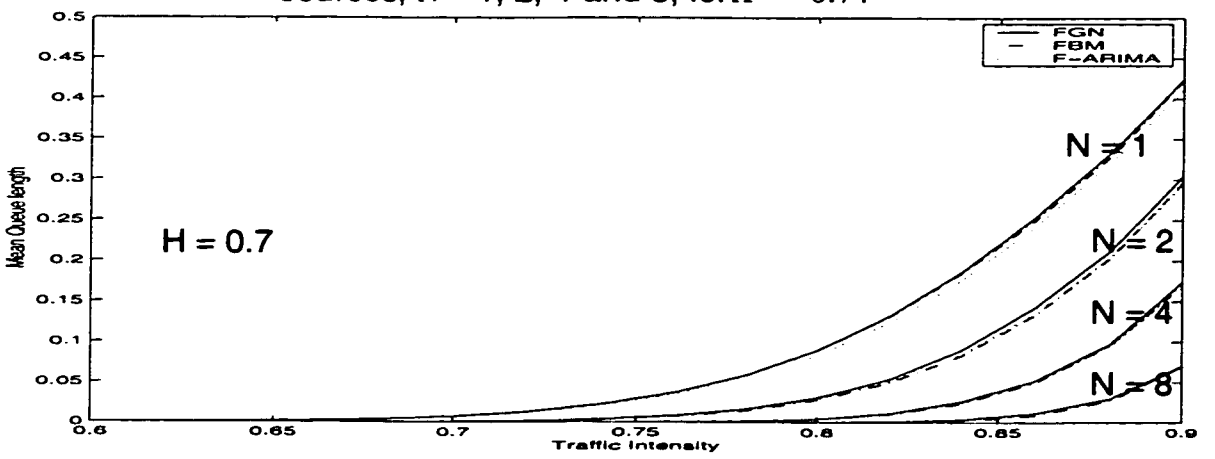

FIGURE.3.15. Mean queue length for multiplexed FGN, FBM and F-ARIMA sources, N= 1, 2, 4 and 8, for H = 0.7.

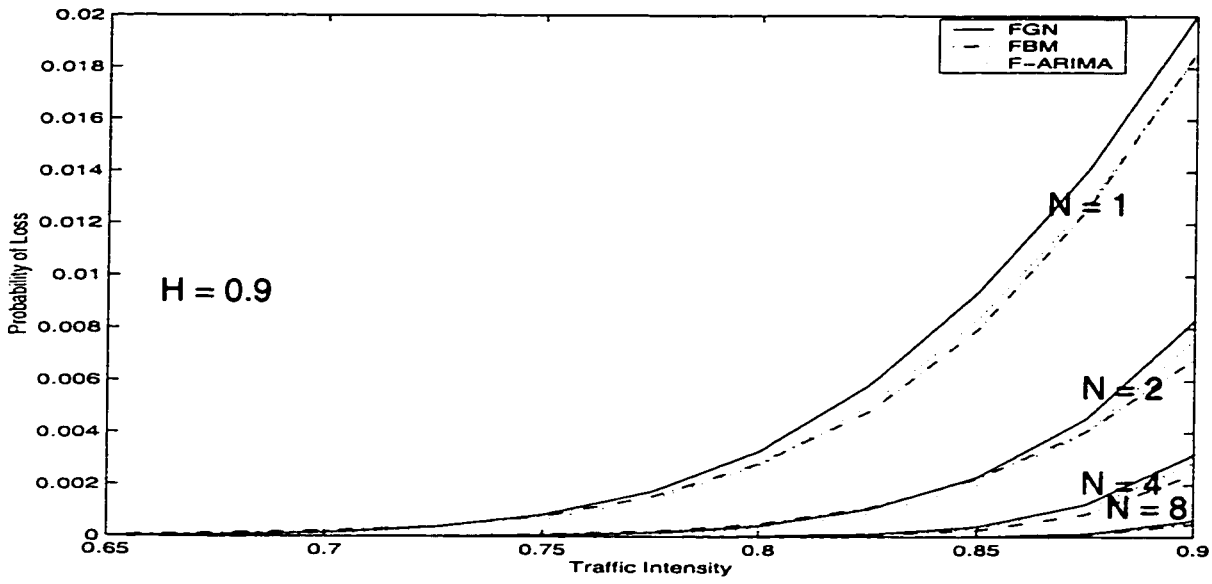

FIGURE.3.16. Probability of loss for multiplexed FGN, FBM and F-ARIMA sources, N= 1, 2, 4 and 8, for $H = 0.9$

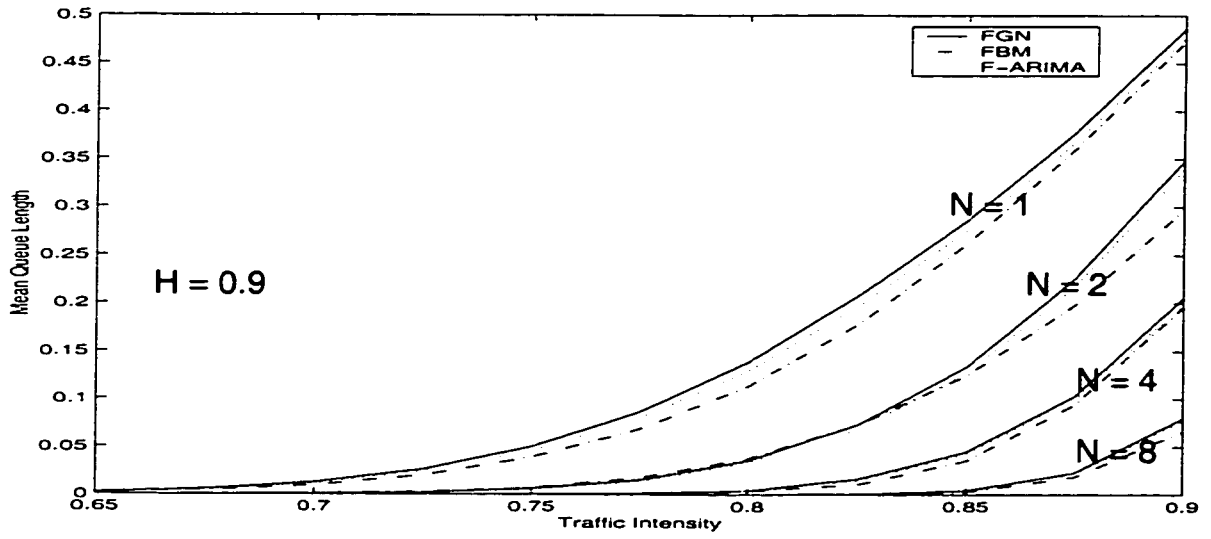

FIGURE.3.17. Mean queue length for multiplexed FGN, FBM and F-ARIMA sources, N= 1, 2, 4 and 8, for $H = 0.9$

Figure 3.16 and 3.17 shows the probability of loss and mean queue length for FGN, FBM and F-ARIMA for high Hurst parameter of value 0.9. As expected, higher values of Hurst parameter will reduce the accuracy of the prediction, espe-

cially when the load is high, as compared to the case of low or medium Hurst parameter that we show in figure 3.14 and figure 3.15.

We have the following interesting results. First, as the correlation index increases, the proability of loss and the mean queue length becomes larger; and second, as the number of multiplexed sources increase, the correlation in the traffic reduced, which has the impact of having smoother traffic and this will lead to reduction in both the probability of loss and the mean queue length. Another important fact that comes out because of multiplexing is that, in some cases, very highly correlated traffic is very difficult to model using the models that are available to us. However, multiplexing several sources together, which is the case in practice, will smooth the traffic and this will make it possible to model this kind of traffic. The results we obtained in this section will be used throughout the presentation. In the next chapter we will compare the QoS of the simulation models along the analytical models to the real data.

3.6 The OPNET model

In this section we present a short review for the OPNET modeler that we will use in generating different traffic such as MMPP, PMPP and ON-OFF sources, which will be introduced in the following chapters. This generated traffic will be used to compare some traffic statistical indices and performance measures with that of the real data. OPNET is a modeling and simulation tool [MIL31] that provides an environment for analysis of communication networks. The tool provides a three layer modeling hierarchy. The highest layer is called the network domain, which allows the definition of the network topologies. The second layer known as the node domain, which allows definition of node architectures (data flow within the node). The third layer is the process domain, which specifies logic or control flow among components in the form of a finite state machine. Figure 3.18 shows OPNET phases of the modelling and simulation cycle.

Typical simulation studies of ATM networks involve either the use of a Poisson process, an ON-OFF model, or an MMPP to generate the network traffic.While these traffic models vary in many fundamental aspects, they rely on different assumptions for determining essential parameters.

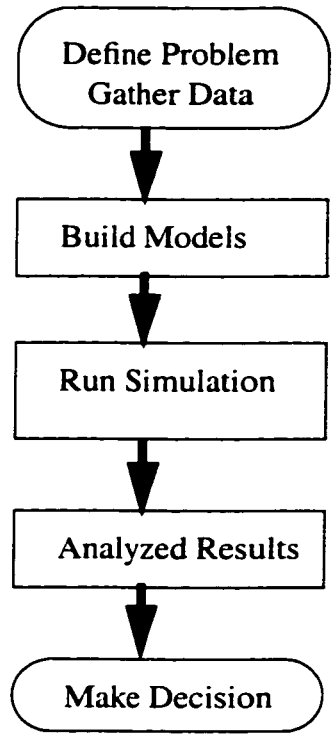

FIGURE.3.18. OPNET phases of the modelling and simulation cycle.

We will use OPNET to generate different traffic data based on the estimated parameters of the models. The output generated data is fed into an ATM multiplexer to find out statistical indices and performance measures; i.e., the IDC, covariance, probability of loss and mean queue length.

3.7 Discussion

In this chapter we presented the most common analytically tractable models, Poisson processes, Bernoulli processes, Markov chain models, MMPP and fluid flow models, that are used to model traffic in ATM networks. We also discussed simulation models that exhibits long-range dependent behavior such as autoregressive processes, PMPP, FGN, FBM and F-ARIMA. We presented the genera-

tion of long-range dependent stochastic traffic using the simulation models, FGN, FBM and F-ARIMA. We plane to asses the effectiveness of each of the models.We presented FGN, FBM and F-ARIMA traces with input $H = 0.5, 0.6, 0.7, 0.8$ and 0.9. Comparison of the autcorrelations, IDC's and variances for the three models shows their self-similar behavior. The covariance decays less exponentially, the IDC increases monotonically and the variance decays more slowly than the inverse of the size. As the correlation between samples increases, the slope of the IDC increases which affects both the value of the slope of the IDC and the variance. As the samples become more correlated, the time constant of the covariance increases, which affects the rate of the decay of the covariance function. Moreover, we presented how the probability of loss and the mean queue length for FGN, FBM and F-ARIMA reacts to the variation of the Hurst parameter. As the Hurst parameter increases, the probability of loss and the mean queue length increases. Moreover, increasing the number of sources to be multiplexed, will reduce the correlation in the stream and this will reduce the probability of loss and the mean queue length. As the traffic intensity and the index of the correlation increases, the three simulation models, FGN, FBM and F-ARIMA becomes slightly different in their prediction of the probability of loss and mean queue length. For low traffic intensity and low correlation index, the three models behave almost the same. We ended this chapter by giving a brief introduction to the OPNET simulation package that will be used to generate traffic of different models throughout the next chapters. In the next chapter, we will present a comparison of the simulation and analytical techniques with the traces from the real data.

CHAPTER IV

Modeling of Ethernet and VBR Video data

The source modelling of Ethernet and VBR video is an active area of research. This is because Ethernet and video services are forecast to be a large portion of the traffic on emerging broadband digital networks. There are several issues that must be solved for broadband digital networks. Two main issues are the problem of deciding whether a new connection can be admitted into an ATM network and the determination of the bandwidth that must be allocated to the connection to ensure adequate QoS, which will be discussed in chapter 7. This requires the finding of models that accurately characterize the traffic flows through ATM networks.

In this chapter we will consider how to model real Ethernet data and VBR video data as an MMPP and PMPP. We also consider how simulation models such as FGN, FBM and F-ARIMA fit the real data. We calculate and compare the index of dispersion for counts, covariances, probability of loss, mean queue length of the Ethernet and VBR video data with that of the models.

4.1 Introduction

In section 3.2.1.4, we have presented a brief introduction to the MMPP processes. We have shown that for this kind of model, we need to know the mean duration of each state, and the Poisson rate in each state. Knowing these parameters will completely lead to the characterization of the traffic.

As has been observed, it is the variability of the variance that makes the process deviate significantly from that of Poisson random process [LEL91]. A promising approach to characterize the variability of the arrival process is to approximate the superposition non-renewal point process by a renewal process characterized by the indices of dispersion for intervals and counts.

The index of dispersion for counts (IDC) of an MMPP increases with the observation time but eventually approaches a fixed value. However, as we have discussed, recent measured studies, indicates that data traffic is self-similar. Its IDC is monotonically increasing with the observation time. However, as we will show in our study that, the MMPP can adequately represent the burstiness of the real video data when the correlation index is not large. For this reason and the simplicity of the analysis, we will restrict our self to the two-state MMPP. In section 3.2.2.2 we also introduced the PMPP process, which is similar to the MMPP, but the duration of the transition from one state to the other is Pareto. In the following, we show how to model Ethernet and video data. We fit the two state MMPP, PMPP, FGN, FBM and F-ARIMA* to the Ethernet and video data. We show that Ethernet can be modeled as FGN, FBM, F-ARIMA and PMPP but not MMPP. On the other hand, medium correlated video traces are modeled as MMPP.

4.2 Fitting a two-state MMPP, two-state PMPP, FGN and FBM to the Bellcore data

In this section we consider the statistical analysis for the real Ethernet data. In the analysis, we compare the MMPP, and PMPP, FGN and FBM generated traffic process with a trace of actual Bellcore Ethernet LAN traffic data from October 1989 and August 1989 discussed in chapter 2. We calculate and compare statistical and performance measures of the generated traffic and the real data.

We choose to consider the two LAN traces, pOct.TL and pAug.TL shown in Table 2.1 of chapter 2. We have estimated the packets counts for the traces pOct.TL and pAug.TL in slots of size 10 ms for LAN interconnection [GUS91,HEY96] resulting in estimated H parameter of 0.781 and 0.80 respectively as shown in Table 4.1. We have found that estimating the packets over dif-

*We chose to consider FGN and FBM only since F-ARIMA has similar results.

ferent intervals (10 ms, 25 ms, 50 ms,...) does not change the slope of the IDC and therefore the value of H. However, the value of the IDC itself becomes larger as the interval over which we make the estimation becomes larger. Moreover, in Table 4.1 we show our estimation over an intervals of 10 ms each, for the peak arrival rate, mean arrival rate, ratio of peak arrival rate to mean arrival rate, variance, the autocorrelation coefficient and the Hurst parameter. These statistics are presented here to give an idea of the LAN traffic averaged over 10 ms intervals and see how it differs from the original LAN traces given in Table 2.2. In section 4.2.1 and section 4.2.2 we present, respectively, the procedures that are used to estimate MMPP and PMPP parameters. In section 4.2.3, the autocorrelation, and IDC for MMPP and PMPP, FGN and FBM are presented. The performance measures of the four traffic models are presented in section 4.2.4.

TABLE 4. 1 Estimated peak arrival rate, mean arrival rate, ratio of peak arrival rate to the mean arrival rate, variance, autocorrelation coefficient and Hurst parameter for pOct.TL and for pAug.TL traces averaged over 10 ms intervals.

Parameter	Peak arrival rate (packet/ 10ms)	Mean arrival rate (packet /10 ms)	Peak arrival rate / Mean arrival rate	Variance	Autocorrelation coefficient	Hurst Parameter
pOct.TL	20	4.876	4.101	13.460	0.765	0.781
pAug.TL	20	3.414	5.858	7.361	0.838	0.80

4.2.1 Estimation of the two state MMPP parameters

The MMPP process was discussed in section 3.2.1.4, where an aggregate arrival process is characterized by two alternating states. When the Markov chain is in state i, $(i = 1, 2)$ the arrival process is Poisson with rate λ_i, and the transition rate of going out of state i is given by σ_i. The IDC for the two state MMPP is derived by Heffes and Lucantoni [HEF86] and it is given by:

$$I_t = 1 + \frac{2\sigma_1\sigma_2(\lambda_1 - \lambda_2)^2}{(\sigma_1 + \sigma_2)^2(\lambda_1\sigma_2 + \lambda_2\sigma_1)} - \frac{2\sigma_1\sigma_2(\lambda_1 - \lambda_2)^2}{(\sigma_1 + \sigma_2)^3(\lambda_1\sigma_2 + \lambda_2\sigma_1)t}\left(1 - e^{-(\sigma_1 + \sigma_2)t}\right) \quad (4.1).$$

The mean arrival rate λ is given by:

$$\lambda = \frac{\lambda_1\sigma_2 + \lambda_2\sigma_1}{\sigma_1 + \sigma_2} \quad (4.2).$$

In the following, we show how to find the four parameters that are needed to characterize the MMPP model. First, we estimate the mean λ, variance ν and the third moment μ_3 of the arrival rate for the two state MMPP from the video data.

The four parameters are given below through equations 4.3-4.6 [HEF80],

$$\sigma_1 = \frac{1}{\tau_c(1 + \zeta)} \quad (4.3).$$

$$\sigma_2 = \frac{\eta}{\tau_c(1 + \zeta)} \quad (4.4).$$

$$\lambda_1 = \lambda + \sqrt{\frac{\upsilon}{\zeta}} \quad (4.5).$$

and

$$\lambda_2 = \lambda - \sqrt{\upsilon\zeta} \quad (4.6).$$

where

$$\zeta = 1 + \frac{\delta}{2}[\delta - \sqrt{4 + \delta^2}] \quad (4.7).$$

and

$$\delta = \frac{\mu_3 - 3\lambda\nu - \lambda^3}{\nu^{1.5}} \quad (4.8).$$

The quantity δ corresponds to a measure of skewness, and τ_c is the time constant and equal to the reciprocal of the sum of σ_1 and σ_2.

The covariance function is given by:

$$C(t) = ve^{-\frac{t}{\tau_c}} \tag{4.9}.$$

The discrete covariance $C(n)$ is obtained by sampling equation (4.9) at intervals of $\frac{n}{\tau_c}$, which has the following form:

$$C(n) = ve^{-\frac{n}{\tau_c}} \tag{4.10}.$$

Let $n(T)$ denote the number of arrivals over time interval of duration T. The IDC at infinity is given by:

$$\lim_{T\to\infty} \frac{Var[n(T)]}{E[n(T)]} = I_\infty = 1 + 2\frac{\nu}{\lambda}\tau_c \tag{4.11}.$$

Thus for a given time constant the variability of the number of arrivals is directly related to the variability of the arrival rate. Also, for a given arrival rate variability, a longer time constant simply implies more variability in the number of arrivals.

Our procedure for finding the two state MMPP parameters λ_1, λ_2, σ_1 and σ_2 is as follows:

1. Estimate the mean λ, variance υ, third moment μ_3 and I_∞ of the arrival rate from the real data.

2. Given I_∞, λ and ν, find τ_c using (4.11). Calculate ζ and δ using (4.7) and (4.8).

3. Find the MMPP parameters $\sigma_1, \sigma_2, \lambda_1$ and λ_2 using equations (4.3)-(4.6).

Table 4.2 shows, along with some statistical values, the resulting two state MMPP parameters (packets are averaged over intervals of 10 ms). That is, all of them rates with dimensions $[10\ ms]^{-1}$

TABLE 4. 2 Estimated values of pOct.TL and pAug.TL for two stage MMPP parameters (all of them rates with dimension $[10\ \mathrm{ms}]^{-1}$)

Parameter	$(\lambda$	(ν)	μ_3	I_∞	τ_c	σ_1	σ_2	λ_1	λ_2
pOct.TL	4.876	13.46	342.79	254.3	45.9	0.012	0.010	8.86	1.49
pAug.TL	3.414	7.36	132.70	209.3	48.3	0.015	0.006	7.56	1.64

4.2.2 Estimation of the two state PMPP parameters

The Pareto-modulated Poisson process (PMPP), which was presented in section 3.2.2.2, can be used to characterize the self-similar traffic [SUB95]. As in the MMPP case, the traffic is modeled as a Poisson process with rate λ_i $i = 1, 2$. However, the duration of each state is independent and identically distributed with Pareto distribution of parameter α (symmetric case), which denotes the thickness of the tail of the distribution [ARN85]. The relationship between α and H is given by [COX84]:

$$H = (3-\alpha)/2. \tag{4.12}$$

For the symmetric case of PMPP, the IDC is given by [SUB95]:

$$I_t = 1 + \frac{(\lambda_1-\lambda_2)^2}{\lambda_1+\lambda_2}\frac{(\alpha-1)}{\alpha}\cdot t^{2-\alpha} \tag{4.13}$$

The mean arrival rate λ is given by:

$$\lambda = \frac{\lambda_1+\lambda_2}{2} \tag{4.14}$$

Our procedure for finding the PMPP parameters is as follows:

i) Estimate the mean arrival rate λ, and the IDC for the real data.

ii) From the IDC time plot, estimate the Hurst parameter H, from which α can be found using (4.12).

iii) From the estimate IDC curve of the real data, choose a value I_t, the IDC at time t.

iv) Substitute the estimated value of the mean arrival rate λ in (4.14), and find λ_1 in terms of λ_2 or vice versa.

v) Using (4.13) find the two parameters λ_1, and λ_2.

vi) Compute based on the current values of the parameters, the goodness of fit of the approximation by comparing the estimated IDC with the theoretical one calculated by (4.13).

vii) Repeat *iii, iv, v,* and *vi* until a satisfactory approximation is obtained.

Table 4.3 shows the resulting PMPP parameters for LAN traces, pOct.TL and pAug.TL, all of them rates with dimensions $[10\ ms]^{-1}$ as we have mentioned in section 4.2 [GUS91].

TABLE 4. 3 Estimated values of pOct.TL and pAug.TL for two state PMPP parameters (all of them rates with dimension $[10\ ms]^{-1}$)

Parameter	Mean λ	Hurst parameter H	α	λ_1	λ_2
pOct.TL	4.874	0.781	1.438	8.717	1.030
pAug.TL	3.414	0.80	1.417	6.123	0.705

4.2.3 Autocorrelation and IDC for MMPP, PMPP, FGN and FBM

Given the MMPP and PMPP parameters shown in Table 4.2 and Table 4.3, respectively, we used OPNET to generate the synthetic traffic. MMPP and PMPP traffic are those obtained from matching to the real data, while FGN and FBM are generated from simulation using Matlab software with input to the model, the Hurst parameter, mean value and variance estimated from the traces pOct.TL and pAug.TL all averaged over intervals of 10 ms as given in Table 4.1.

Figure 4.1 and figure 4.2 shows a comparison of the autocorrelation function for the real Ethernet pOct.TL and pAug.TL sequences averaged over intervals of 10 ms with that of the synthetic MMPP, PMPP, FGN and FBM. Simulation models FGN and FBM have good prediction for the autocorrelation function of the Ethernet data. The PMPP prediction of the real data autocorrelation is clear and is not as good as the prediction of the FGN and FBM models. The autocorrelation function of the MMPP is larger than the autocorrelation functions of the real Ethernet traffic. However, it matches the real data for the first few lags.

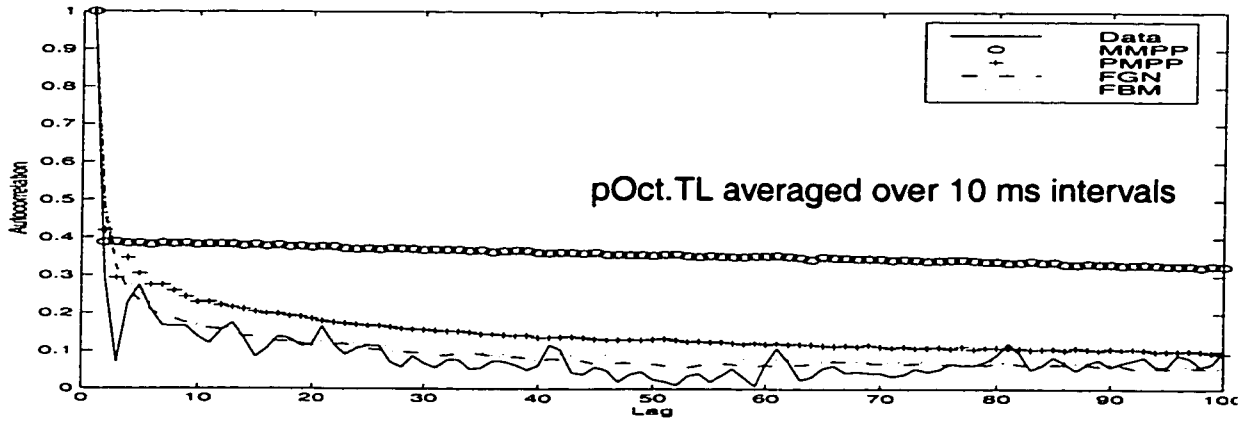

FIGURE.4.1. Autocorrelation function for pOct.TL Ethernet data averaged over 10 ms intervals, MMPP, PMPP, FGN and FBM.

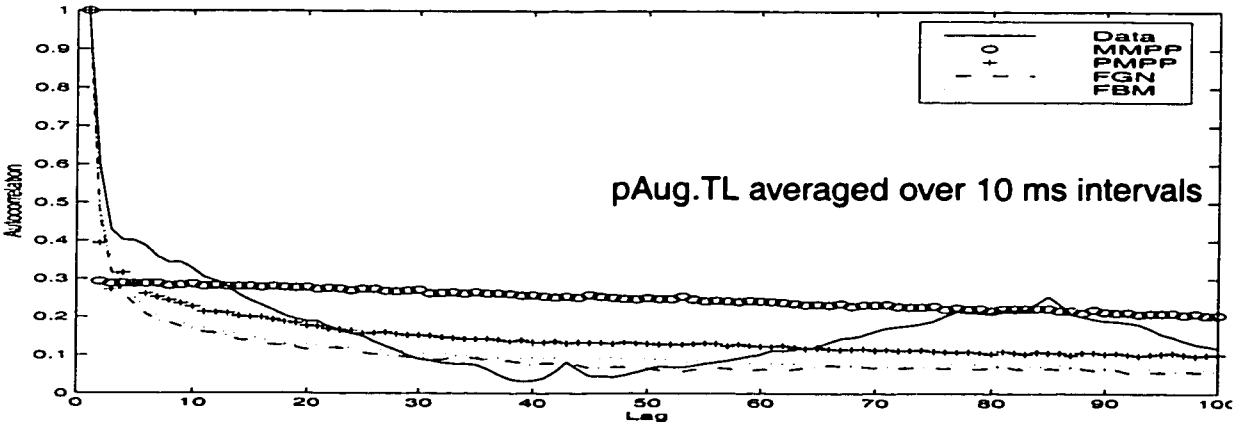

FIGURE.4.2. Autocorrelation function for pAug.TL Ethernet data averaged over 10 ms intervals, MMPP, PMPP, FGN and FBM.

Figure 4.3 and figure 4.4 shows, respectively, the IDC's for pOct.TL and pAug.TL averaged over 10 ms intervals, fitted MMPP, fitted PMPP, FGN and FBM models for values of $H = 0.781$ and 0.80 respectively. It is clear that FBM and FGN as well as PMPP are good predictors of the Ethernet data IDC. However, MMPP overestimates the IDC of the traffic.

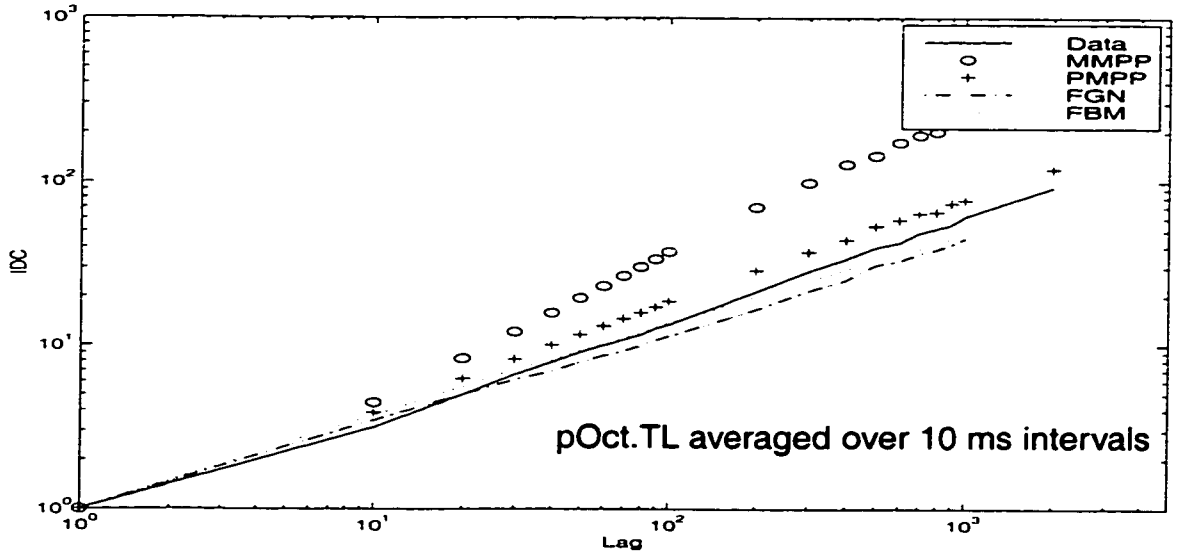

FIGURE.4.3. IDC's of pOct.TL Bellcore data averaged over 10 ms intervals, FGN, FBM, and fitted MMPP, PMPP model, $H = 0.781$.

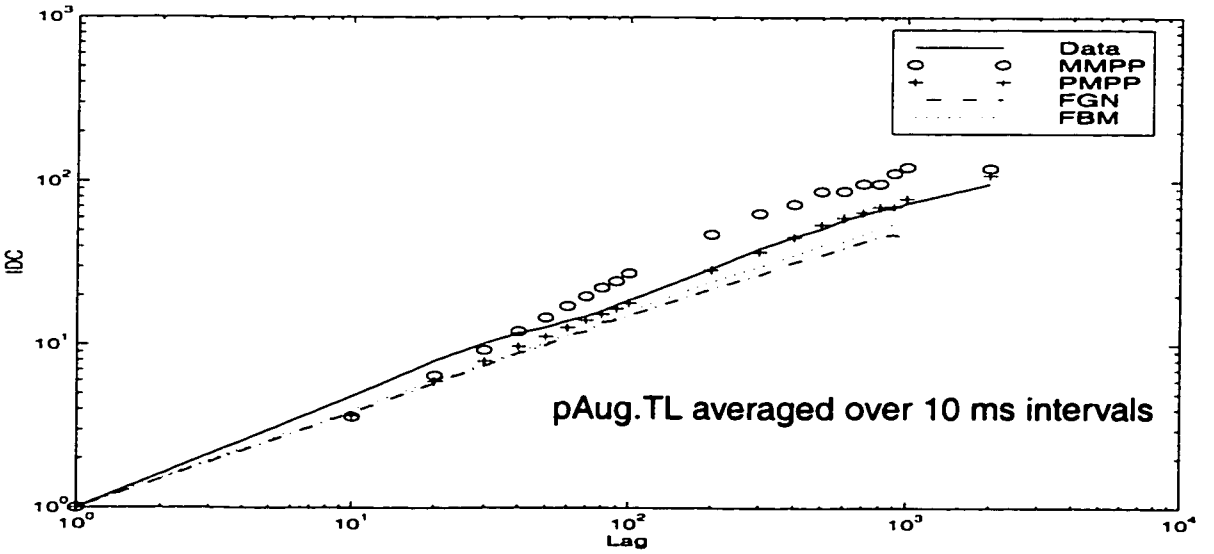

FIGURE.4.4. IDC's of pAug.TL Bellcore data averaged over 10 ms intervals, FGN, FBM, and fitted MMPP, PMPP model, $H = 0.80$.

4.2.4 Performance analysis of the modeled MMPP, PMPP, FGN and FBM for Ethernet data

Given the estimated MMPP and PMPP parameters from the real Ethernet data, OPNET is used to generate synthetic traffic. In addition, we consider how simulation models such as FGN and FBM are tracking the data. We compare the performance measures such as the probability of loss and the mean queue length of the generated MMPP, PMPP, FGN and FBM traffic with the real data.

4.2.4.1 Probability of loss

The MMPP, PMPP, FGN and FBM estimation of the probability of loss for the trace pOct.TL and pAug.TL averaged over 10 ms intervals with a buffer capacity of 10 packets over a large scale of traffic intensity are shown in figure 4.5 and figure 4.6, respectively. The PMPP prediction for the probability of loss of the Ethernet traces is satisfactory for practical engineering design, however, the MMPP is not and has the tendency to underestimate the probability of loss. As shown in both figures, the MMPP does not perform well for the Ethernet data regarding the probability of loss, especially when the traffic intensity is high. FGN and FBM make a good estimation for the probability of loss over the entire range of traffic intensity. The results are expected from those we obtained when we discussed the autocorrelation and IDC in the previous section, where FGN, FBM and PMPP autocorrelation and IDC make a good approximation of the data, but not MMPP.

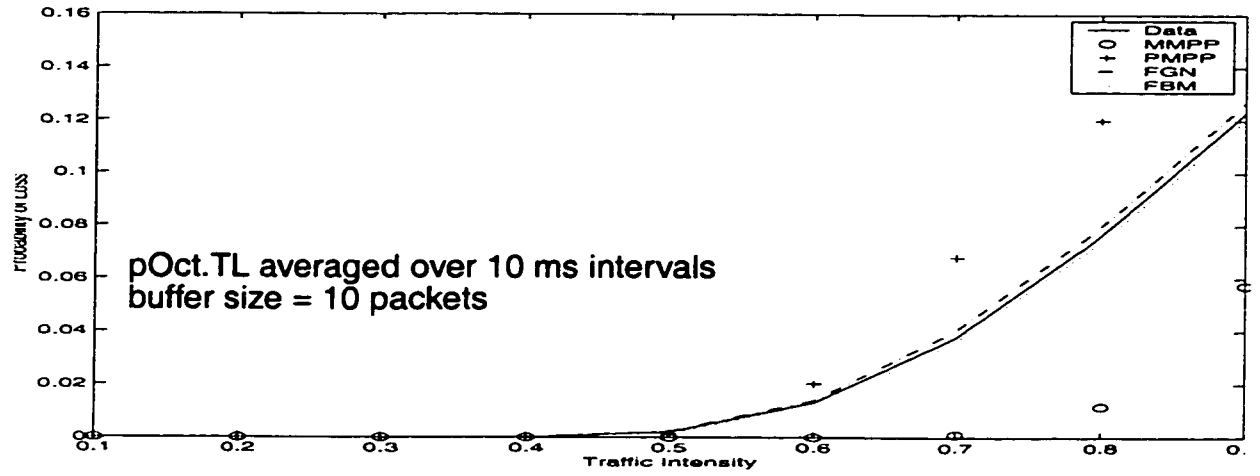

FIGURE.4.5. Probability of loss for the real pOct.TL Bellcore data averaged over 10 ms intervals compared with that using MMPP, PMPP, FGN and FBM models

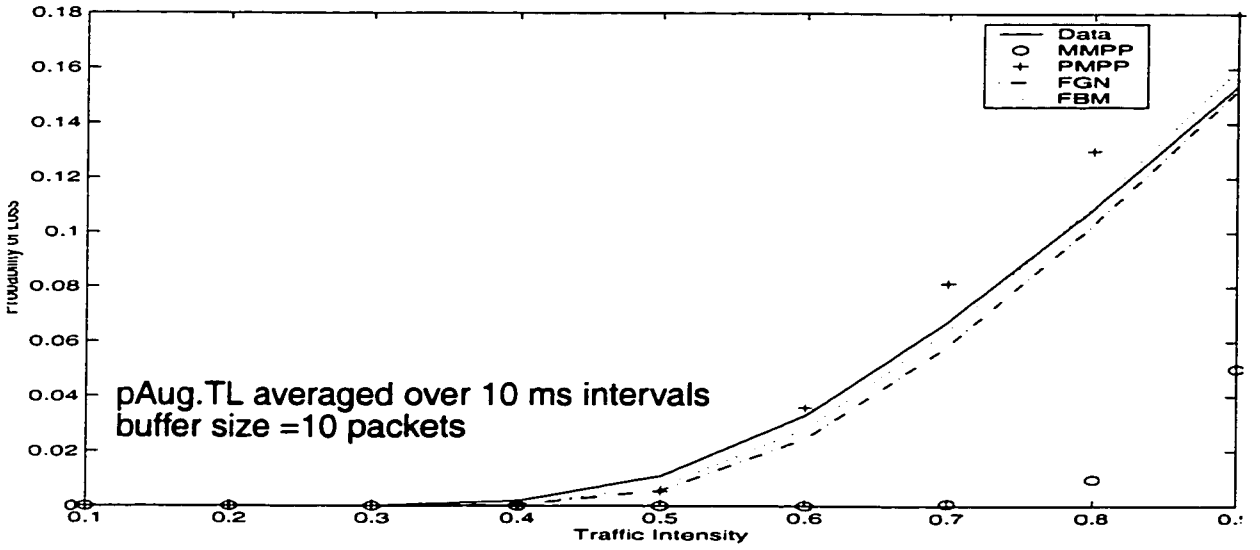

FIGURE.4.6. Probability of loss for the real pAug.TL Bellcore data averaged over 10 ms intervals compared with that using MMPP, PMPP, FGN and FBM models.

4.2.4.2 Mean queue length

The estimation of the mean queue length is now discussed. As for the probability of loss, the mean queue length estimation of the Ethernet traces, pOct.TL and pAug.TL averaged over 10 ms intervals, based on the PMPP model is also good for practical engineering design. See figure 4.7 and figure 4.8. The MMPP

does not preform as well as the PMPP. However, the MMPP prediction of the queue length is much better than that for the probability of loss. This is in agreement with the results that were obtained by Heffes and Lucantoni used to predict queuing delay of a packet speech multiplexer [HEF86]. Moreover, as for the probability of loss, FGN and FBM makes a better prediction for the queue length as compared with PMPP and MMPP.

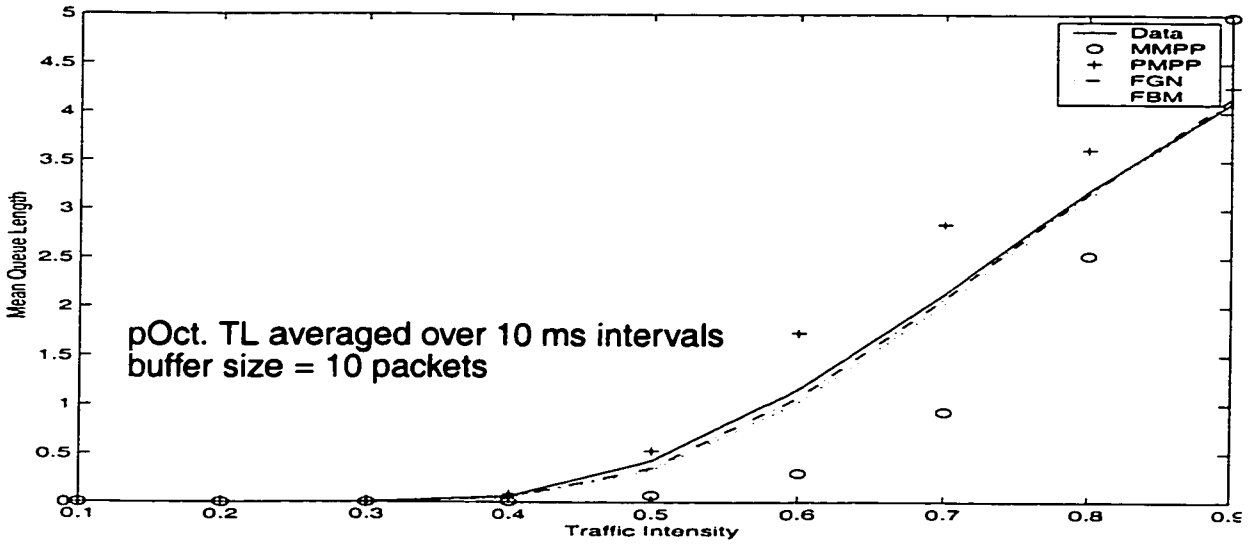

FIGURE.4.7. Mean queue length for the real pOct.TL Bellcore data averaged over 10 ms intervals compared with that using MMPP, PMPP, FGN and FBM models.

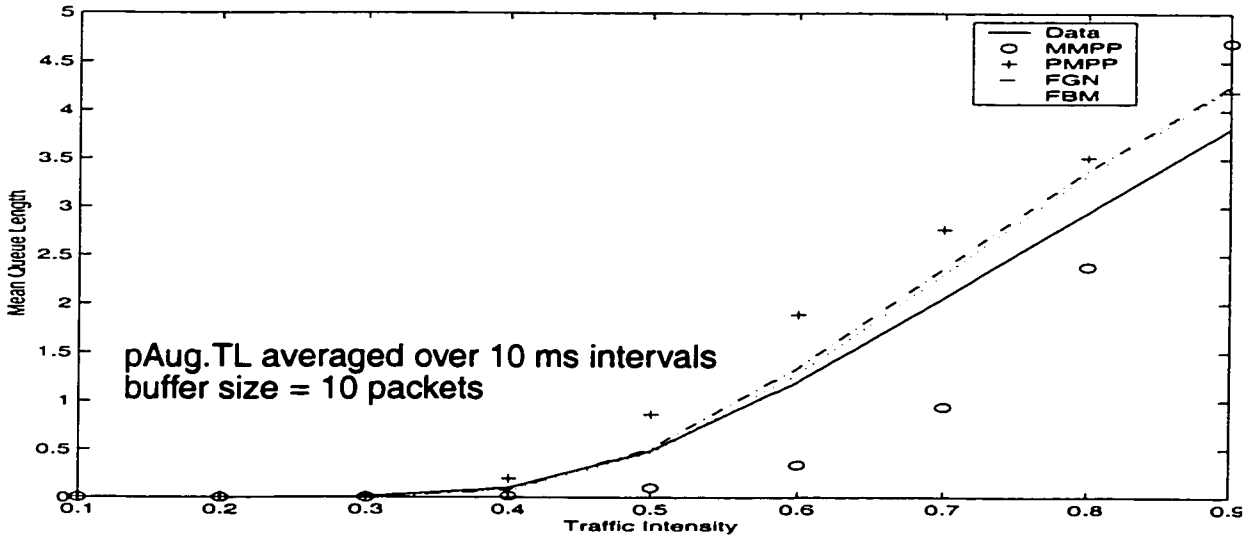

FIGURE.4.8. Mean queue length for the real pAug.TL Bellcore data averaged over 10 ms intervals compared with that using MMPP, PMPP, FGN and FBM models.

4.3 Modeling multiplexed sources

In this section, we show that multiplexing several statistically independent and identical highly correlated sources will result in a good prediction and reduction of the probability of loss and mean queue length. We present multiplexing of the Ethernet data pOct.TL and pAug.TL averaged over 10 ms intervals.

As shown in figure 4.9 and figure 4.10, it is clear how multiplexing improves the accuracy of the matching when several sources are multiplexed. The paths of the multiplexed FGN and FBM traces tracks the path of the real data, which indicates that FGN and FBM are good models for characterizing data traffic. PMPP estimation is not as good as FGN and FBM, however, its prediction is reasonable for practical engineering design. Good improvement is achieved when comparing the results of the probability of loss and the mean queue length obtained when multiplexing 10 sources with that when we have only a single source as shown in figures 4.5 - 4.8. Multiplexing of many sources has the advantages of reducing the probability of loss and the mean queue length, which we discussed in section 3.4.3, as shown below for the Ethernet traces pOct.TL and pAug.TL averaged over 10 ms intervals.

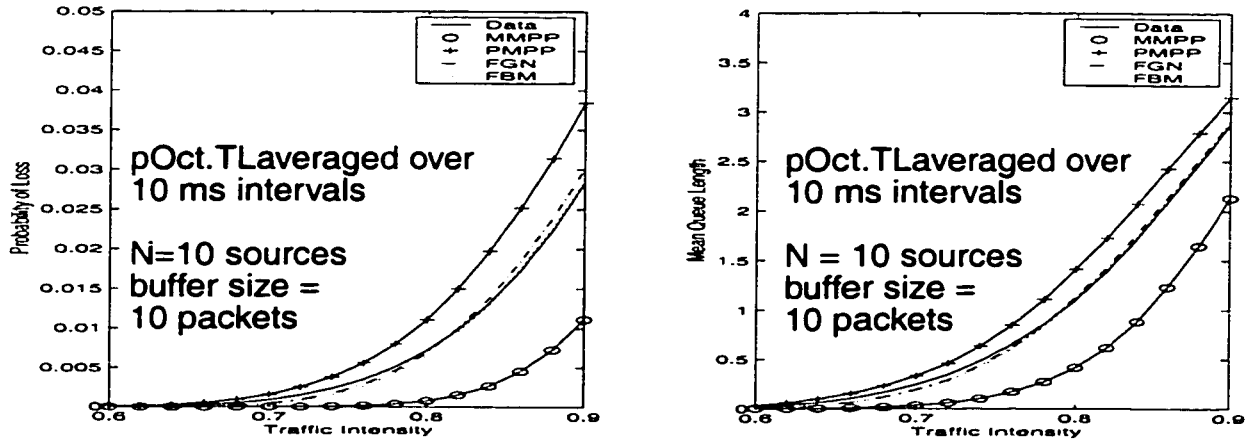

FIGURE.4.9. Comparison of probability of loss and mean queue length for10 multiplexed sources MMPP, PMPP, FGN and FBM with that of the original 10 multiplexed pOct.TL sequence averaged over 10 ms intervals, buffer capacity =10 packets

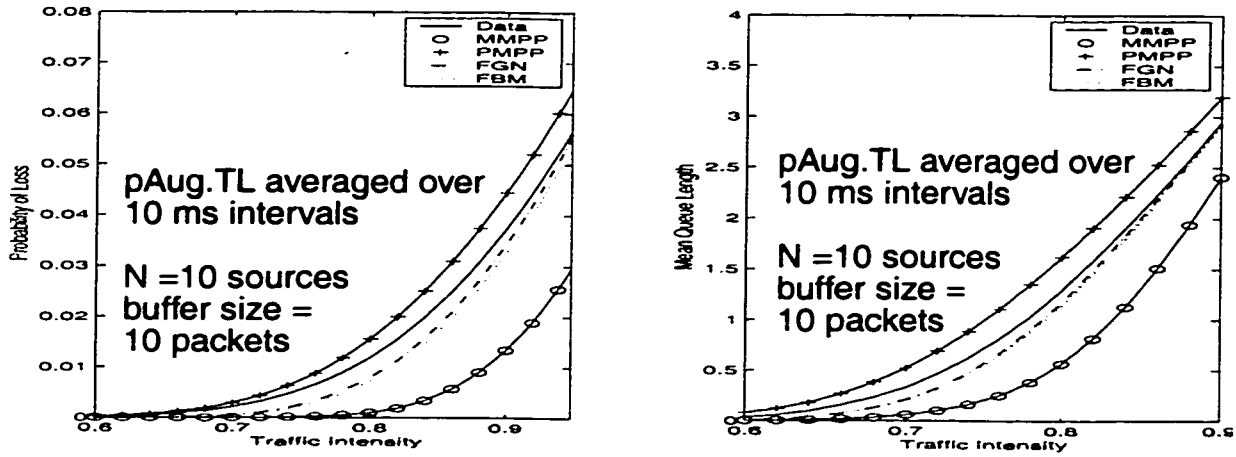

FIGURE.4.10. Comparison of probability of loss and mean queue length for10 multiplexed sources MMPP and PMPP with that of the original 10 multiplexed pAug.TL sequence averaged over 10 ms intervals, buffer capacity =10 packets

4.4 Fitting a two-state MMPP and FBM* to the video data

In section 4.2, we have shown that Ethernet data can be modeled as PMPP but not MMPP. We also showed that simulation models such as FGN, FBM and F-ARIMA are good models in characterizing and modeling Ethernet data. In this section we will present how we can fit an MMPP model to the real video data. However, PMPP, FGN, FBM and F-ARIMA are not good models for characterizing video data, which will be shown below. We use the procedure that was introduced in section 4.2.1 to fit the MMPP model to the real video data. Since the simulation models FGN, FBM and F-ARIMA have similar results as we have seen in section 3.3, we only concentrate on using one of them in modeling video data. We chose the FBM model. Therefore, in this section we concentrate on studying how to model video data as FBM and as an MMPP. The video traces are presented in section 2.3 of chapter 2. The FBM traffic is generated using Matlab software with Hurst parameter H, mean and variance (see Table 2.4) as inputs to the model. For the MMPP case, the mean λ, variance v and the third moment μ_3 of the arrival rate for the two state MMPP are estimated from the video data. Some of

*We chose to consider FBM only since FGN and F-ARIMA have similar results.

statistical values and MMPP parameters for four video traces are shown in Table 4.4.

TABLE 4. 4 Description of some statistical values and MMPP parameters for video data

Video sequence	mean m	variance ν	third moment μ_3	I_∞	τ_c	σ_1	σ_2	λ_1	λ_2
video-conferencing	130.3	5.54e+03	4.898e+06	6071.6	71.4	0.010	0.003	265.5	89.4
video-phone	170.6	1.15e+04	1.254e+07	9751.6	72.5	0.011	0.003	374	114
TV series	5336	1.35e+06	1.754e+011	96365	190	0.004	0.001	7342	4662
Movie	5948	1.56e+06	2.407e+011	13512	257	0.003	0.001	8195	5254

Given the Hurst parameter H, mean and variance of the video traces, we generate FBM using Matlab software. Moreover, given the estimated MMPP parameters σ_1, σ_2, λ_1 and λ_2 of the video traces shown in Table 4.4, we generate the MMPP traffic using OPNET. Then, we compare the covariance, IDC, probability of loss, mean queue length for the FBM, MMPP and for the real video data. In Table 4.4 and according to our estimation of the Hurst parameter shown in Table 2.4 of chapter 2, we see that as the variance to mean ratio increase, the traffic becomes more bursty and this will lead to a larger value of Hurst parameter. Also in Table 4.4, and as we have presented in section 4.2.1, we see that as the traffic becomes more bursty, the time constant τ_c of the processes becomes larger, which means that, the autocorrelation (or covariance) function of the highly correlated traffic will decay more slowly than uncorrelated traffic. This is an important observation, which can be used to see how the traffic is correlated by looking at their time constants. The autocorrelation functions for the video traces are shown in figure 2.14 of chapter 2.

4.4.1 Covariance and IDC

The covariance of generated FBM, MMPP traffic models and that of the real video-conferencing and video-phone video data are shown in figure 4.11 and figure 4.12, respectively. The approximation for the MMPP is very good over a large range of lag. However, the FBM covariance falls far below the covariance of the real data.

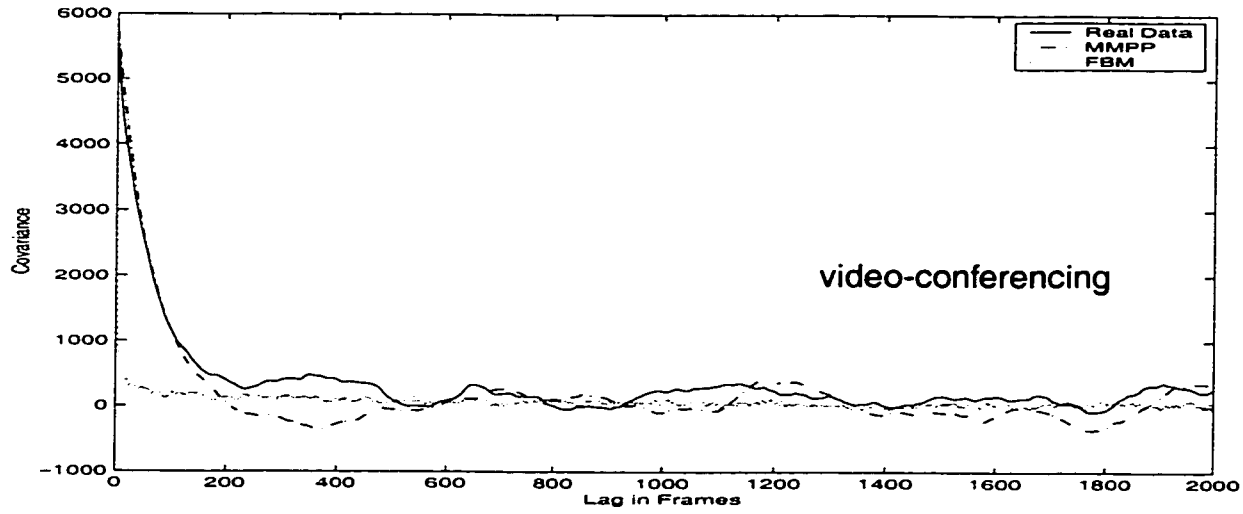

FIGURE.4.11. Covariance functions of FBM and MMPP compared with that of histogram of the original video-conferencing sequence

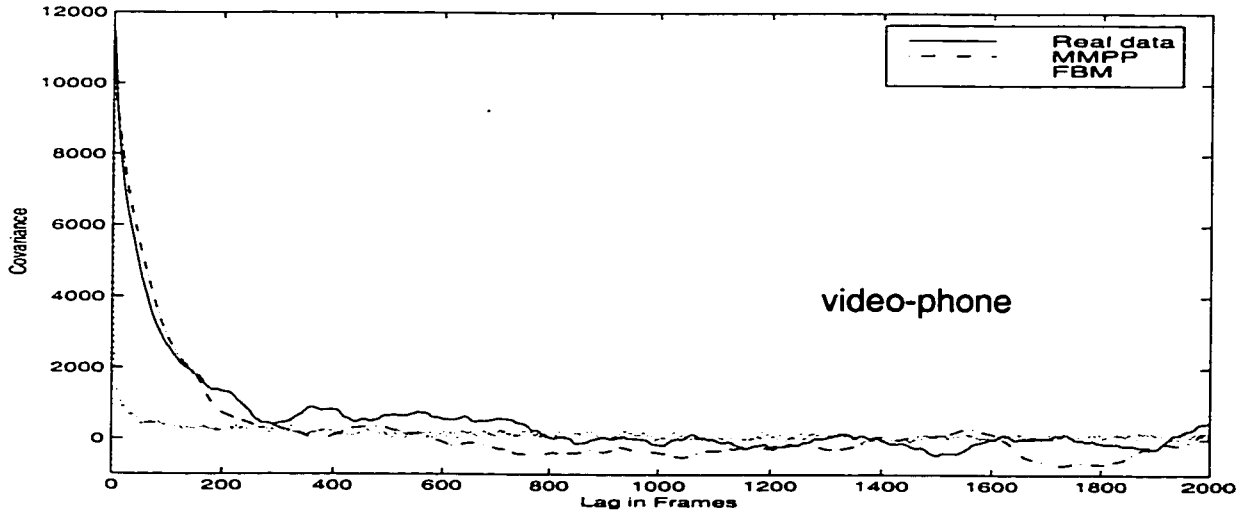

FIGURE.4.12. Covariance functions of FBM and MMPP compared with that of histogram of the original video-phone sequence

For highly correlated traffic such as TV series and Movie and as shown in figure 4.13 and figure 4.14, the covariance of the FBM, also as the case for the tele-

conferencing data, falls far below that of the real data. The MMPP model underestimates the covariance of real data. However, for small lag, the covariance of the generated MMPP traffic accurately predicts the covariance of the highly correlated entertainment traffic.

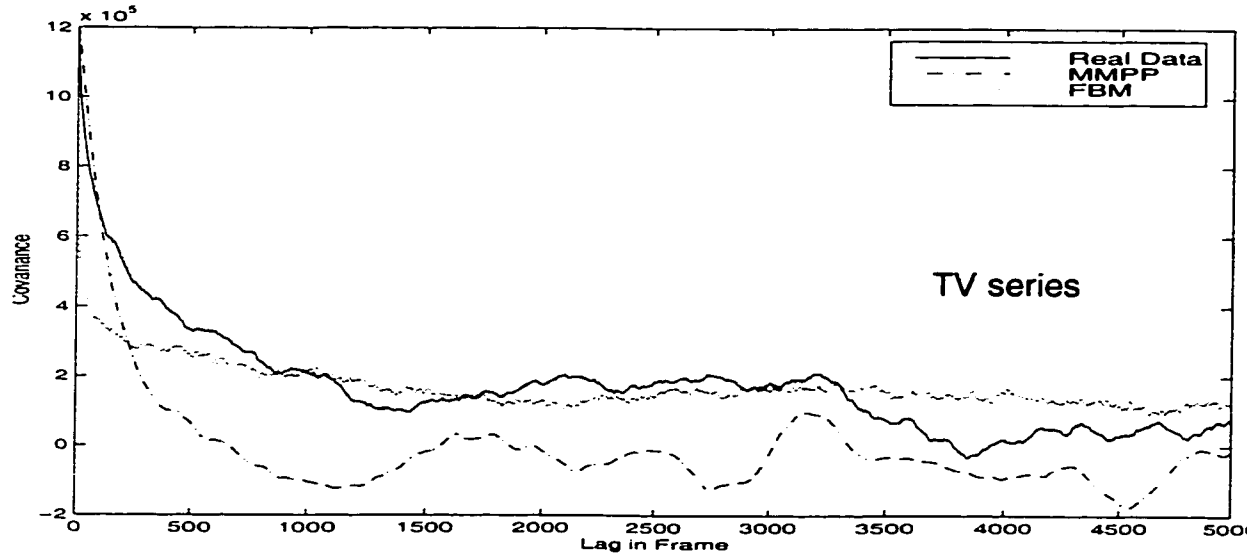

FIGURE.4.13. Covariance functions of FBM and MMPP compared with that of histogram of the original TV series sequence.

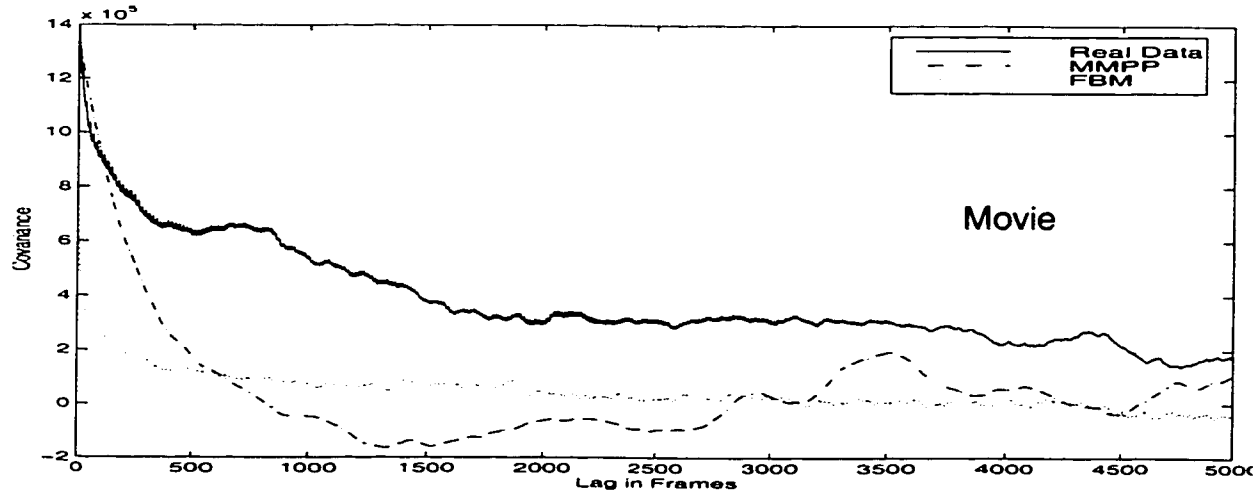

FIGURE.4.14. Covariance functions of FBM and MMPP compared with that of histogram cf the original Movie sequence.

As for the covariance of the video-conferencing and video-phone, the IDC for the generated FBM does not match that of the real data. However, the IDC of the MMPP traffic model and that of the real data video-conferencing and video-phone

data are in good agreement over a large range of frames. This is shown in figure 4.15 and figure 4.16, respectively, for video-conferencing and video-phone.

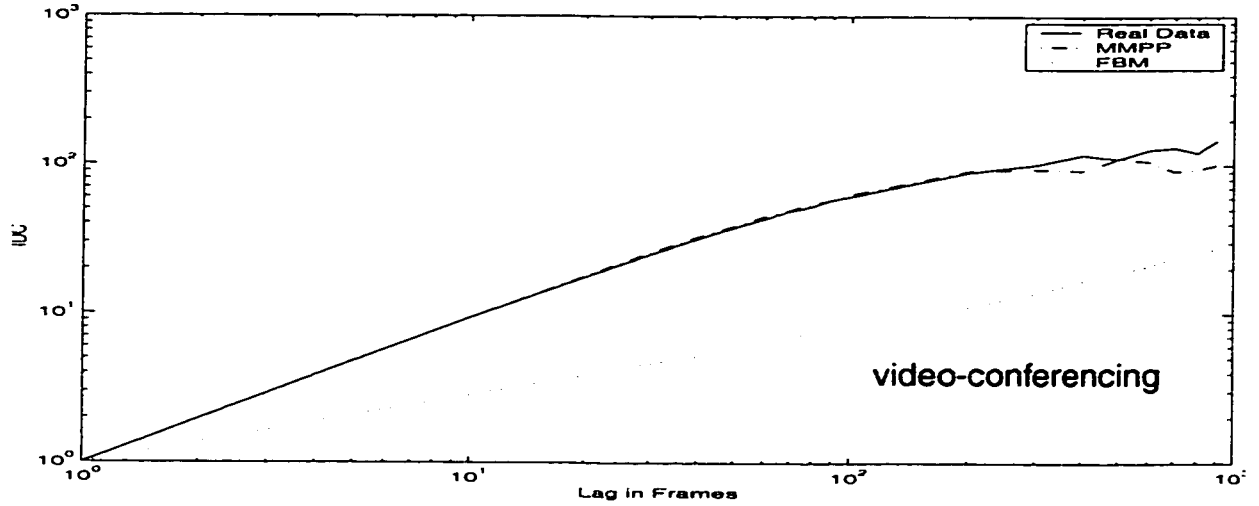

FIGURE.4.15. IDC of FBM and MMPP compared with that of histogram of the original video-conferencing sequence

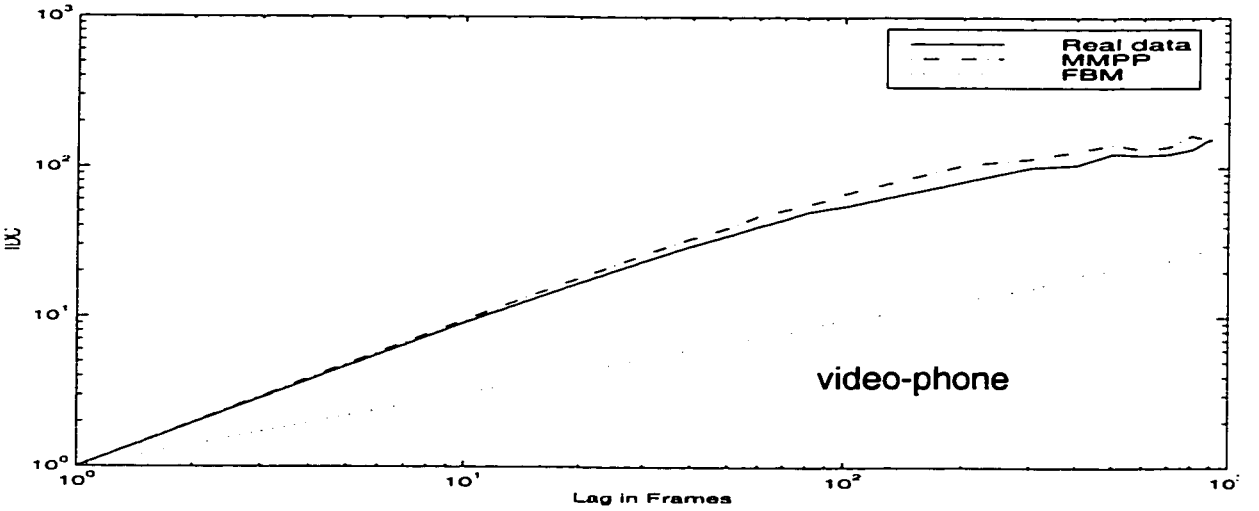

FIGURE.4.16. IDC of FBM and MMPP compared with that of histogram of the original video-phone sequence.

The IDC for the TV series and Movie are shown in figure 4.17 and 4.18 respectively. Generated traffic FBM IDC for TV series and Movie also does not match that of the real data. The IDC of the MMPP slightly overestimates the IDC

of the TV series and Movie data. However, the discrepancy between the model and the data is less than that based on FBM mode.

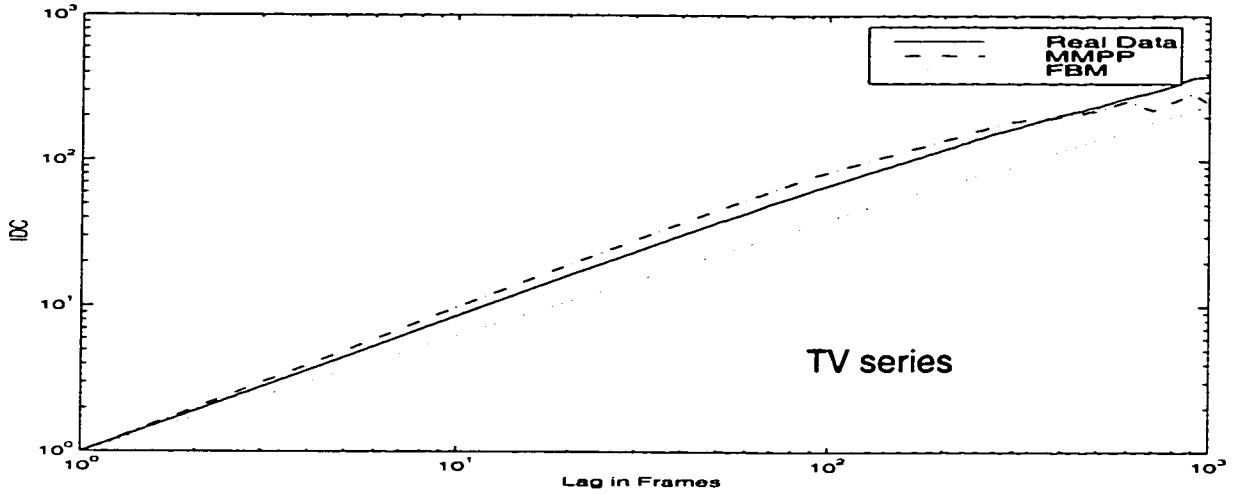

FIGURE.4.17. IDC of FBM and MMPP compared with that of histogram of the original TV series sequence.

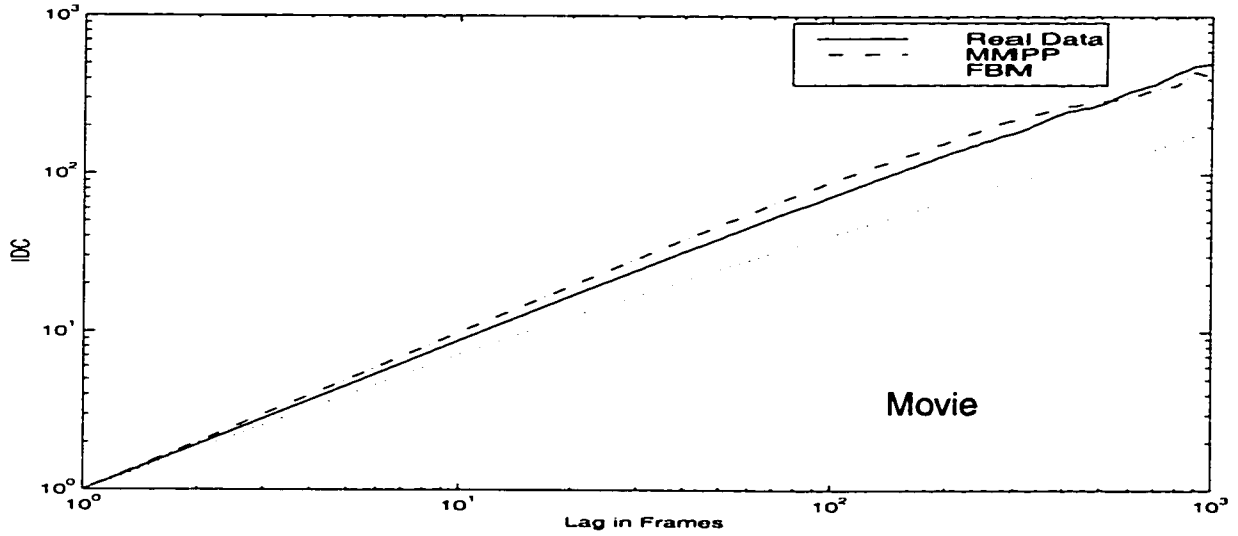

FIGURE.4.18. IDC of FBM and MMPP compared with that of histogram of the original Movie sequence

4.4.2 Performance analysis of the modeled FBM and MMPP video data

Given the Hurst parameter H, mean and variance of the video traces, we generate FBM using Matlab software. Moreover, given the estimated MMPP parameters from the real data, which is shown in Table 4.4, OPNET is used to generate synthetic traffic. We compare the performance measures such as the

probability of loss and the mean queue length of the generated FBM and MMPP traffic and that for the real data.

4.4.2.1 Probability of loss

In our analysis, we consider the probability of loss for the video traces, which we discussed in chapter 2, as a function of the traffic intensity. The buffer capacity is constant. Figure 4.19 and figure 4.20 shows, respectively, the probability of loss for a number of multiplexed video-conferencing and video-phone calls as a function of the traffic intensity with the buffer size treated as a parameter of value 100 cells. The FBM largely underestimates the real traces. The prediction of the MMPP model is good over a large interval of traffic intensity. The probability of loss for the video-phone trace is larger than that of the video-conferencing trace. This is due to the fact that, video-phone trace is more bursty than the video-conferencing trace. As the load increases, the discrepancy between the MMPP model and the real data increases. This is expected since heavy load means the chance of losing cells becomes more. As the number of multiplexed sources increases, the matching between the real data and the MMPP model is more accurate. However, increasing the number of multiplexed sources for the FBM model has no effect on the accuracy of the matching. In figure 4.21 and figure 4.22, we show the probability of loss for a number of multiplexed TV series and Movie sources and the buffer size is 4000 cells. For small number of multiplexed sources, the discrepancy between the model and the data is clear. As the number of multiplexed sources increases, the prediction of the model probability of loss to the TV series and Movie data improves. As for the teleconferencing data, FBM does not do well for entertainment video data such as TV series and Movie. In other words, FBM shows poor prediction for the telconferencing and entertainment data that we con-

sidered. Moreover, multiplexing a number of FBM sources has no effect on improving the matching.

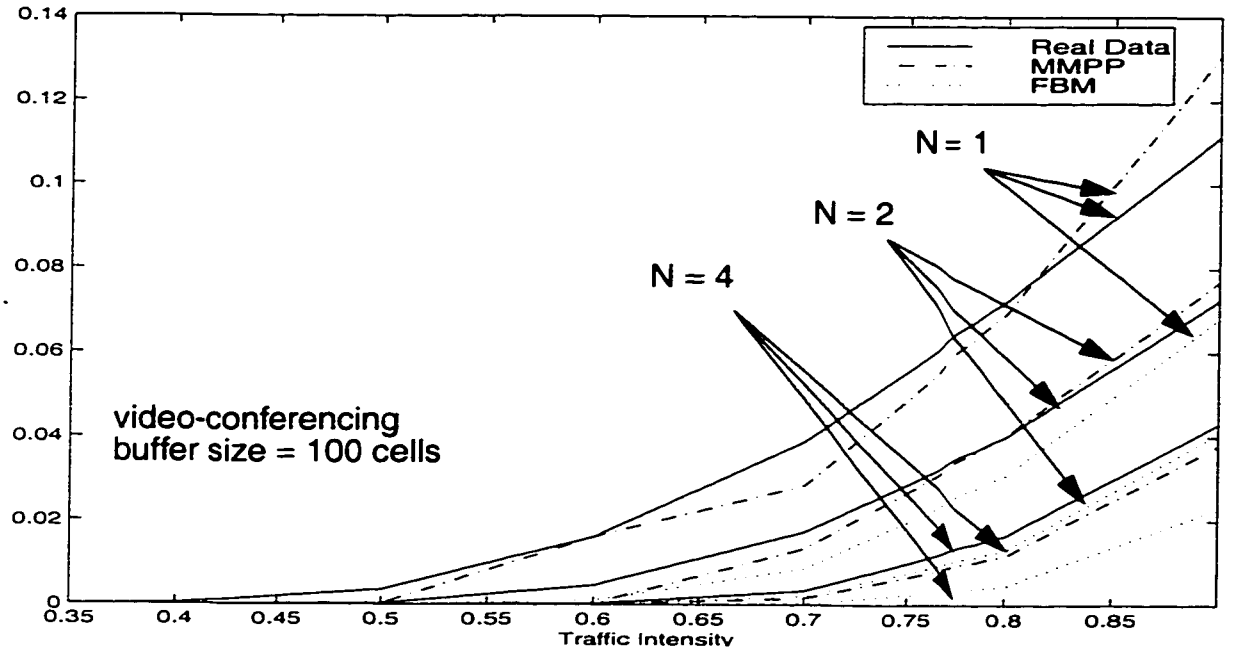

FIGURE.4.19. Comparison of the probability of loss of real video-conferencing data and that of FBM and MMPP for N =1, 2 and 4 multiplexed source

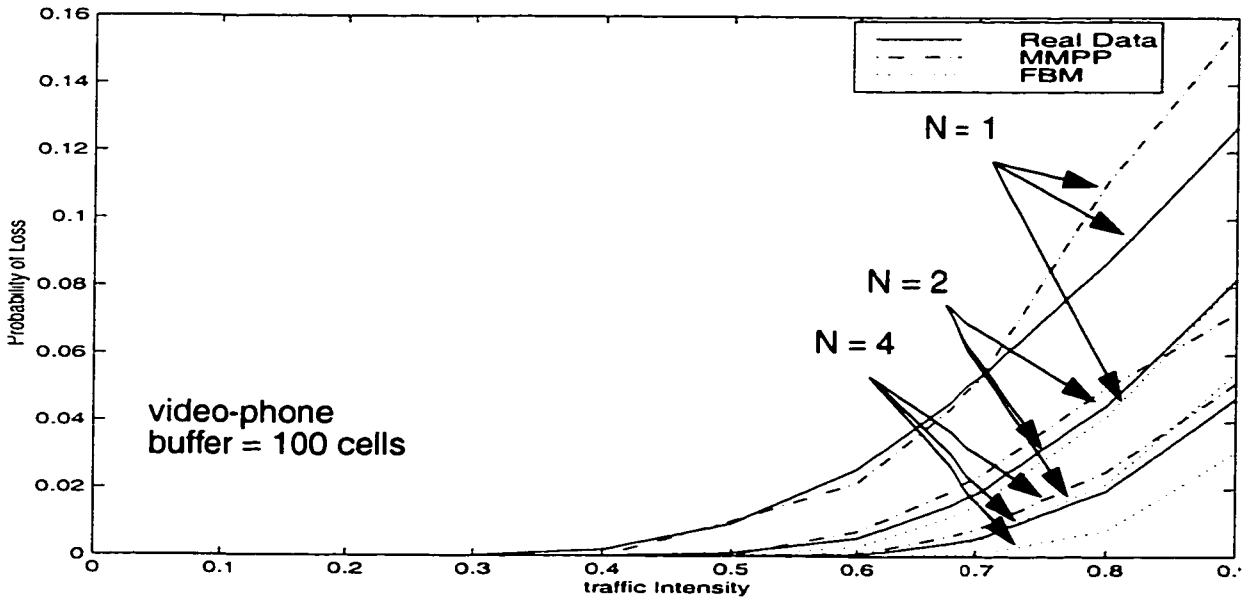

FIGURE.4.20. Comparison of the probability of loss of real video-phone data. and that of FBM and MMPP for N =1, 2 and 4 multiplexed sources

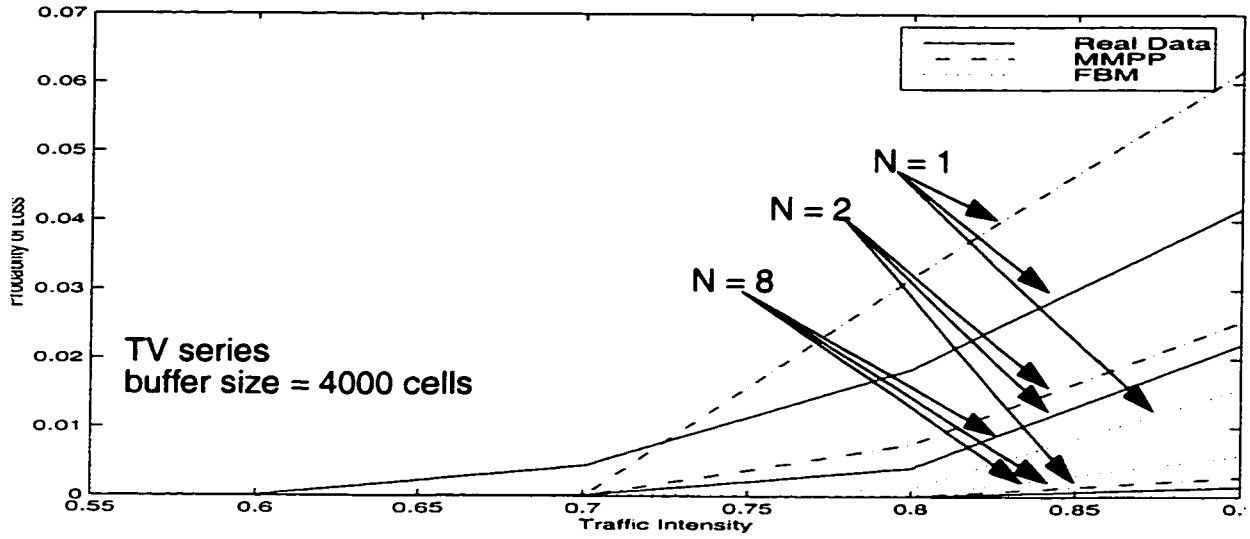

FIGURE.4.21. Comparison of the probability of loss of real video-phone data and that of FBM and MMPP for N =1, 2 and 8 multiplexed sources

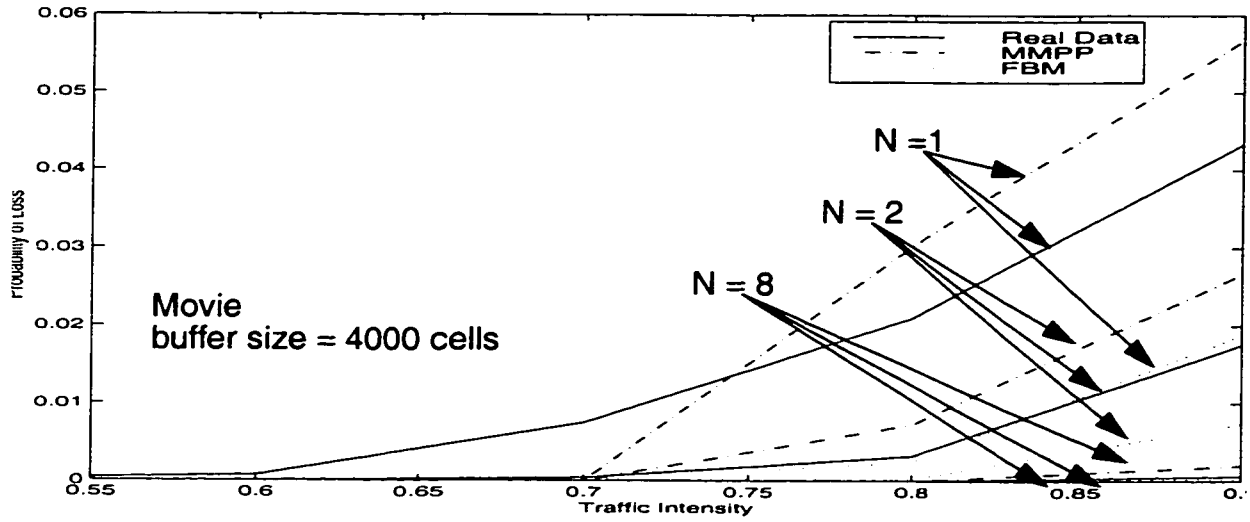

FIGURE.4.22. Comparison of the probability of loss of real Movie data. and that of FBM and MMPP for N =1, 2 and 8 multiplexed sources

4.4.2.2 Mean queue length

The same thing applies to the mean queue length for a number of multiplexed sources as shown in figures 4.23 and figure 4.24, where the queue length is a function of the traffic intensity and the buffer capacity is infinite. The FBM queue length does not match the mean queue length of the real data. However, as can be seen, the mean queue length for the generated MMPP traffic and that of

the real data are in agreement when the load is not large. Moreover, the mean queue length increases as the correlation index increases. Video-phone is more correlated than video-conferencing and as shown video-phone has larger mean queue length than video-conferencing given the same capacity, which is fixed at a very large number for both cases. Figure 4.25 and figure 4.26 show the mean queue length for a number of generated multiplexed sources of FBM and MMPP model and that of the real TV series and Movie video data as a function of the traffic intensity. The buffer size is infinite. The results of the mean queue length for the two entertainment video data based on MMPP and FBM are similar to that of probability of loss.

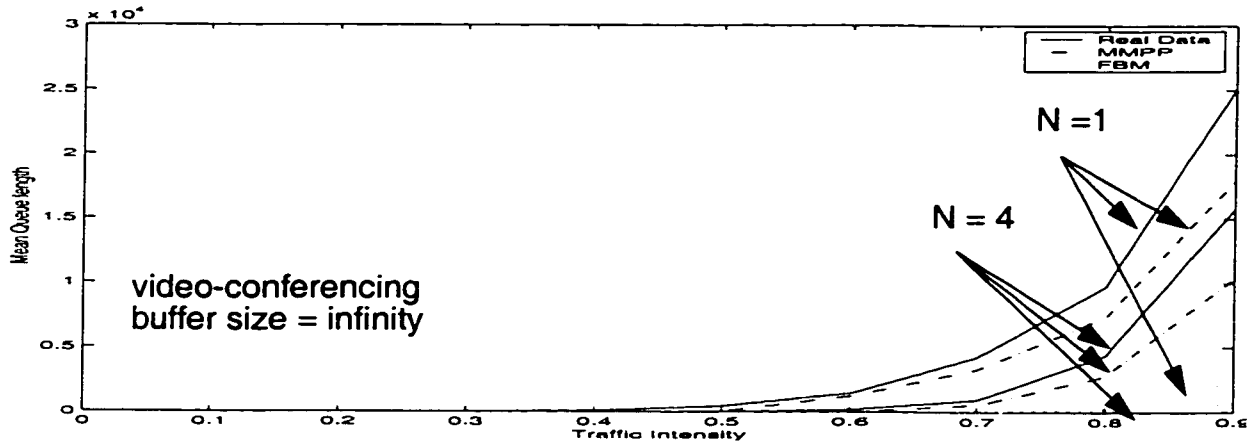

FIGURE.4.23. Comparison of the mean queue length of real video-conferencing data. and that of FBM and MMPP for N =1, 4 multiplexed sources, infinite buffer size.

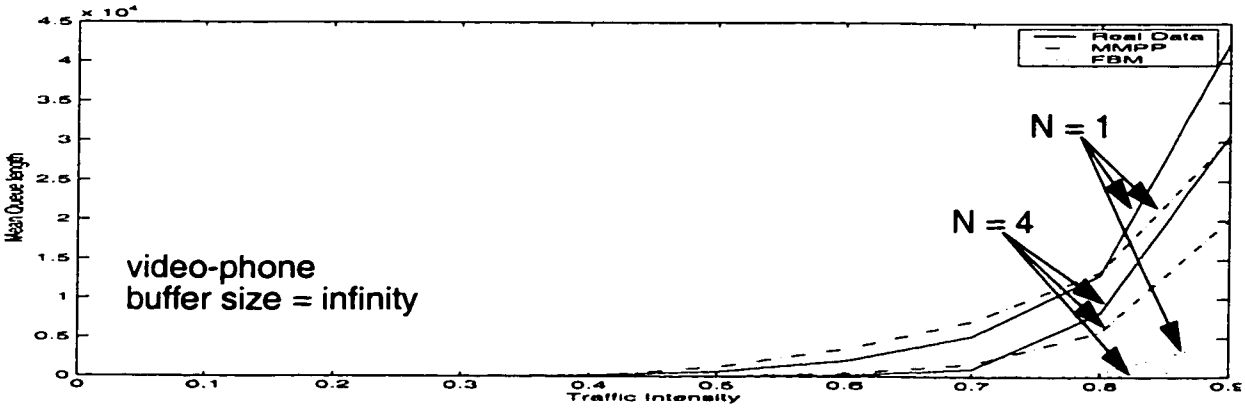

FIGURE.4.24. Comparison of the mean queue length of real video-phone data. and that of FBM and MMPP for N =1, 4 multiplexed sources., infinite buffer size

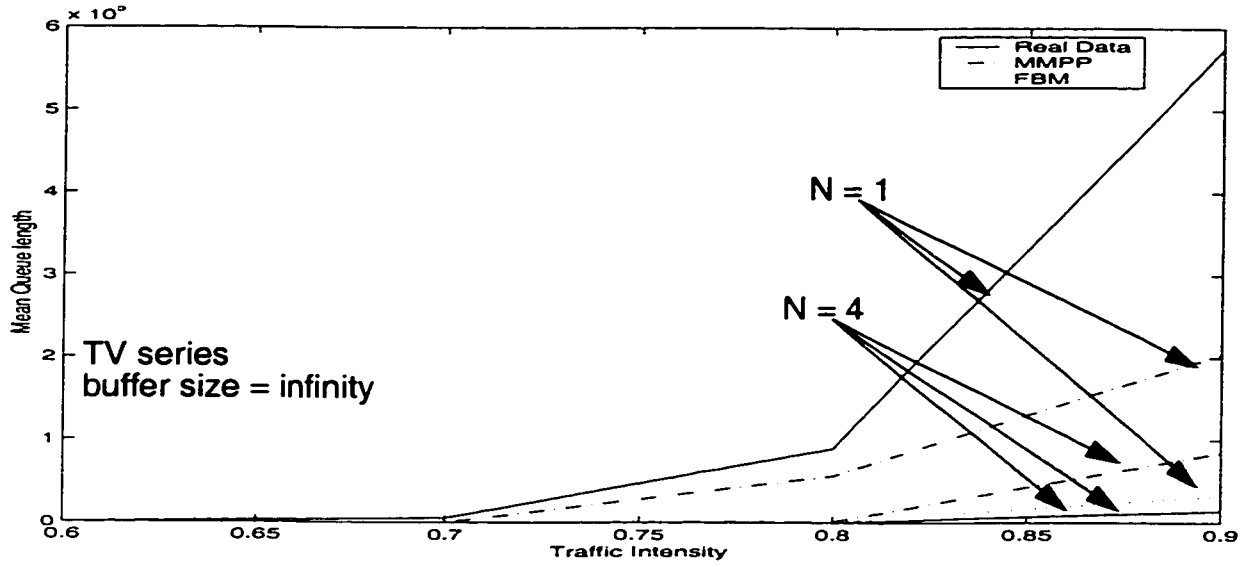

FIGURE.4.25. Comparison of the mean queue length of real Movie data. and that of FBM and MMPP for N =1, 4 multiplexed sources, infinite buffer size.

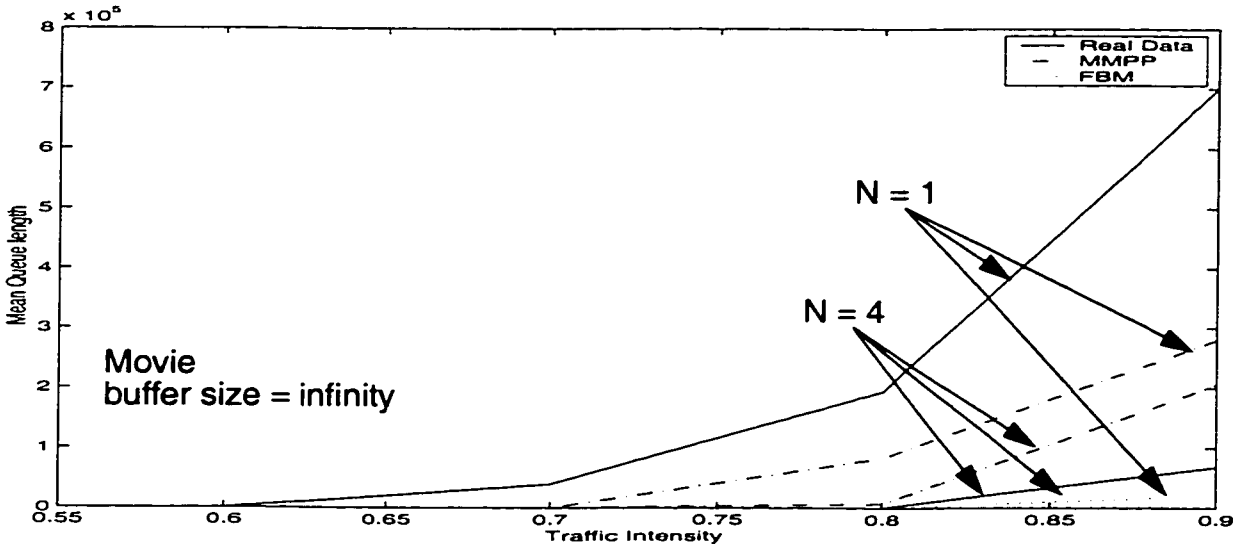

FIGURE.4.26. Comparison of the mean queue length of real Movie data. and that of FBM and MMPP for N =1, 4 multiplexed sources, infinite buffer size.

4.5 Discussion

The covariance and IDC are used to characterize arrival processes consisting of packets in an Ethernet LAN and frames in VBR video data. A procedure is introduced for utilizing index of dispersion for counts to fit a PMPP model to the

measured Ethernet data. We have fitted the IDC for FGN, FBM, F-ARIMA, PMPP, and MMPP models to the Bellcore data. FGN, FBM and F-ARIMA fitting is quite good over a large time scale, while PMPP prediction is reasonable for practical engineering design. The IDC for MMPP overestimates that of the Ethernet data. However, the fitting is reasonable when the lag is not large. Therefore, over a short time interval, MMPP matchs the Ethernet traffic. Simulation models FGN, FBM and F-ARIMA have the best matching of the probability of loss and mean queue length for the Ethernet data. We also showed that PMPP can reasonably predict the probability of cell loss and mean queue length for the Bellcore data over a large range of traffic intensity. However, the analytical performance measures for the PMPP are still unknown. MMPP estimation of the mean queue length as a function of traffic intensity is better than its estimation of the probability of loss. This is in agreement with the results that obtained by [HEF86, TUC88]. However, the matching of the probability of loss is reasonable when the traffic intensity is not large and the deviation of the MMPP prediction from that of the real Ethernet data increases when the traffic becomes more correlated and the buffer size is large.

Simulation models FGN, FBM, F-ARIMA and PMPP are not good models to characterize and predict video traffic. Their covariance, IDC, probability of loss and mean queue length largely fall below that of the real video traffic. The MMPP gives an excellent estimate of the covariance and the IDC of the teleconferencing VBR video data over a large number of frames. It fails to predict the covariance of the highly correlated entertainment traffic such as TV series and Movie. MMPP covariance underestimates the covariance of the real data, although, the MMPP prediction of the IDC for the highly correlated traffic is good. Moreover, MMPP accurately predicts the proability of loss and the mean queue length of the VBR video traffic when the number of multiplexed sources is large.

CHAPTER V

The Markov Chain and Self-Similar Traffic

5.1 Introduction

If the future evolution of the system depends only on its current state, the system may be represented by a Markov process. The information that is most often sought from such a model is the probability of being in a given state at a certain time after the system becomes operational. Often this time is taken to be sufficiently long that all influence of the initial starting state has been erased. The probabilities thus obtained are referred to as stationary probabilities. Probabilities at a particular time are called transient probabilities. When the number of states is small, it is relatively easy to obtain transient and stationary solutions quickly and accurately and from these to predict the behavior of the system. However, as models become more complex the process of obtaining these solutions becomes much more difficult.

It is well known that Markov chains do not have long range dependence behavior. The autocorrelation function for a Markovian chain decays exponentially. The autocorrelation of the self-similar traffic decays hyperbolically (obeying some power law) as the lag k increases rather than exponentially (see condition *iv* of section 1.7.2) [LEL94]. However, it is possible to model self-similar traffic as a Markov chain when the traffic correlation index is not large [HEY92] and be able to

predict the cell-loss rate and the mean queue delay for this kind of correlated traffic.

In this section, we show the results of approximating a given sequence by its quantized equivalent, consisting of L levels. We calculate the marginal probability distribution in two ways. First, we estimate the transition probabilities P from an actual sequence and use these to calculate the steady-state probabilities π according to equations (3.6) and (3.7) given in chapter 3:

$$\pi P = \pi \tag{5.1}.$$

$$\pi_1 + \pi_2 + \ldots\ldots\ldots\ldots + \pi_N = 1 \tag{5.2}.$$

Secondly, we compile a histogram of the samples. We perform these calculations on the video trace data presented in chapter 2. The effect of the long-range dependency on the estimate of the steady state probabilities, the performance measures and traffic indices of the Markov chain will be considered.

Once the steady-state probability vector π has been obtained, the computation of most of the interesting performance measures, such as the cell loss probability and the average delay, is straightforward. The drawbacks to the Markov chain model are that it has too many parameters and there is no apparent connection between the parameters and some easily measured statistics of the data [HEY92]. This limits the use of Markov chain in analysis, however, it can also be used in simulation.

Figure 5.1 shows part of the original and the quantized version of one of the video traces, video-conferencing, presented in chapter 2 over an interval of 500 frames and $L = 8$ quantization levels. The video-conferencing trace has maximum and minimum frame size of 629 and 23 cells per frame, respectively. For $L = 8$ quantization levels, the quantization step is calculated as:

$$\Delta = (max(confcam) - min(confcam))/number\ of\ levels$$

In this case, the quantization step Δ is approximately equal to 76 cells per frame.

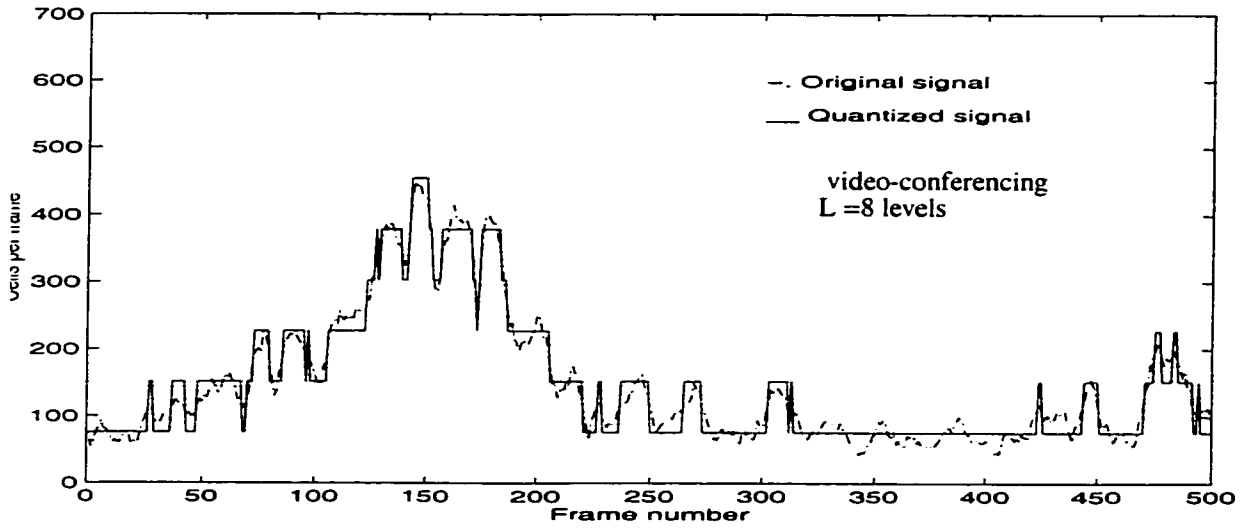

FIGURE.5.1. video-conferencing sequence. original and quantized version, 8 quantization levels.

5.2 Probability distribution

The Markov chain steady state probability distribution compared with the histogram of the original video-conferencing sequence is shown in figure 5.2 for L = 8, 16 and 32 quantization levels. As can be seen, as the number of levels increases the closeness of the match also increases. From Table 2.4, we see that the Hurst parameter H for the video-conferencing sequence is 0.72.

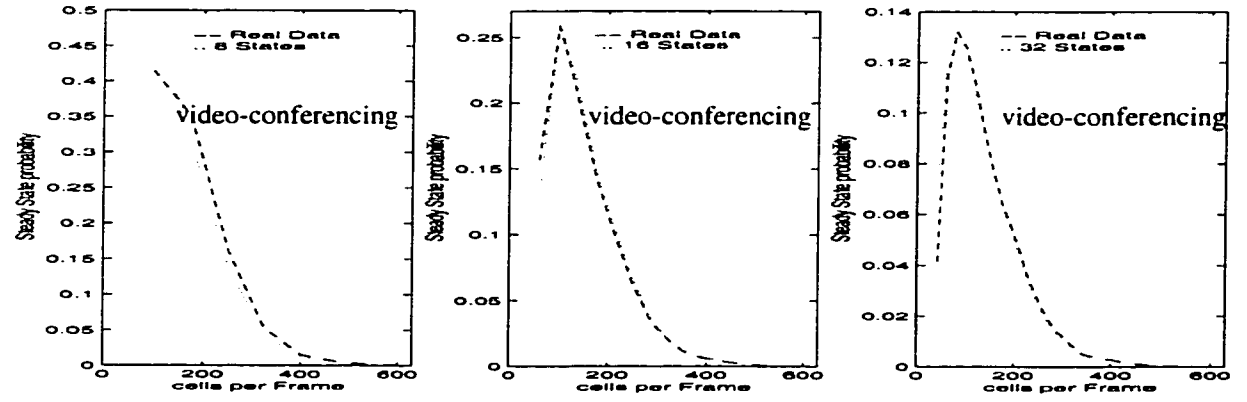

FIGURE.5.2. Comparison of the Markov chain steady state probability with that of histogram of the original video-conferencing sequence, 8, 16 and 32 quantization levels

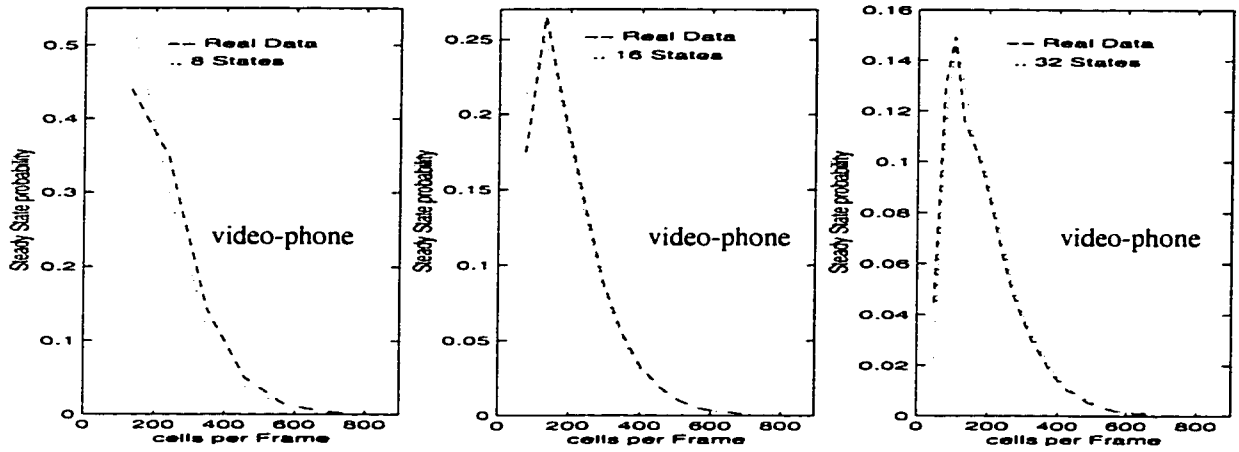

FIGURE.5.3. The Markov chain steady state probability compared with that of histogram of the original video-phone sequence, 8, 16 and 32 quantization levels

Figure 5.3 shows the Markov chain steady state probability distribution compared with the histogram of the original video-phone sequence for L = 8,16 and 32 quantization levels. As for the video-conferencing sequence, the matching is very close especially when the number of quantization levels increases. The Hurst parameter H for this sequence is 0.74.

The TV series Markov chain steady state probability distribution compared with that of the histograms for different quantization levels is shown in figure 5.4 for L = 8,16 and 32 quantization levels. Movie Markov chain steady state probability distribution compared with that of the histograms for different quantization levels is shown in figure 5.5 for L = 6, 12, 18 and 24 quantization levels. The matching for these two sequences is also as good as that for the video-conferencing and video-phone sequences. The Hurst parameter H for TV series and Movie sequences are 0.9 and 0.96, respectively.

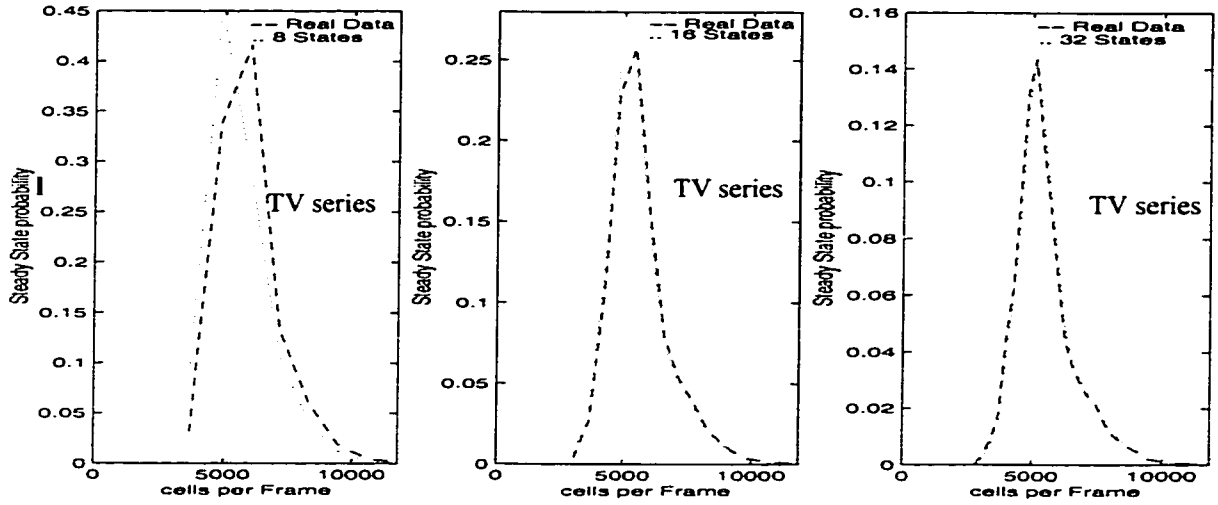

FIGURE.5.4. The Markov chain steady state probability compared with that of histogram of the original TV series sequence, 8, 16 and 32 quantization levels

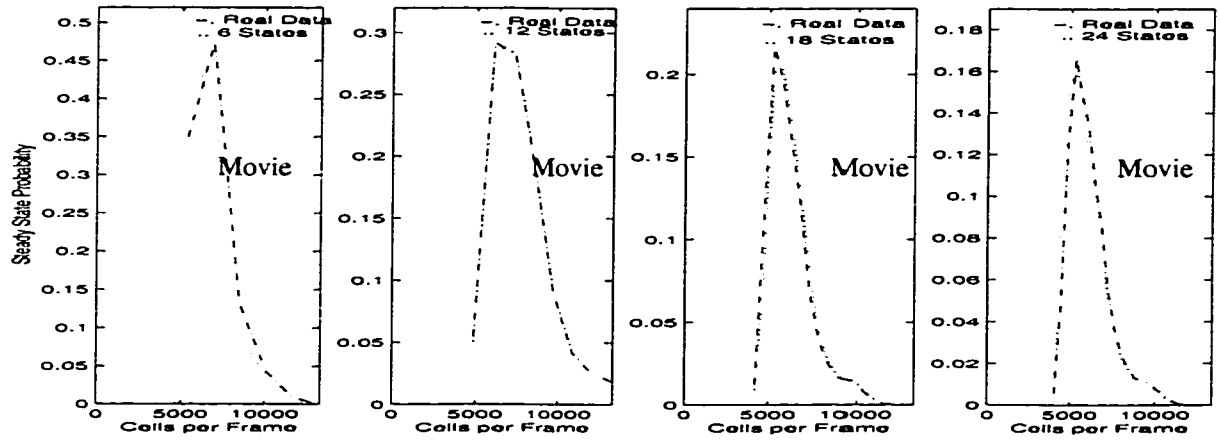

FIGURE.5.5. The Markov chain steady state probability compared with that of histogram of the original Movie sequence, 6, 12, 18, 24 quantization levels

5.3 Covariance and IDC

Given the transition probabilities for the Markov chain, we generate traffic using Matlab software based on the values of the estimated transition matrix from the real video data and compare the covariance and IDC of the generated Markov chain traffic with the real video data. The number of Markov chain frames

generated is 50000 frames, which is approximately the same as that of the real video traces. See Table 2.3. The covariance and the IDC are considered over a lag of length 1000 frames, which is 40 sec, since 25 frames are generated per second.

Consider the covariance of two real video teleconferences data, video-conferencing and video-phone, and the Markov chain model of the data. The covariances for the video-conferencing and video-phone data compared with that of the Markov chain counterpart quantized signal for $L = 8, 16$ are shown in figure 5.6 and figure 5.7, respectively. We see that the covariance of the Markov chains, which is a short range dependent processes, matches the covariance of the data especially when we increase the number of quantization levels. The covariance of the Markov chain typically is smaller than those of the data. Very good matching is achieved for this kind of data. We do the same for the two entertainment video sequences, Movie and TV series. The covariance functions are shown in figure 5.8 and figure 5.9, respectively. The covariance function of the Markov chain models are not a good representation of the covariance function of the data. The covariance functions of the Markov chain for the two entertainment video sequences are much smaller than the covariance functions of the data even for a larger number of quantization levels, that we used for the teleconferencing data. Moreover, the covariance function for TV series and Movie does not decline geometrically to zero as the video conference and video-phone scenes do.

The approximation of the Markov chain covariance to the teleconferencing video data becomes better as the number of quantization levels increases. This is because video-phone and video-conferencing sequences have a medium long-

range dependence parameter near 0.7. However, increasing the number of quantization levels for the highly correlated TV series and Movie sequences, those having a long range dependence parameter near 0.9, have very little effect on improving the matching (see Table 2.4).

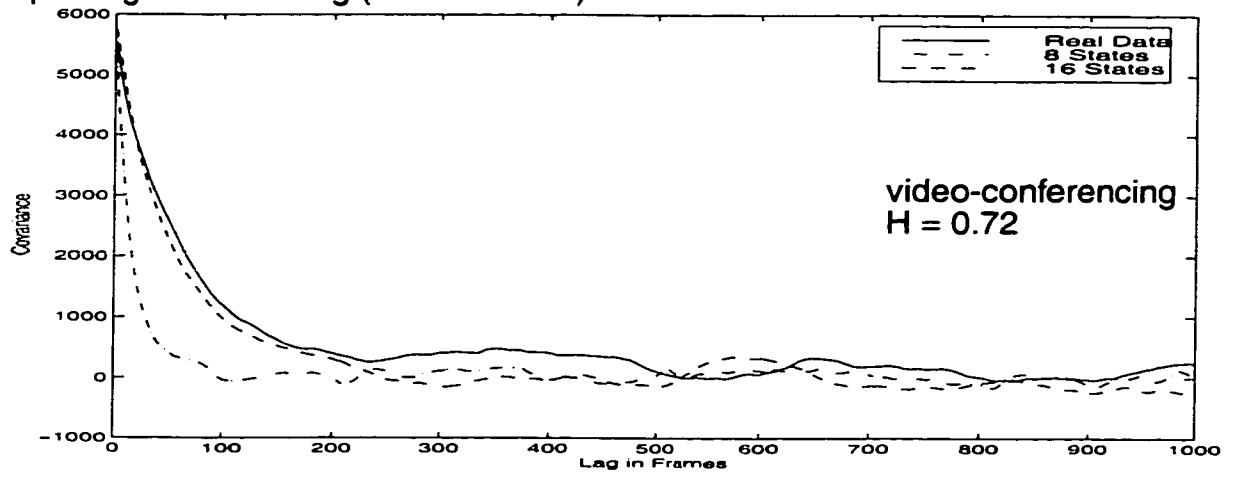

FIGURE.5.6. Covariance functions of the Markov chain compared with that of histogram of the original video-conferencing sequence for, 8 and16 quantization levels.

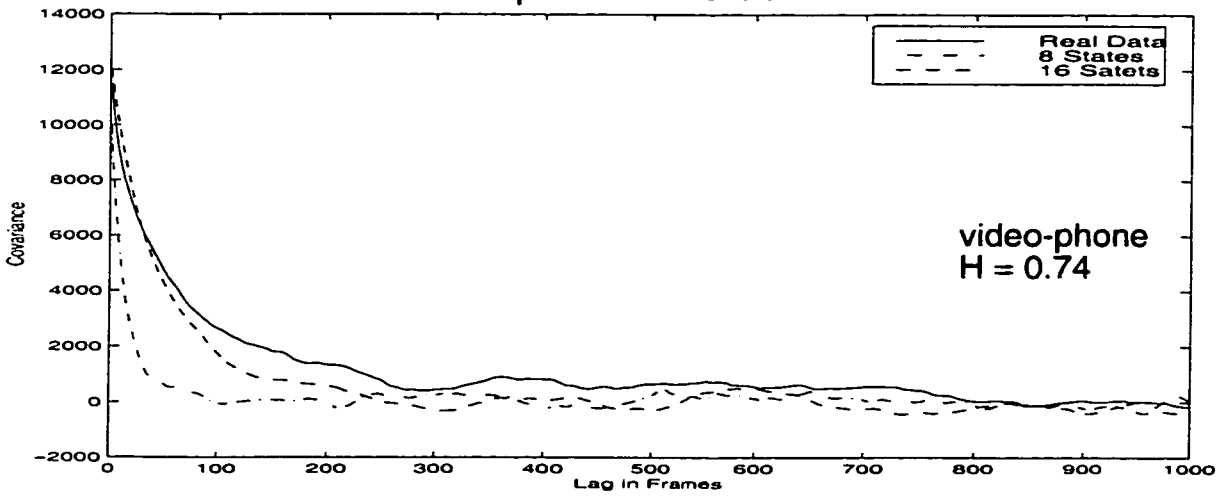

FIGURE.5.7. Covariance functions of the Markov chain compared with that of histogram of the original video-phone sequence for, 8 and16 quantization levels.

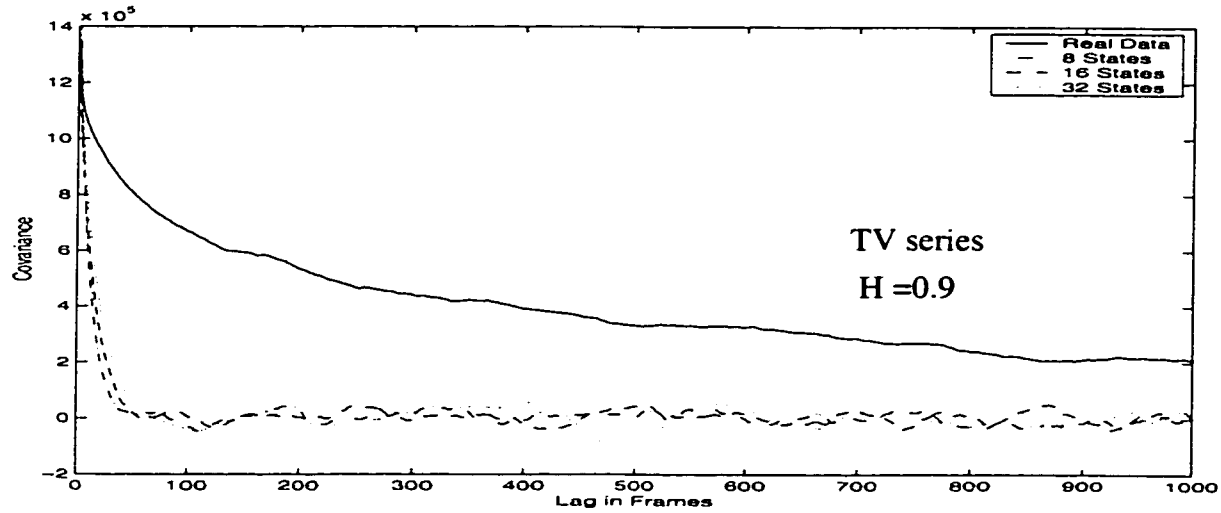

FIGURE.5.8. Covariance functions of the Markov chain compared with that of histogram of the original TV series sequence for, 8 and16 and 32 quantization levels

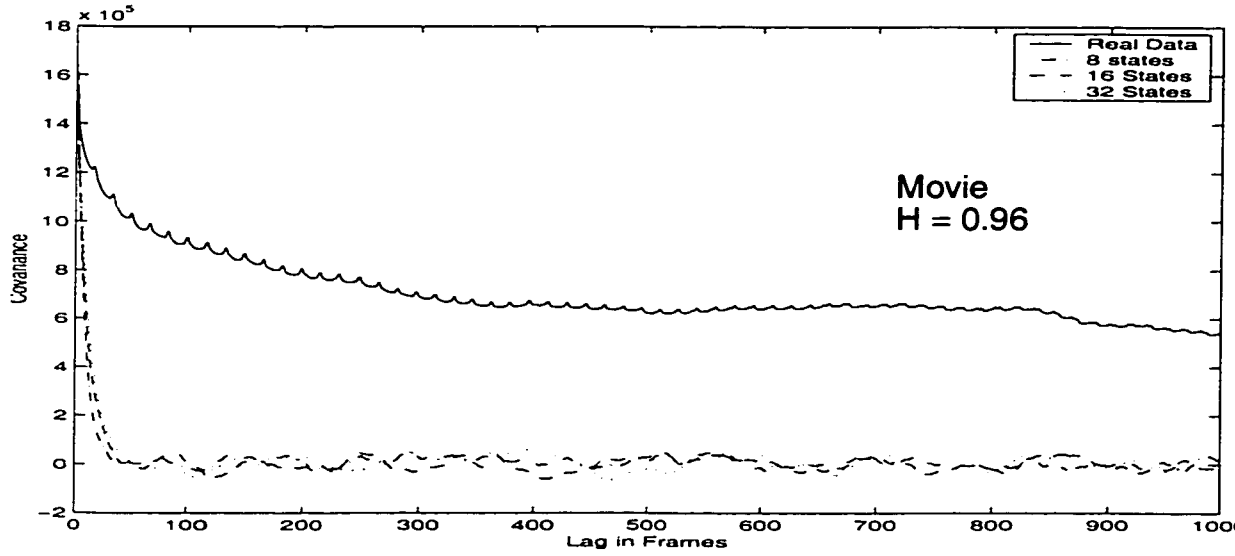

FIGURE.5.9. Covariance functions of the Markov chain compared with that of histogram of the original Movie sequence for, 8 and16 and 32 quantization levels

The same thing applies to the IDC, where the Markov chain approximation works for those with medium Hurst parameter such as teleconferencing as shown in figure 5.10 and figure 5.11. The matching is very good for the whole range of lags especially when we increase the number of quantization levels. However, for entertainment video such as TV series and Movie sequences shown in figure 5.12

and figure 5.13, the matching is good up to a small lag of value approximately 10 frames or 0.4 sec, since 25 frames are generated per second. Beyond this lag, the matching is poor even for a large number of quantization levels. As shown for entertainment data in figure 5.12 and figure 5.13, there is a little improvement for the fitted IDC when we increase the number of quantization levels from 8 to 16 and then to 32 levels. However, we see in figure 5.10 and figure 5.11, for the tele-conferencing data, there is a big improvement in the fitting of the IDC when increasing the number of quantization levels from 8 to 16.

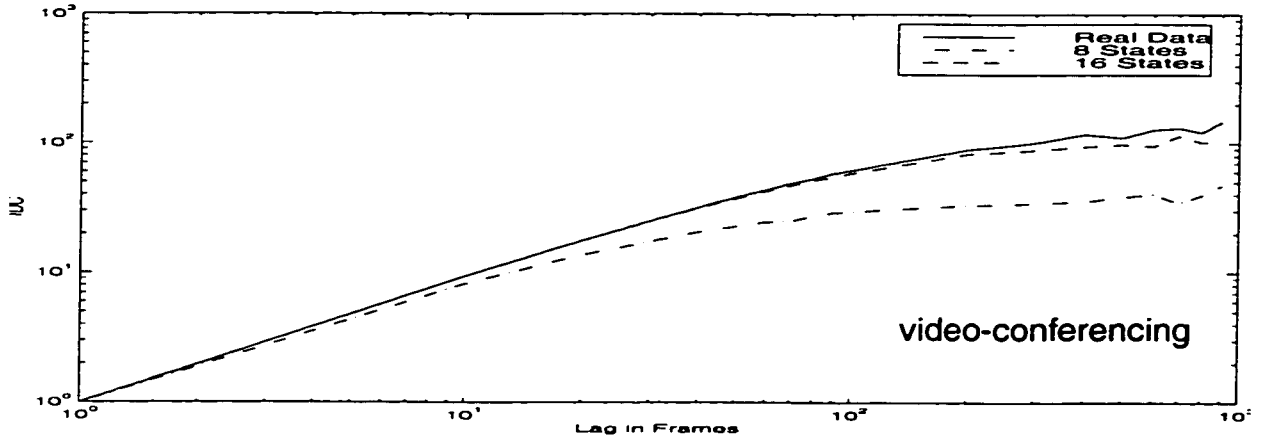

FIGURE.5.10. IDC of the Markov chain compared with that of histogram of the original video-conferencing sequence for, 8 and16 quantization levels

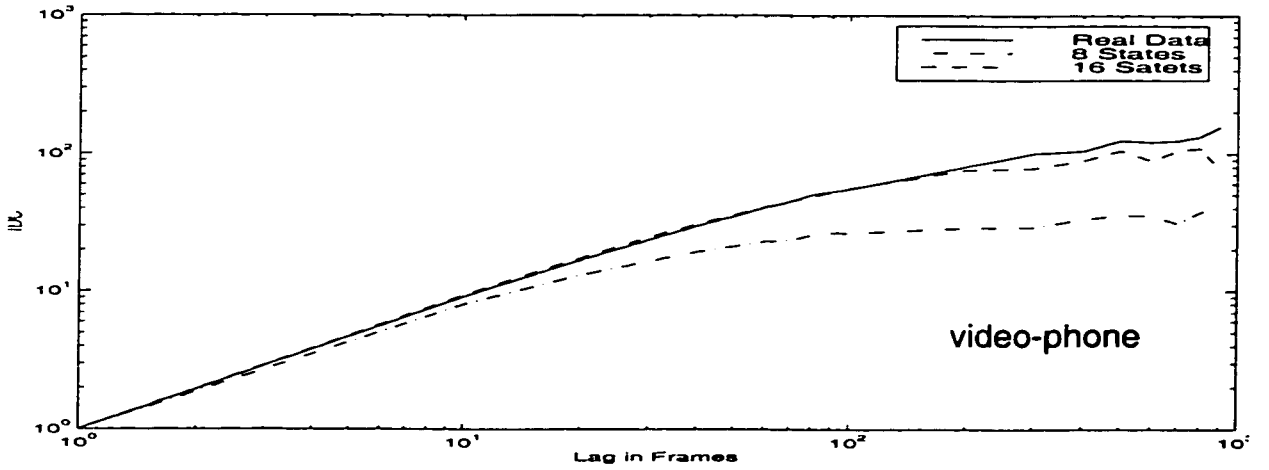

FIGURE.5.11. .IDC of the Markov chain compared with that of histogram of the original video-phone sequence for, 8 and16 quantization levels

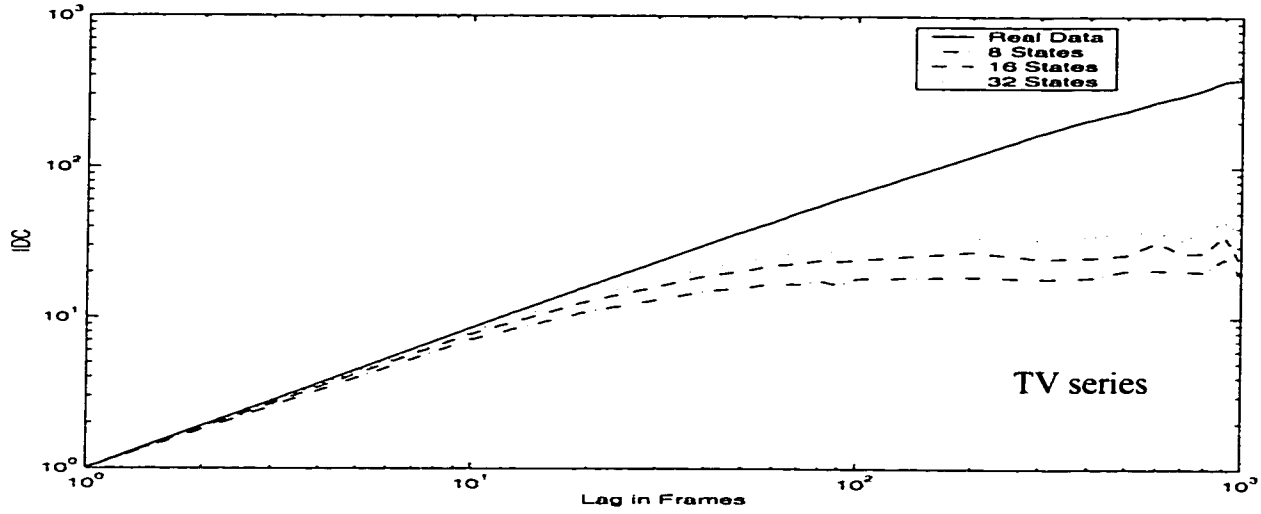

FIGURE.5.12. .*IDC* of the Markov chain compared with that of histogram of the original TV series sequence for, 8,16 and 32 quantization levels

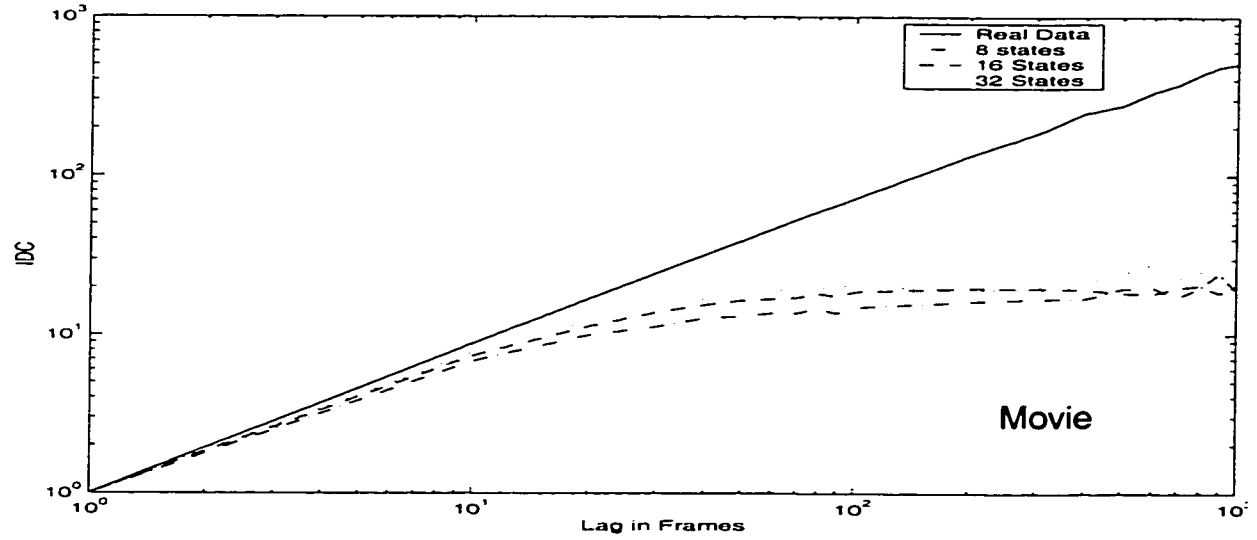

FIGURE.5.13. .*IDC* of the Markov chain compared with that of histogram of the original Movie sequence for, 8,16 and 32 quantization levels

5.4 Performance analysis

Given the transition probabilities for the Markov chain, to evaluate the cell loss, we use Matlab software to generate synthetic Markov traffic based on the values of the transition matrix and compare the performance measures with those of the real data.

5.4.1 Probability of loss

The mean number of cells per frame for video-conferencing and video-phone are 130 and 170, respectively as shown in Table 2.4. In our analysis, we consider the probability of loss for four video traces; video-conferencing, video-phone, TV series and Movie. figure 5.14 and figure 5.15 shows, respectively, the probability of loss for video-conferencing and video-phone as a function of the traffic intensity with the buffer size treated as a parameter of value equal to 100 cells. The results shown are excellent over the entire range of the traffic intensity. As the number of quantization levels increases, the approximation of the Markov chain to the real data becomes more accurate. The probability of loss for video-phone trace is larger than that of the video-conferencing trace, assuming the same buffer size and same number of quantization levels. This is due to the fact that the video-phone trace is more bursty (larger Hurst parameter H) than the video-conferencing trace.

The simulation experiments we report have cell-loss rates larger than 10^{-6} because the data do not have enough cells to reliably estimate cell-loss rates any smaller. This result predicts that if we could do experiments with smaller cell-loss rates, the accuracy of the Markov chain models would be better than the accuracy we are achieving now.

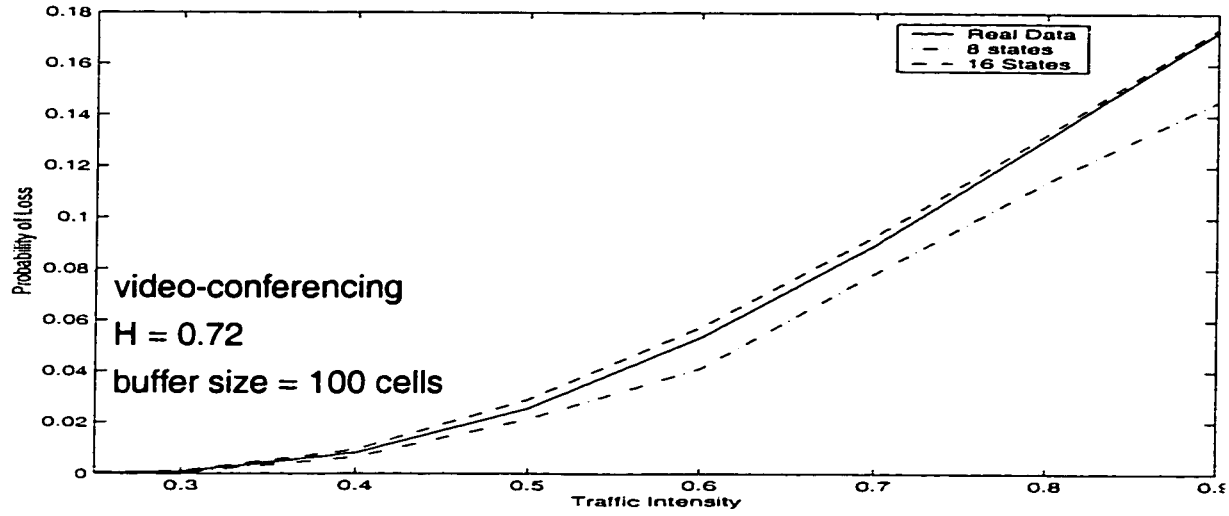

FIGURE.5.14. Comparison of the Markov chain probability of loss with that of histogram of the original video-conferencing sequence, 8 and16 quantization level.

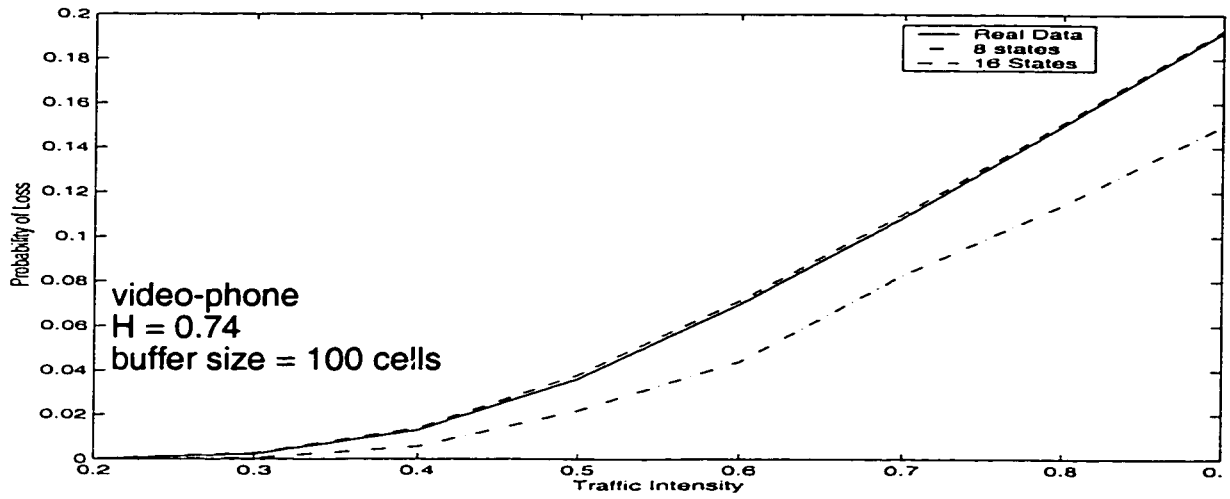

FIGURE.5.15. Comparison of the Markov chain probability of loss with that of histogram of the original video-phone sequence, 8 and16 quantization level

The mean number of cells per frame for TV series and Movie are, 5948 and 5336, respectively. See Table 2.4. We choose a buffer size for both entertainment traces of 4000 cells. In figure 5.16 and figure 5.17, the probability of loss for the highly correlated traffic TV series and Movie are shown, respectively. As expected, from the comparison of the covariances and the IDC for these two data shown in figures 5.8 - 5.9 and figures 5.12 - 5.13, the matching of the probability

of loss between these two video sequences and the Markov chain is not in good agreement. This is as pointed out above due to the high correlation index of the traffic which is 0.9 and 0.96 for TV series and Movie sequences, respectively. It is interesting to see that the matching of the covariance and the IDC seems to play a major rule in the approximation of the probability of loss. In other words, if the matching of the covariance or the IDC is good over a large interval of frames, then the probabilities of loss also seems to be in good agreement.

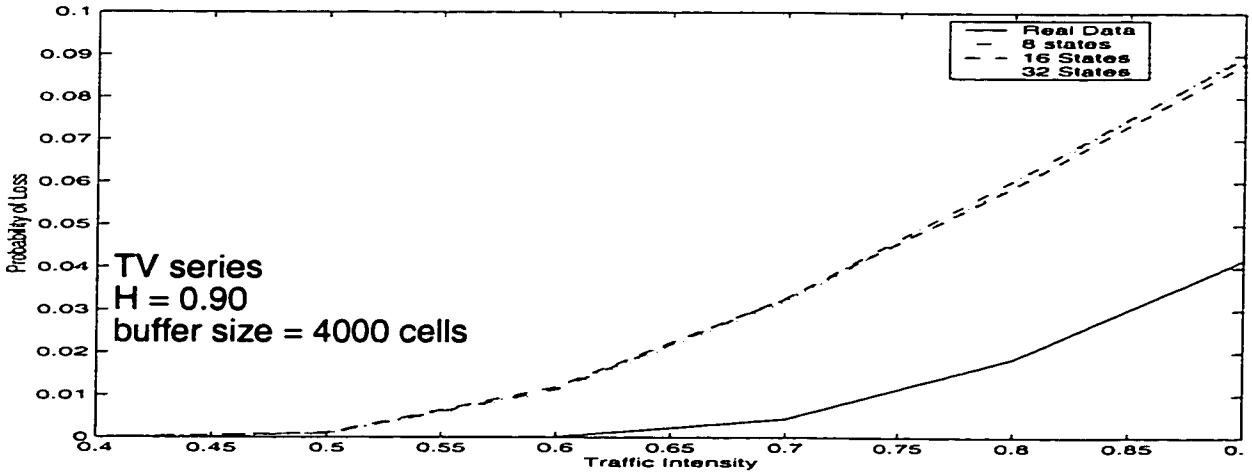

FIGURE.5.16. Comparison of the Markov chain probability of loss with that of histogram of the original TV series sequence, 8 and16 and 32 quantization level.

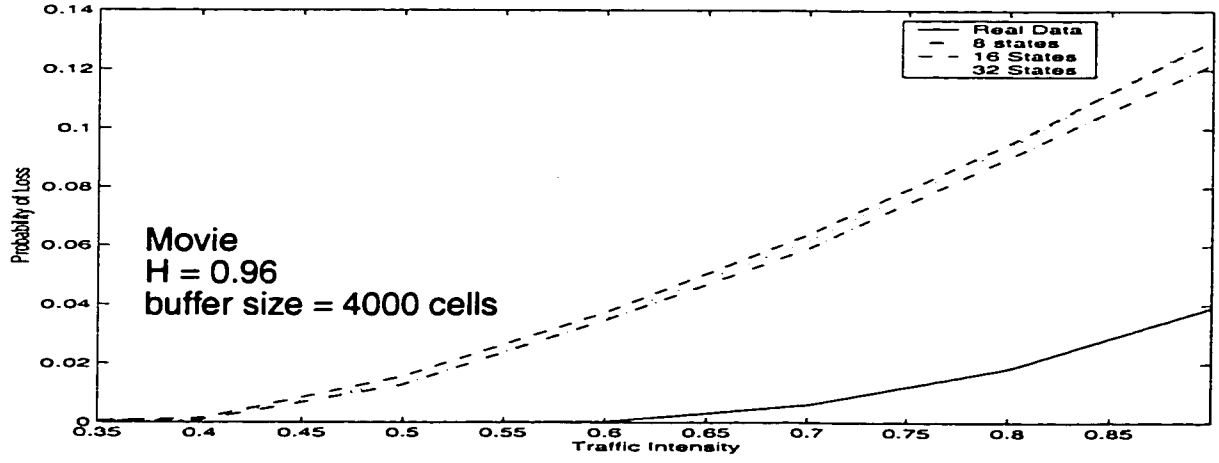

FIGURE.5.17. Comparison of the Markov chain probability of loss with that of histogram of the original Movie sequence, 8, 16 and 32 quantization level.

5.4.2 Mean queue length (finite and infinite buffer)

Another important performance measure that we discuss in this section is the mean queue length, from which the mean queueing delay can be obtained using Little's formula. We consider two cases: finite buffer and infinite buffer.

5.4.2.1 Finite buffer

We show in figure 5.18 and figure 5.19 the mean queue length for video-conferencing and video-phone sequences as a function of the load with buffer capacity of 100 cells. Two Markov chain paths are shown, along with a curve of the real data. As the number of quantization levels increases, the prediction of the mean queue length gets better. The 16-state Markov chain path follows the curve of the data very well and the knee of the curves are aligned as shown in figure 5.18 and figure 5.19 for video-conferencing and video-phone sequences, respectively. The mean queue length for TV series and Movie Markov chain paths as a function of the load and buffer size of 4000 cells does not match that of the original data even for a large number of quantization levels (32 levels) as compared with the good prediction of the teleconferencing sequences for smaller number of quantization levels (16 levels). The results are shown in figure 5.20 and figure 5.21. As indicated before, this is because of the high correlation index in this kind of traffic.

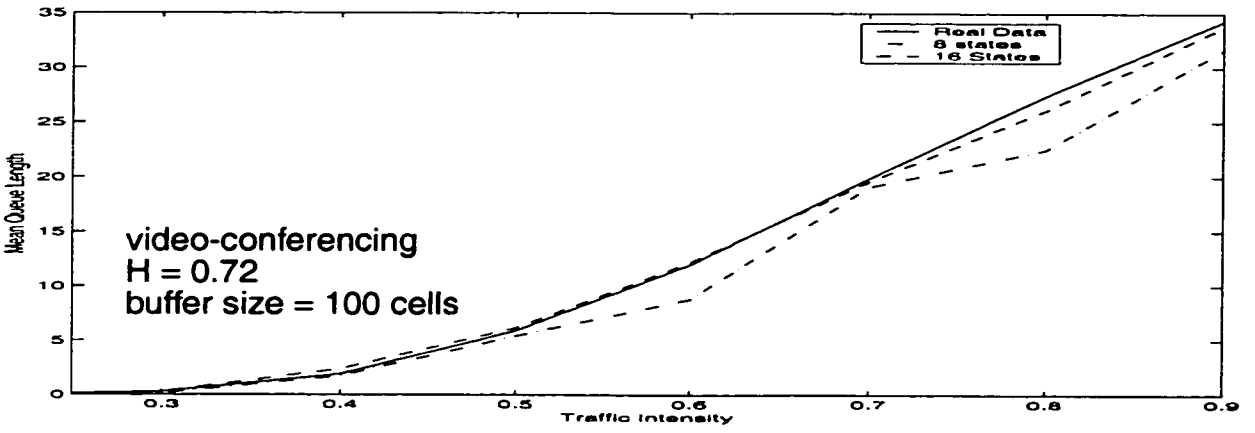

FIGURE.5.18. Comparison of the Markov chain mean queue length versus traffic intensity with that of the histogram of video-conferencing for 8 and 16 quantization levels.

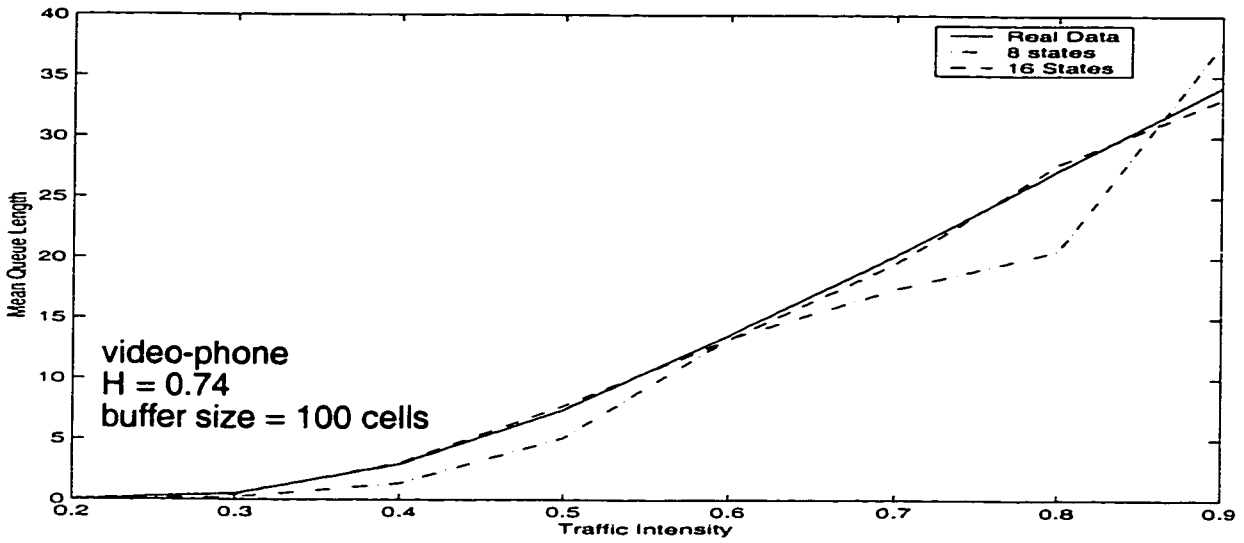

FIGURE.5.19. Comparison of the Markov chain mean queue length versus traffic intensity with that of the histogram of video-phone for 8 and 16 quantization levels.

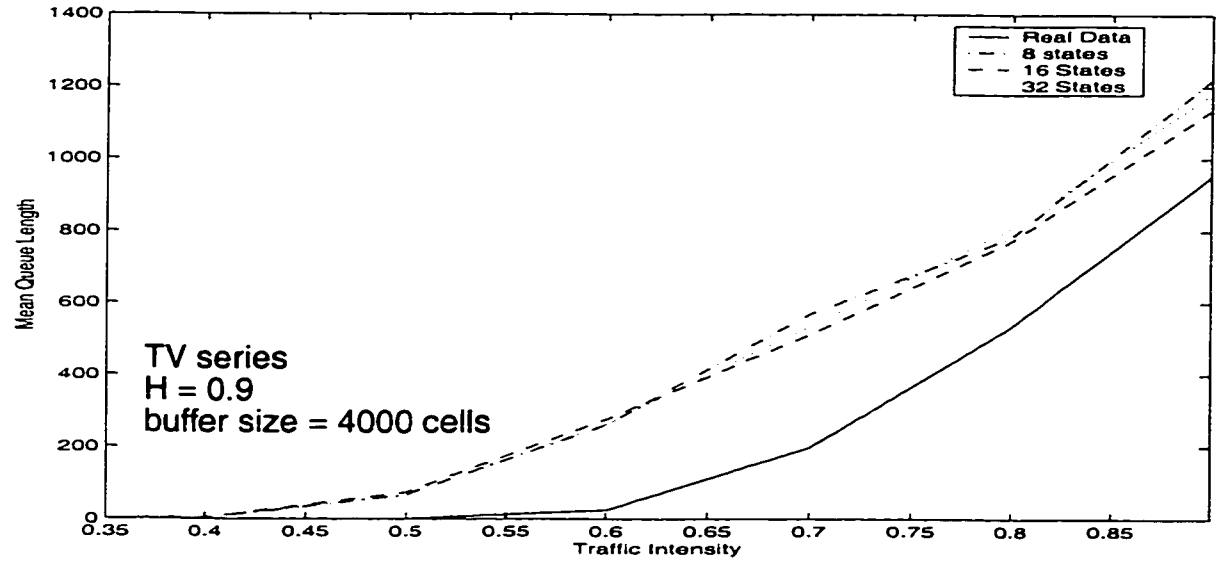

FIGURE.5.20. Comparison of the Markov chain mean queue length versus traffic intensity with that of the histogram of TV series for 8,16 and 32 quantization levels.

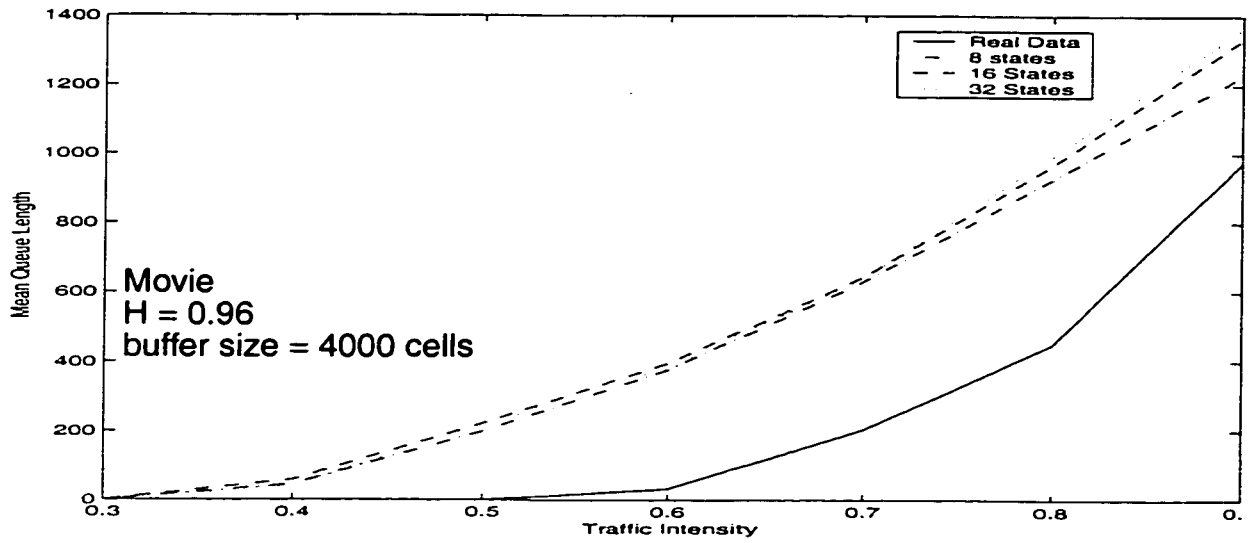

FIGURE.5.21. Comparison of the Markov chain mean queue length versus traffic intensity with that of the histogram of Movie for 8,16 and 32 quantization levels.

5.4.2.2 Infinite buffer

For the infinite buffer case and as shown in figure 5.22 and figure 5.23, the Markov chain paths mean queue length becomes close to that of the real video-conferencing and video-phone data as the number of quanization levels increases and the load is not large. As the load increases, the matching becomes less accurate. This mismatch in this case is because we have a large buffer, and therefore, correlation is more dominant. However, for light loading of up to 0.7 and number of quantization levels around 32 levels, the prediction is satisfactory for practical engineering design. The Markov chain mean queue length paths of the real TV series are shown in figure 5.24. As shown the paths are in agreement with that of the real sequence load for small load values and beyond that the discrepancy becomes clear. As the load increases, the discrepancy between the real path and that of the Markov chain becomes larger than that when we have finite buffer as shown in figure 5.16. The Markov chain paths mean queue length of the Movie trace shown in figure 5.25 are matching the path of the real data for small loading. That is, it broke earlier than the case for the TV series. This is because Movie has

higher Hurst parameter than TV series. The large mismatch for TV series and Movie, respectively, are due to the high correlation of this kind of video traffic and to the large size of the buffer.

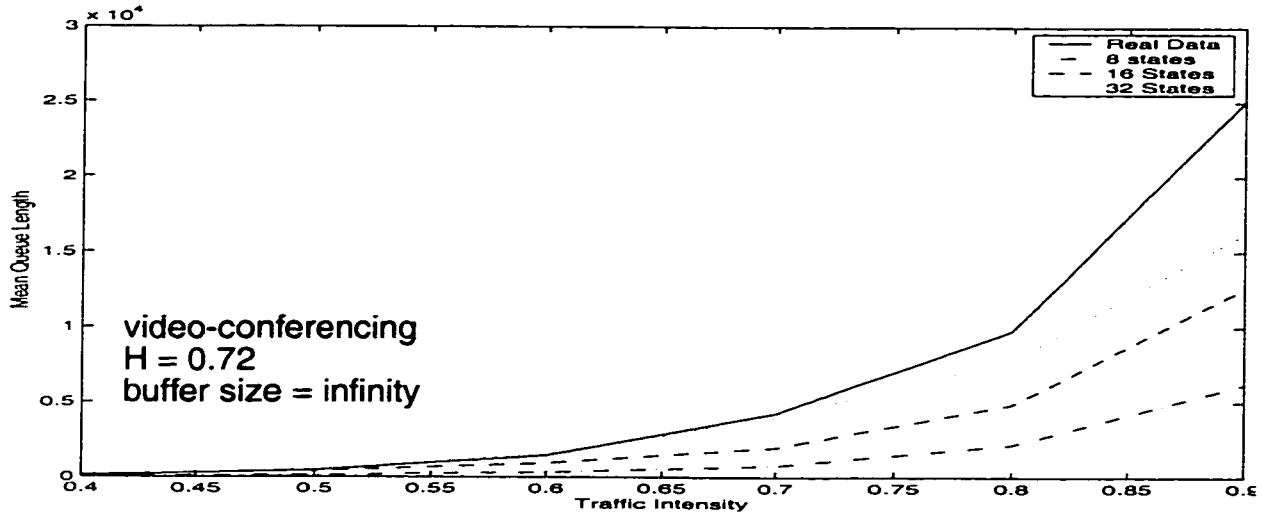

FIGURE.5.22. Comparison of the Markov chain mean queue length for infinite capacity versus traffic intensity with that of the histogram of video-conferencing for 8,16 and 32 quantization levels

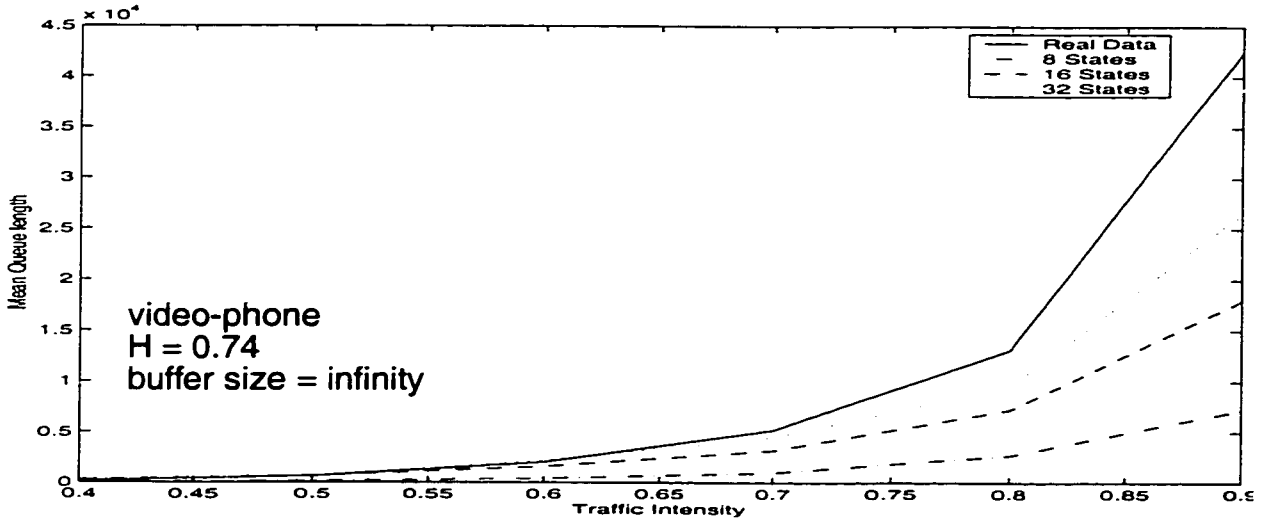

FIGURE.5.23. Comparison of the Markov chain mean queue length for infinite capacity versus traffic intensity with that of the histogram of video-phone for 8 and 16 quantization levels.

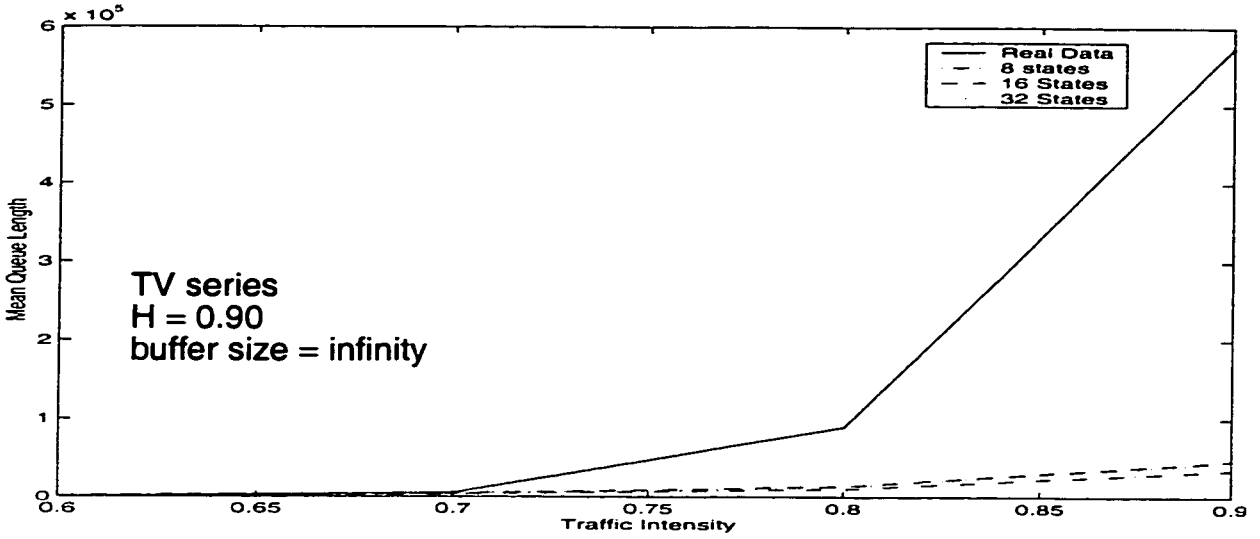

FIGURE.5.24. Comparison of the Markov chain mean queue length for infinite capacity versus traffic intensity with that of the histogram of TV series for 8,16 and 32 quantization levels.

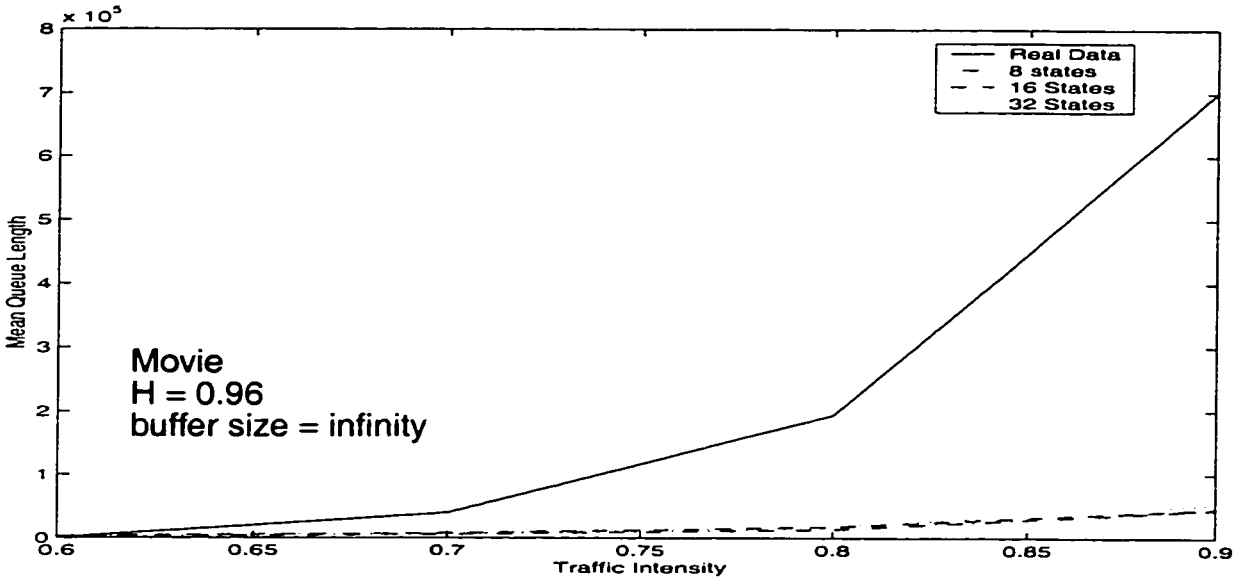

FIGURE.5.25. Comparison of the Markov chain mean queue length for infinite capacity versus traffic intensity with that of the histogram of Movie for 8,16 and 32 quantization levels.

5.5 Modeling multiplexed sources

In this section, we show that multiplexing several statistically independent and identical highly correlated sources will result in a good prediction of the probability

of loss and mean queue length. We represent multiplexed sources by superposing them, i.e., by adding the number of cells that arrive in the same frame interval. That is, when a new frames from a single source arrive every 40 ms, we model the multiplexed sources by adding the number of cells that arrive in each 40 ms interval. As we explained in chapter 3, with this model the advantage of multiplexing is due to smoothing the peaks and valleys of traffic as a result of averaging. We present multiplexing of the video sources video-conferencing, video-phone, TV series and Movie. First we show how highly correlated sources discussed in the previous section such as TV series and Movie can be modeled as a Markov chain when several sources are multiplexed. We simulate many video sources in many traffic environments such as, traffic intensities, number of multiplexed sources, and buffer size. We show the effect of multiplexing on the reduction of the probability of loss and the mean queue length for video-conferencing and video-phone sequences.

For TV series data, figure 5.26 shows the IDC for 10 multiplexed real sources and 10 multiplexed 8-state Markov chains. As shown, the matching is excellent over a long rang of lag. We also obtained good results for the Movie data, which is shown in figure 5.27. In figure 5.27, we show the IDC for 10 multiplexed Movie sources and 10 multiplexed 8-state Markov chains. We see that the model does track the curve obtained from the data. This improvement in the matching of the IDC as compared to those shown in figure 5.12 and 5.13, is because of reducing the correlation between frames due to multiplexing. Our explanation was that, the multiplexing of the sources distorts the correlation between frames, and when this correlation is eliminated, the IDC for the model and the data agree.

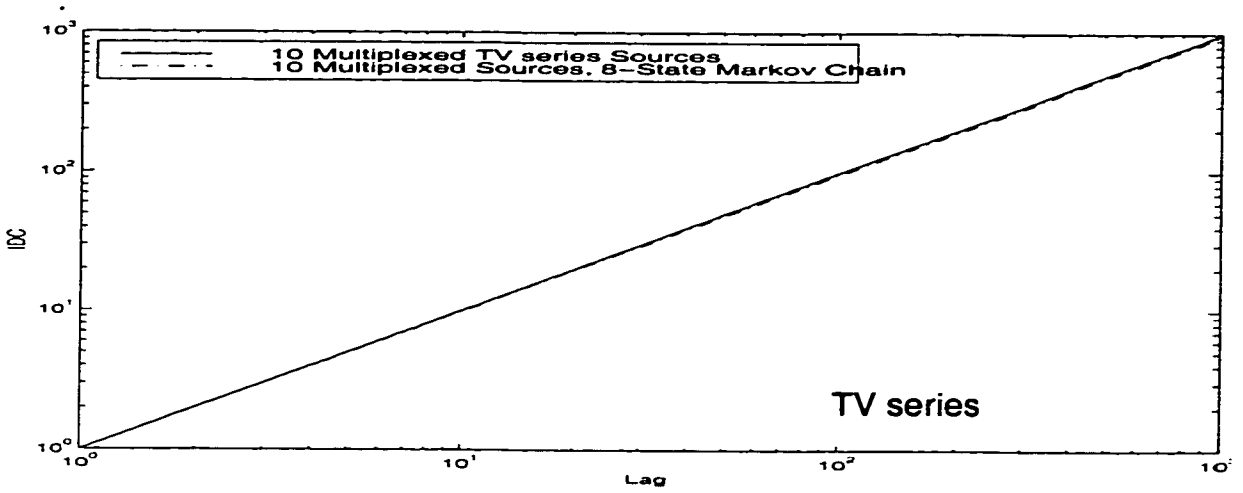

FIGURE.5.26. Comparison of IDC for10 multiplexed sources 8-state Markov chain with that of the original 10 multiplexed TV series sequence.

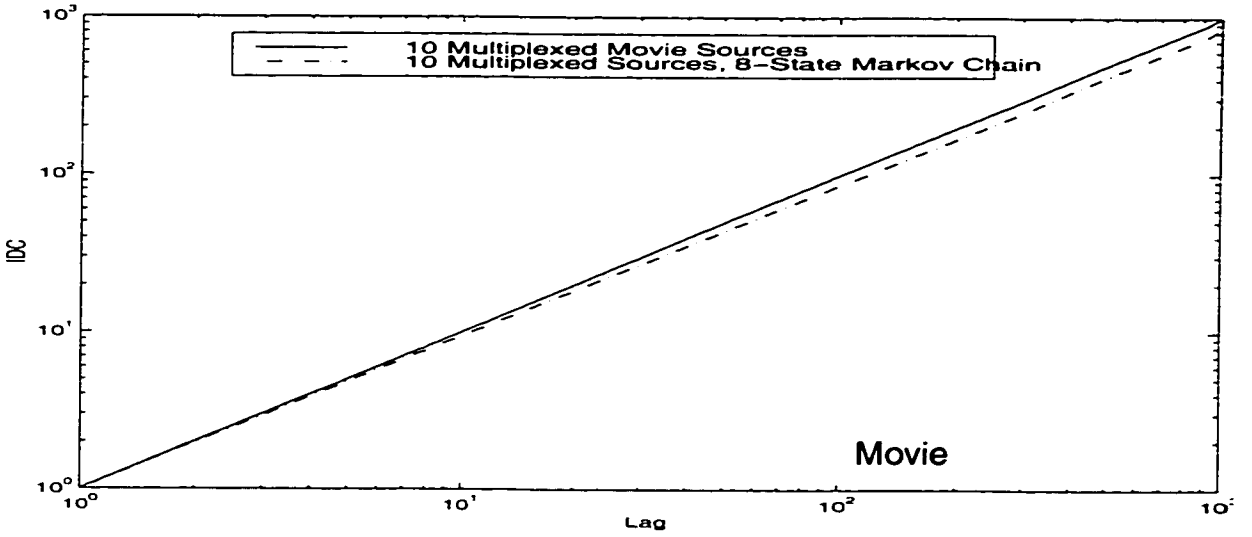

FIGURE.5.27. Comparison of IDC for10 multiplexed sources 8-state Markov chain with that of the original 10 multiplexed Movie sequence.

The probability of loss and the mean queue length for the TV series and Movie sequence are shown in figure 5.28 and 5.29, respectively. We choose the buffer size for both sequences to be a finite of value 4000 cells. As shown in the figures, we see that the model does an excellent job of tracking the curve obtained from the data. It is clear that the results we achieved because of multiplexing are

good over large traffic intensity intervals. The model does a very good job of matching the probability of loses and mean queue lengths.

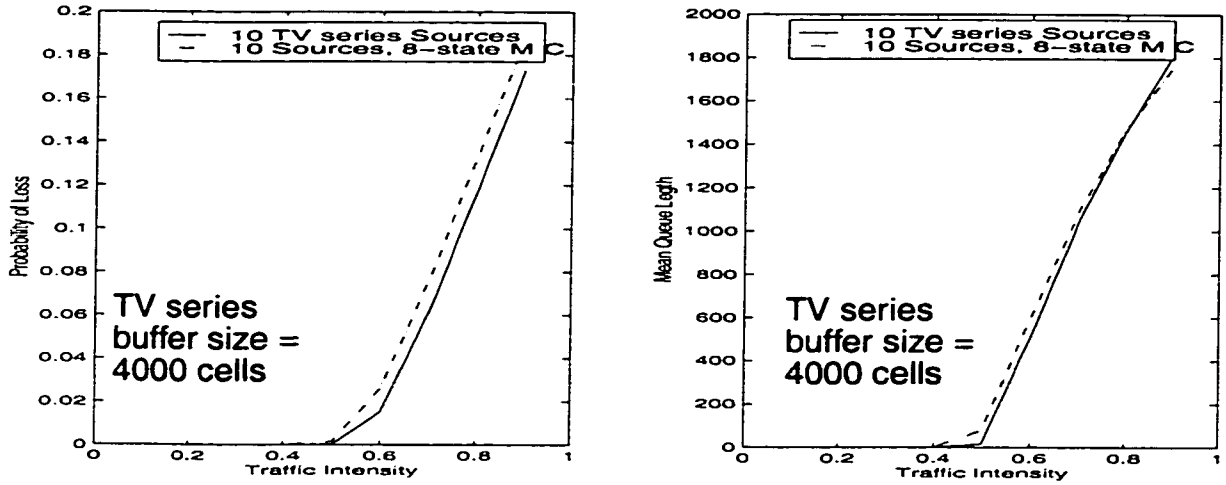

FIGURE.5.28. Comparison of probability of loss and mean queue length for10 multiplexed sources 8-state Markov chain with that of the original 10 multiplexed TV series sequence, buffer capacity = 4000 cells.

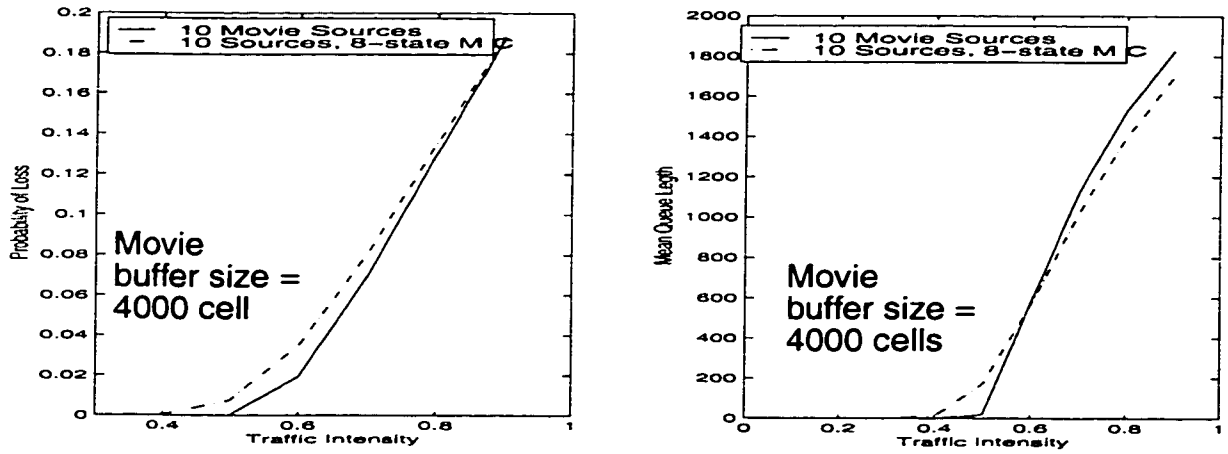

FIGURE.5.29. Comparison of probability of loss and mean queue length for10 multiplexed sources 8-state Markov chain with that of the original 10 multiplexed Movie sequence, buffer capacity = 4000 cells.

In figure 5.30 and figure 5.31, we show the behavior of statistically multiplexed sources. We chose the number of sources being multiplexed N=1, 2, 4 and 8. We considered the video-conferencing and video-phone sequences probability of loss and mean queue length as a function of the traffic intensity. The mean number of cells per frame of the two sequences are 130 and 170 cells, respectively. See

Table 2.4. We choose in our simulation the buffer size of both sequences to be finite of value150 cells. As shown in the figures, the reduction of the probability of loss and the mean queue length increases as the number of multiplexed sources increases. This is due to the fact that as we multiplex more sources the total traffic stream becomes less bursty.

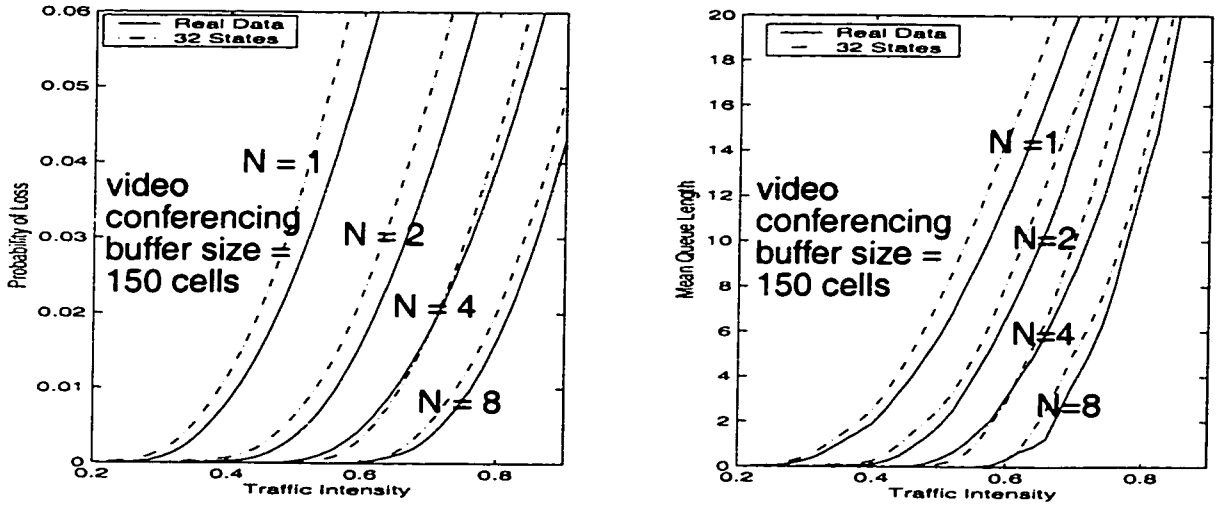

FIGURE.5.30. Behavior of statistically multiplexed video-conferencing sources. Probability of loss and mean queue length versus traffic intensity, constant buffer size =150 cells for number of sources N=1, 2, 4 and 8.

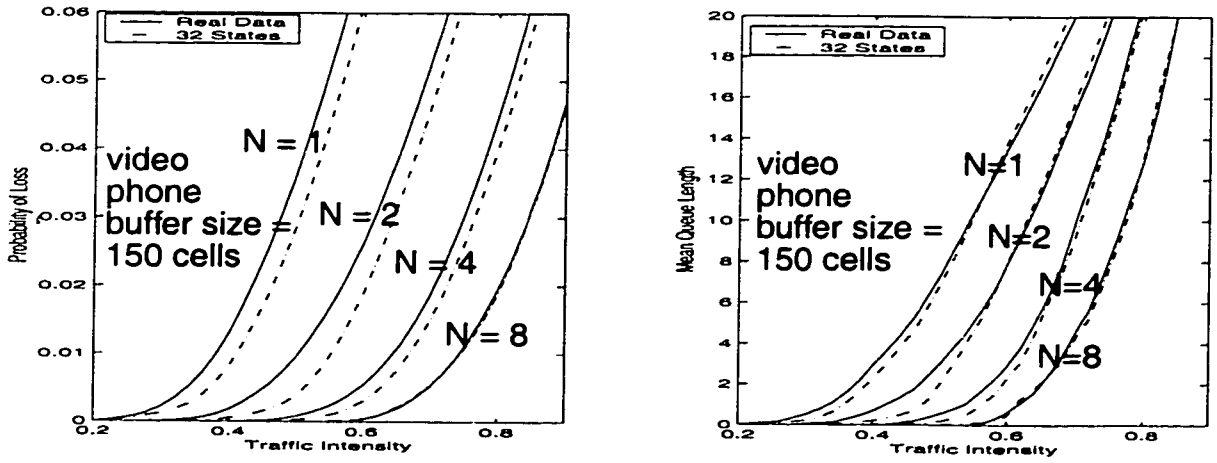

FIGURE.5.31. Behavior of statistically multiplexed video-phone sources. Probability of loss and mean queue length versus traffic intensity, constant buffer size =150 cells for number of sources N=1, 2, 4 and 8.

5.6 Discussion

We examined four video data sequences. Two of them are video teleconferences and the cther two are video entertainment. The teleconferencing data has a medium Hurst parameter, while the entertainment has a high Hurst parameter. It should be noted that we make comparison between video data that has comparable statistics such as mean, variance and number of cells per frame. That is we compare video-conferencing and video phone on one hand and we compare TV series and Movie on the other hand. See section 2.31 and Table 2.4 for more detail of the video traffic data statistics.

When the Hurst parameter is not large, as when we have the teleconferencing data, Markov chain models have an excellent estimates of covariance, IDC, cell loss probabilities and mean queue length. Moreover, as the number of quantization levels L increases, the accuracy of the matching between the Markov chain and that of histogram of the original teleconferencing sequence increases. For entertainment data, which has high Hurst parameter, the covariance, IDC, cell loss probabilities and the mean queue length are not estimated well when the traffic intensity is large. Moreover, increasing the number of quantization levels for the entertainment data has little effect on improving the matching of the statistical indices and the QoS. However, multiplexing several correlated video sources will improve the characterization and the prediction of the real traffic.

From these results, we draw the tentative conclusion that there is a link between the Hurst parameter and the utility of the Markov chain approximation. Even if we have a good estimation for the steady state probabilities between the Markov chain and the real data for different video sequences ranging from 0.72 to 0.96, the matching for the covariance, the index of dispersion for counts, probability of loss and mean queue length works only for those of medium Hurst parameter around 0.7. Multiplexing several highly correlated source will smooth the traffic and good results are achieved regarding traffic characteristic indices and performance measures.

CHAPTER VI

Modeling of Self-Similar Traffic using Heterogeneous ON-OFF Source model

6.1 Introduction

The criteria that determine the utility of a model are analytical tractability, simplicity in terms of the number of parameters that are involved and goodness of fit to actual data. In our work the goodness of fit is based on the ability of the model to capture the covariance function of real data and then to predict probability of loss and mean queue length for real data.

Until recently it has not been clear whether Markov based models could be used to model self-similar traffic. It has been claimed that the large number of states needed to model the traffic makes Markov models inapplicable for all practical purposes. This has initiated the search for other models that might be more suitable for modeling self-similar traffic such as FGN, FBM, F-ARIMA and chaotic maps [PRU95]. For these models, however, the analytical tools for analyzing queueing behavior do not exist. However, they may be used in simulation.

There are basically two analytical models, Batch Markov Arrival Processes (BMAP), which has MMPP [HEF86] as a special case, which we presented in chapter 4 to model video data, and fluid flow models. Fluid models characterize traffic as a continuous stream [ANI82]. A fluid model that is normally used to model traffic is the Markov modulated fluid model. In this model, the current state

of the underlying Markov chain determines the flow rate with a different rate for each state of the chain. This model is a Markov modulated constant rate model and is used to model variable bit rate video [MAG88, ELW93].

The ON-OFF source model is the most popular model for voice. It was used to model video traffic based on the minsources approach by Maglaris [MAG88]. Anick, Mitra and Sondhi [ANI82] used the ON-OFF sources to analyze bursty traffic. The ON-OFF source model is tractable for analysis when the transitions from the ON state to OFF state and from OFF state to ON state are exponentially distributed.

In our case, we will use m classes of heterogeneous ON-OFF sources to model video data. This model is based on matching the total covariance of the heterogeneous sources to the real data. The covariance of the m heterogeneous sources is composed of m different exponential functions, while in the homogenous case it is just one exponential [MAG88]. The model is very attractive, because as we will see for a small number of ON-OFF sources it is possible to get a good results for the probability of loss and mean queue length. Moreover, the small number of parameters makes the analysis in finding the covariance and the parameters of the sources simple.

Anderson, et. al. [AND97] presented a fitting method for modeling second order processes. The fitting method is based on fitting to the autocorrelation function of counts for a second order self-similar process. They have shown that it is possible to match the autocorrelation function of counts for a second order self-similar traffic over 3 - 5 time scales with 8 - 16 state Markovian Arrival Processes

(MAPs). However, they have shown that the second order properties of counts for the arrival process are not sufficient for predicting queueing performance.

The model we developed is different from that of Andersen, et. al. We used the Feldmann algorithm [FEL97] for approximating a long-tail covariance function by a finite mixture of exponentials. However, Feldman, et. al., used the algorithm to fit probability distribution. Our model is simpler than Anderson's model. The matching of the covariance and IDC for the real data to the traffic generated using the model is quite good. The prediction of the queueing performance such as the probability of loss and mean queue length for different VBR video traffic is fair.

6.2 The mathematical model

In this section we consider m independent classes of ON-OFF sources, let N_i $(i = 1, 2,....m)$ denote the number of sources in class i to model long-range dependence traffic such as video. Within a class the sources are identical and independent. In this model, for the *ith* class, packets are generated during talk spurts which is the ON state, and no packets are generated during the OFF state. The times spent in the ON and OFF states are exponentially distributed with means $1/\beta_i$ and $1/\alpha_i$, $i = 1, 2,....m$, respectively. When the source is in the ON state it generates data at rate of R_i, $i = 1, 2, ...m$.

The ATM multiplexer consists of a server transmitting cells at a specified line rate and a buffer whose size is determined by the delay constraints on cell trans-

mission. Cells arrive at the multiplexer from a number N_i $(i = 1, 2,....m)$ of sources. See figure 6.1.

The basic idea of the 3-class model is that there are three time frames for transitions: short term, medium term and long term, respectively. The transition rates are such that $\alpha_1 \gg \alpha_2 \gg ..... \gg \alpha_m$ and $\beta_1 \gg \beta_2 \gg ..... \gg \beta_m$, where for our model we have $m = 3$, so that the shorter the time frame, the more rapid the transition. For example, in the case of three level Markov chain the possible transitions are illustrated in figure 6.2. From each state there are three possible transitions: from state 1 there is a short term transition that will take us to state 2 given by α_1, a medium term transition to state 3 given by α_2 and finally a long term transition that will take us to state 5 given by α_3. The most likely transition is to state 2. As shown in figure 6.2, the three level two state model is complicated.

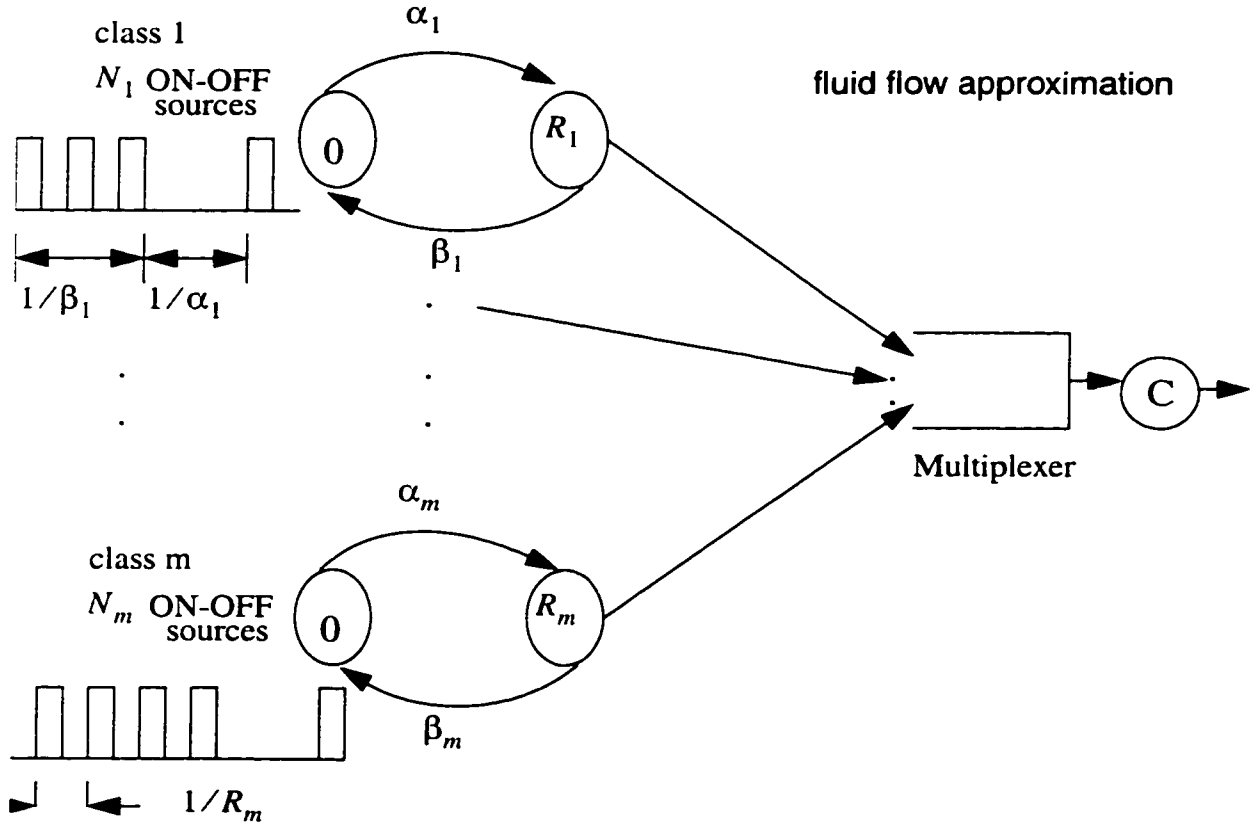

FIGURE.6.1. m-class ON-OF source mode

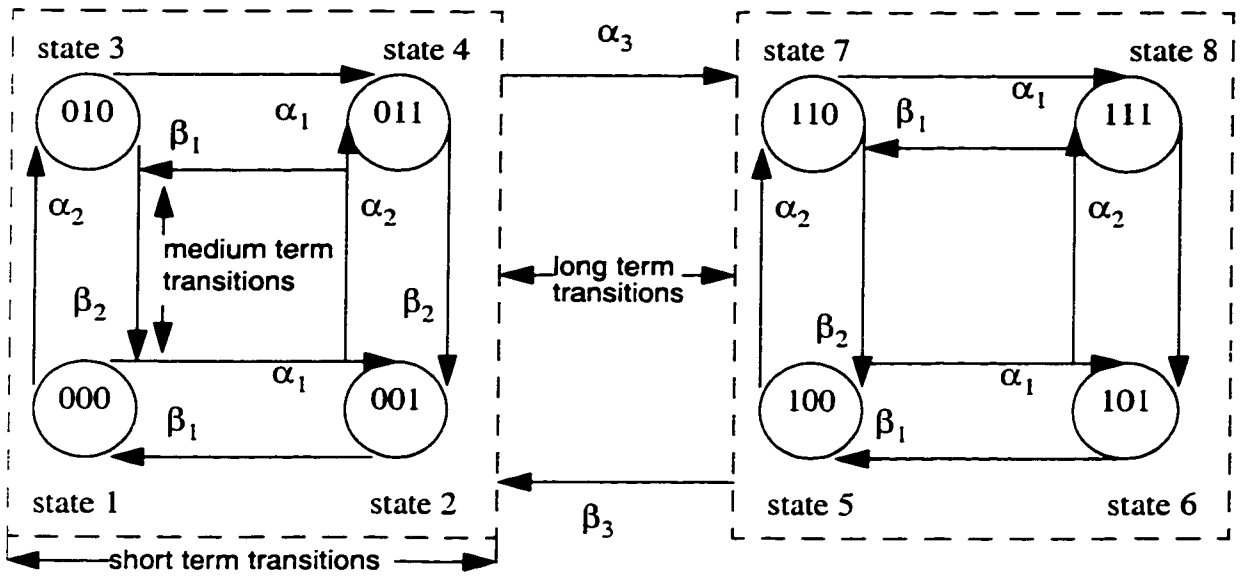

FIGURE.6.2. Three level Markov Chain

Using the state transition diagram, the 3 level Markov chain shown in figure 6.2 can be mapped into the simple three independent ON-OFF sources model shown in figure 6.3. This can be verified by finding out the infinitesimal generator matrix from figure 6.2 and that from figure 6.3.

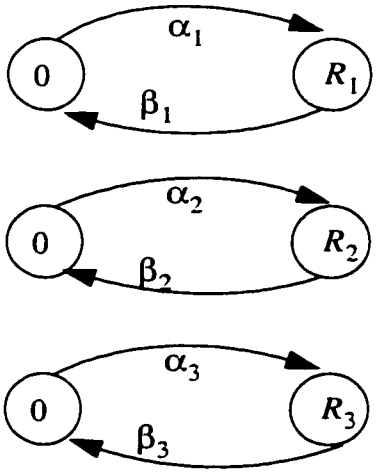

FIGURE.6.3. Three independent ON-OFF sources

Let $[n_1 n_2 \ldots n_m; \mathrm{u}]$ be the state with n_i source in class i ON and the buffer content does not exceed u and $p_{n_1 n_2 \ldots n_m}(u)$ be its equilibrium probability. Pack-

ets are served at a rate of C packets per time unit. We utilize the fluid flow approximation [ANI82], which has shown much promise in the analysis of ATM networks. Similarly to [ANI82], and as we have presented in section 3.2.1.5, we have the following:

$$\left(\sum_{i=1}^{m} R_i n_i - C\right)\frac{dp_{\dots\dots}(u)}{du} = \sum_{i=1}^{m} [\alpha_i(N_i - n_i + 1)]p_{\dots n_i - 1 \dots}(u) - \{\alpha_i(N_i - n_i) + \beta_i n_i\}p_{\dots\dots}(u) + \beta_i(n_i + 1)p_{\dots n_i + 1 \dots}(u) \quad (6.1).$$

We express equation (6.1) in the following familiar matrix form,

$$D\frac{dp(u)}{du} = p(u)M \quad (6.2).$$

where D is an $(N_1 + 1) \times \dots\dots \times (N_m + 1)$ diagonal matrix, M is $(N_1 + 1) \times \dots\dots \times (N_m + 1)$ infinitesimal generator matrix and $p(u)$ is a vector equal to $[p_{00\dots0}(x), \dots\dots\dots\dots, p_{n_1 n_2 \dots\dots n_m}(u)]$.

In the next section, we introduce the covariance function of equation (6.2), which will be used to derive the parameters that characterize the independent m class ON-OFF sources. That is, the parameters determination is based on second order statistics. Moreover, the importance of equation (6.2), when time comes into play, will appear when we present our analysis for congestion control in chapter 7.

6.3 Model parameter determination

Our work is based on finding the total covariance of the independent m classes $N_1, N_2, \dots\dots, N_m$ heterogeneous ON-OFF source model. Then by matching to the real data we find out the parameters that characterize the ON-OFF sources by adapting the Feldman algorithm to the fitting of the covariance [FEL97]. We may view this algorithm as analogous to Gram-Schmidth orthogonalization over

the time axis. The goal is to approximating a long-tail covariance distribution by a finite mixture of exponentials over shorter time scales. That is, we approximate a non-exponential function with a sum of exponential terms that we can easily deal with. The quality of the approximation is based on goodness of fit of the approximation by comparing the covariance function of the model with that of the data.

The covariance CO of the number of packets of a long-tail process as a function of the lag k and the Hurst parameter H behaves asymptotically as [COX84, LEL94]:

$$CO(k) \sim k^{2H-2} \tag{6.3}$$

The covariance function given by equation (6.3) decays hyperbolically (obeying some power law) as the lag k increases rather than exponentially, where k is the lag and H is the Hurst parameter.

The covariance $COV(\tau)$ of independent m class ON-OFF sources described by equation (6.2) is simply given by:

$$COV(\tau) = \sum_{i=1}^{m} \alpha_i \beta_i N_i \frac{R_i^2}{(\alpha_i + \beta_i)^2} e^{-(\alpha_i + \beta_i)\tau} \tag{6.4}$$

Applying the additivity property we find for the mean μ,

$$\mu = \sum_{i=1}^{m} \frac{\alpha_i N_i R_i}{(\alpha_i + \beta_i)} \tag{6.5}$$

and for the variance Var,

$$Var = \sum_{i=1}^{m} \frac{\alpha_i \beta_i N_i R_i^2}{(\alpha_i + \beta_i)^2} \tag{6.6}$$

Let,

$$\lambda_i = \alpha_i + \beta_i, \qquad i = 1, 2, \ldots\ldots m \tag{6.7}$$

and

$$K_i = \alpha_i \beta_i N_i \frac{R_i^2}{\lambda_i^2}, \qquad i = 1, 2, \ldots\ldots m \tag{6.8}$$

Substitute equation (6.7) and equation (6.8) in to equation (6.4) and assume that frames are generated at rate of f frames /sec:

$$COV(\tau) = K_1 e^{\frac{-\lambda_1 \tau}{f}} + K_2 e^{\frac{-\lambda_2 \tau}{f}} + \ldots\ldots + K_m e^{\frac{-\lambda_m \tau}{f}} = \sum_{i=1}^{m} K_i e^{-\frac{\lambda_i \tau}{f}} \tag{6.9}$$

Equation (6.9) is a finite mixture of exponentials that approximate the long-tail distribution function given by equation (6.3). The idea is to approximate equation (6.3) by equation (6.9), because performance models with component long-tail distributions tend to be difficult to analyze [NOR94, PRU95].

As can be seen from equation (6.9) we have $2m$ unknowns and therefore we need $2m$ equations to find them. Since the covariance is composed of m exponential components $\lambda_1, \lambda_2, \ldots\ldots, \lambda_m$, and m arguments $K_1, K_2, \ldots\ldots K_m$, we match at the quantiles: $0 < c_1 < c_2 < \ldots\ldots < c_m$, which represent how many classes that we have. For example, for two classes we have two quantiles c_1, c_2 and for three classes we have three quantiles c_1, c_2, c_3 and so on. In order to solve $2m$ equations to find the $2m$ unknowns, let b be a scaling factor such that $1 < b < \frac{c_{i+1}}{c_i}$ for all i; e.g., we could have $b = 4, c_i = 10^{(i-1)} c_1$ for $2 \le i \le m$. $\frac{c_2}{c_1} = \frac{c_3}{c_2} = \ldots = \frac{c_m}{c_{m-1}}$. See figure 6.4 for the three source case.

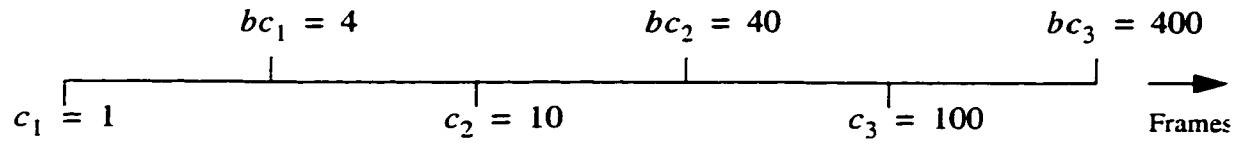

FIGURE.6.4. Illustration of how to choose the quantiles c_1, c_2, c_3 and the scaling factor b $(c_1 = 1, c_2 = 10, c_3 = 100, b = 4)$

Given the real data and using the technique in [FEL97] we can obtain the exponential components $\lambda_1, \lambda_2, ...,\lambda_m$ and the arguments $K_1, K_2, ...K_m$ in reverse order by finding first λ_m and K_m and then λ_{m-1} and K_{m-1} and so on until we find λ_1 and K_1. As mentioned above $\lambda_1 >> \lambda_2 >> .....\lambda_m$. Therefore, at the quantiles c_m and bc_m, only the terms of the covariance that have argument λ_m count and those terms of covariance that have arguments $\lambda_{m-1}, \lambda_{m-2}, ..., \lambda_1$ are negligibly small. Therefore,

$$COV(c_m) = K_m e^{\frac{-\lambda_m c_m}{f}} \qquad (6.10).$$

and

$$COV(bc_m) = K_m e^{\frac{-\lambda_m bc_m}{f}} \qquad (6.11).$$

From equations (6.10) and (6.11), we find the two unknowns, λ_m and K_m.

Now we proceed to find the other two unknowns λ_{m-1} and K_{m-1} at the quantiles c_{m-1} and bc_{m-1}. In this case only the terms of the covariance that have argument λ_{m-1} and λ_m count and the terms of the covariance that have arguments $\lambda_{m-2}, \lambda_{m-3}, ..., \lambda_1$ are negligibly small.

$$COV(c_{m-1}) = k_m e^{\frac{-\lambda_m c_{m-1}}{f}} + k_{m-1} e^{\frac{-\lambda_{m-1} c_{m-1}}{f}} \quad (6.12).$$

$$COV(bc_{m-1}) = k_m e^{\frac{-\lambda_m bc_{m-1}}{f}} + k_{m-1} e^{\frac{-\lambda_{m-1} bc_{m-1}}{f}} \quad (6.13).$$

where λ_m and K_m are already known from equations (6.10) and (6.11).

Given λ_m, k_m, λ_{m-1} and k_{m-1} we find the next two unknowns λ_{m-2} and k_{m-2} at the quantiles c_{m-2} and bc_{m-2},

$$COV(c_{m-2}) = k_m e^{\frac{-\lambda_m c_{m-2}}{f}} + k_{m-1} e^{\frac{-\lambda_{m-1} c_{m-2}}{f}} + k_{m-2} e^{\frac{-\lambda_{m-2} c_{m-2}}{f}} \quad (6.14).$$

$$COV(bc_{m-2}) = k_m e^{\frac{-\lambda_m bc_{m-2}}{f}} + k_{m-1} e^{\frac{-\lambda_{m-1} bc_{m-2}}{f}} + k_{m-2} e^{\frac{-\lambda_{m-2} bc_{m-2}}{f}} \quad (6.15).$$

and so on until we end up with the last two unknowns λ_1 and K_1.

The final step is to find the parameters that characterize the ON-OFF sources, i.e., $\alpha_1, \beta_1, R_1; \alpha_2, \beta_2, R_2; .....; \alpha_m, \beta_m, R_m$. We have a system of $3m$ (where m is the number of classes and the factor 3 comes from the fact that each source has 3 parameters to be determined) parameters to be calculated, however we have in hand only $2m$ known factors $(\lambda_1, k_1; \lambda_2, k_2; .....; \lambda_m, k_m)$. The basic property of our model is given in section 6.2, where the transitions rates are assumed to be such that $\alpha_1 >> \alpha_2 >> .....\alpha_m$ and $\beta_1 >> \beta_2 >> .....\beta_m$ (already from 6.7, we have $\lambda_1 >> \lambda_2 >> .....\lambda_m$). We use this assumption in such a way that we have fewer unknowns to evaluate.

Let* $\alpha_i = 10^{-(i-1)}\alpha_1, \qquad i = 2, ...m$

From (6.8) we find R_i in terms of K_i, α_i and β_i,

$$R_i = \sqrt{\frac{\lambda_i^2 K_i}{\alpha_i \beta_i N_i}} \tag{6.16}$$

Substitute for R_i , β_i ($\beta_i = \lambda_i - \alpha_i$) in (6.5) we have, after some manipulation, the following,

$$\sqrt{\frac{\alpha_1 N_1 k_1}{(\lambda_1 - \alpha_1)}} + \sqrt{\frac{\alpha_2 N_2 k_2}{(\lambda_2 - \alpha_2)}} + ... + \sqrt{\frac{\alpha_i N_i k_i}{(\lambda_i - \alpha_1)}} + ... + \sqrt{\frac{\alpha_m N_m k_m}{(\lambda_m - \alpha_m)}} = \mu \tag{6.17}$$

Since $\alpha_i = 10^{-(i-1)}\alpha_1$, equation (6.17) can be written in the following form:

$$\sqrt{\frac{\alpha_1 N_1 k_1}{(\lambda_1 - \alpha_1)}} + \sqrt{\frac{\alpha_1 N_2 k_2}{(10\lambda_2 - \alpha_1)}} + ... + \sqrt{\frac{\alpha_1 N_i k_i}{(10^{i-1}\lambda_i - \alpha_1)}} + ...+ \sqrt{\frac{\alpha_1 N_m k_m}{(10^{m-1}\lambda_m - \alpha_1)}} \tag{6.18}$$

The number of sources N_1, N_2, ...,N_m is given in advance. Also, we know the values of λ_1, λ_2, ...λ_m, K_1, K_2, ... K_m from matching to the data, and also we know μ the estimated mean value of the real data. Therefore, the non-linear equation (6.18), which is a function of only one unknown parameter α_1, can be solved numerically. Knowing α_1, the parameters β_1, R_1; α_2, β_2, R_2;.....; α_m, β_m, R_m can be obtained very easily using equations (6.7) and (6.8).

6.4 Numerical results

The model can be applied to any number of classes and any number of sources per class, however, as the number of sources increases the solution of equation (6.18) becomes more difficult. Because of this, we apply the model to the three classes and one source per class. Using the video data presented in section 2.3, we calculate the covariance function for the real data and apply the procedure

* *We also assumed* $\alpha_1/\beta_1 = \alpha_2/\beta_2 = ...=\alpha_m/\beta_m$ *and similar results are obtained*

presented in section 6.3 to find the parameters for the heterogeneous ON-OFF source model. These three heterogeneous ON-OFF source models are used to generate video traces in OPNET program. We generate almost the same number of real video data frames of approximately 50,000 frames for video-conferencing, video-phone, TV series and Movie video sequences. From the real traffic we find the covariance function, IDC, probability of loss and mean queue length and compare them with that of the generated traffic. Also, as a reference, we calculate covariance, IDC, probability of loss and mean queue length based on the Maglaris model with 20 minisources [MAG88].

The values of the parameters depends on the quantiles c_i's, the scaling factor b and the number of frames over which the matching is going to be done. As the number of quantiles, the scaling factor and the number of frames over which the matching is going to be done increases the accuracy will be increased. However, this is not always possible since increasing the number of quantiles means increasing the number of classes (for the three class model, three quantiles are needed), which makes the solution of equation (6.18) more difficult.

Tables 6.1 - 6.8 show, respectively, the estimated parameters α_i, β_i, R_i for two and three class single ON-OFF sources needed to model the video-conferencing, video-phone, TV series, and Movie data that we presented in chapter 2. They also show the estimated exponential components λ_i's and the arguments K_i's. Given the parameters α_i's, β_i's and R_i's for ON-OFF sources for the VBR video traffic traces, a replica of the traffic is generated using OPNET. From the real traffic we find out the covariance function, IDC, probability of loss and mean

queue length and compare them with that of the generated traffic. We also plot the covariance based on equation (6.9).

TABLE 6. 1 Parameters for the two class single ON-OFF sources matched to the video-conferencing trace over 768 frames with $c_1 = 1, c_2 = 192$ and $b = 4$

Parameter	λ	k	α	β	R
source 1	0.588	5034.2	0.348	0.24	144.43
source 2	0.046	620.78	0.0348	0.011	58.68

TABLE 6. 2 Parameters for the three class single ON-OFF sources matched to the video-conferencing trace over 768 frames with $c_1 = 1, c_2 = 16, c_3 = 256$ and $b = 3$

Parameter	λ	k	α	β	R
source 1	2.156	401.69	1.79	0.36	53.81
source 2	0.39	4815.4	0.179	0.22	139.38
source 3	0.03	429.10	0.0179	0.016	41.55

TABLE 6. 3 Parameters for the two class single ON-OFF sources matched to the video-phone trace over 768 frames with $c_1 = 1, c_2 = 192$ and $b = 4$

Parameter	λ	k	α	β	R
source 1	0.94	9346.0	0.494	0.445	193.599
source 2	0.08	2485.7	0.0494	0.026	104.953

TABLE 6. 4 Parameters for the three class single ON-OFF sources matched to the video-phone trace over 768 frames with $c_1 = 1, c_2 = 16, c_3 = 256$ and $b = 3$

Parameter	λ	k	α	β	R
source 1	2.53	1883.4	1.75	0.779	94.035
source 2	0.42	8883.4	0.175	0.24	190.832
source 3	0.05	1042.6	0.0175	0.029	66.702

TABLE 6. 5 Parameters for the two class single ON-OFF sources matched to the TV series video trace over 768 frames with $c_1 = 1, c_2 = 192$ and $b = 4$

Parameter	λ	k	α	β	R
source 1	3.033	746110.0	0.305	2.99	2980.9
source 2	0.031	700350.0	0.030	0.0008	5199.0

TABLE 6. 6 Parameters for the three class single ON-OFF sources matched to the TV series video trace over 768 frames with $c_1 = 1, c_2 = 16, c_3 = 256$ and $b = 3$

Parameter	λ	k	α	β	R
source 1	16.47	461150.0	2.63	13.84	1853.4
source 2	0.408	505950.0	0.263	0.1451	1486.0
source 3	0.027	617570.0	0.0264	0.00097	4233.4

TABLE 6. 7 Parameters for the two class single ON-OFF sources matched to the Movie video trace over 768 frames with $c_1 = 1, c_2 = 192$ and $b = 4$

Parameter	λ	k	α	β	R
source 1	2.8733	817700.0	0.08099	2.7923	5463.4
source 2	0.0083	830860.0	0.008099	0.000201	5937.7

TABLE 6. 8 Parameters for the three class single ON-OFF sources matched to the Movie video trace over 768 frames with $c_1 = 1, c_2 = 16, c_3 = 256$ and $b = 3$

Parameter	λ	k	α	β	R
source 1	21.53	469380.0	0.60	20.93	4147.2
source 2	0.46	592920.0	0.060	0.40	2282.9
source 3	0.006	779260.0	0.0060	0.00015	5674.0

As a reference, we compare the statistical and performance measures, covariance, IDC, probability of loss and mean queue length of the real video data and the generated video traffic based on heterogeneous ON-OFF source model with that of the generated traffic based on Maglaris model. To do that we need to find the parameters that characterize the Maglaris model. Using equations (8), (9), (10) and (11) of [MAG88] we calculated the parameters of the Maglaris model for the four video traces. These are shown in Table 6.9 below.

TABLE 6. 9 Maglaris model parameters for VBR traces, video-conferencing, video-phone, TV series, and Movie. Each video source is characterized by 20 minisources.

Parameter	α	β	R
video-conferencing	0.05	0.30	39.00
video-phone	0.04	0.30	75.81
TV series	0.07	0.06	520.41
Movie	0.05	0.05	559.59

6.4.1 Covariance and IDC

In this section we consider the matching of the covariance and the IDC of the generated traffic to the real data using heterogeneous sources. As a reference, we have generated traffic based on the Maglaris model of 20 minisources [MAG88]. We compare the accuracy of our model with that based on the Maglaris model. Also, we consider the effect of increasing the number of ON-OFF heterogeneous sources to the accuracy of the matching. Given the exponential components λ_i's and the arguments K_i's of the heterogeneous ON-OFF source model, we see how well the covariance represented by equation (6.9) matches the covariance of the real data and that based on the Maglaris model.

The covariances of the real video-conferencing and the generated 3 and 2 class single ON-OFF source are shown in figures 6.5. We also plot the covariance based on Magalris model of 20 minisources. In figure 6.6, we plot the covariance of real video-conferencing and that based on the formula given by equation (6.9) and that given by Maglaris of one exponential term, which is given by equation (5) [MAG88]. We do the same for video-phone data which, are shown in figures 6.7 and figure 6.8, respectively. As expected, the accuracy increases as the number of classes increases. Moreover, in comparison with the Maglaris model, the matching based on the 3 class single ON-OFF source is shown to be better. For both video telconferencing data, the match to the traces based on the generated traffic and that using formula (6.9) is quite good over a large number of lags.

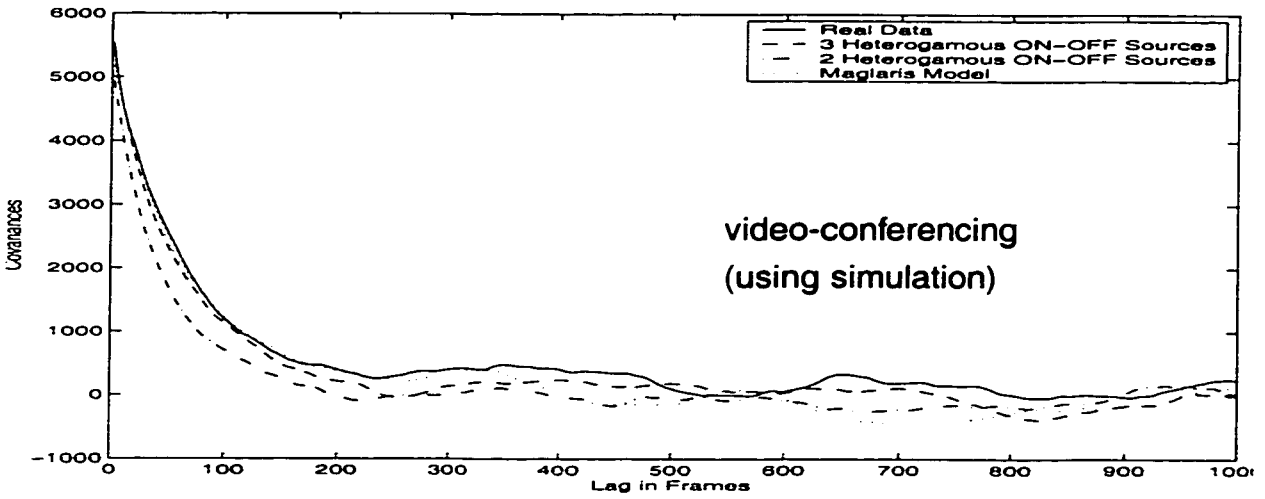

FIGURE.6.5. Covariances functions of real video-conferencing data compared with that of Maglaris and generated 3 and 2 class heterogeneous source model, each class has 1 ON-OFF source.

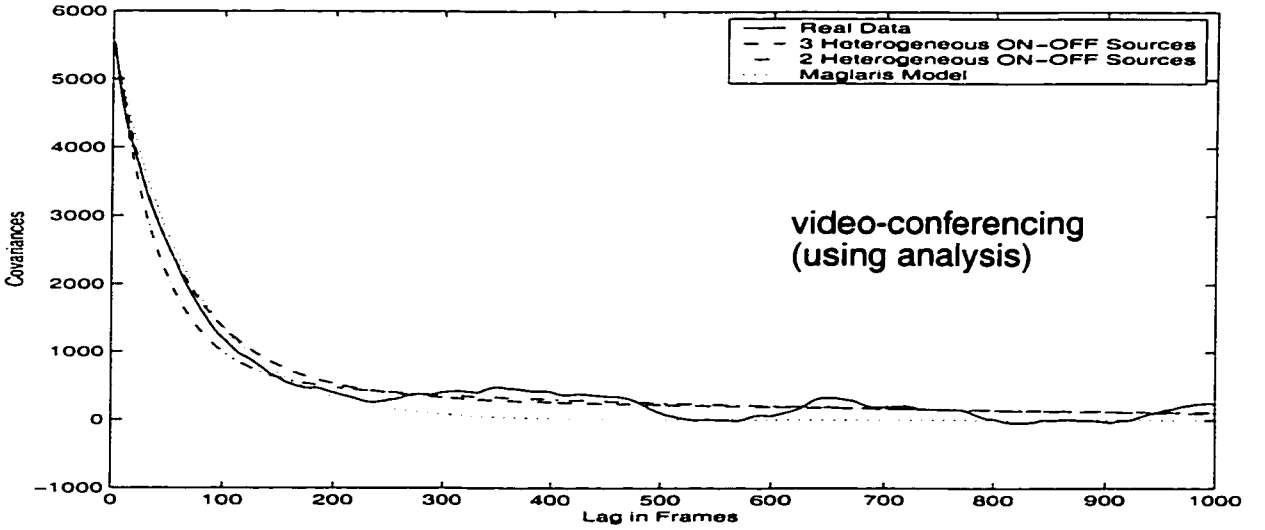

FIGURE.6.6. Covariance functions of real video-conferencing data, and using formula (6.9) for 3 and 2 class heterogeneous source model, each class has 1 ON-OFF source

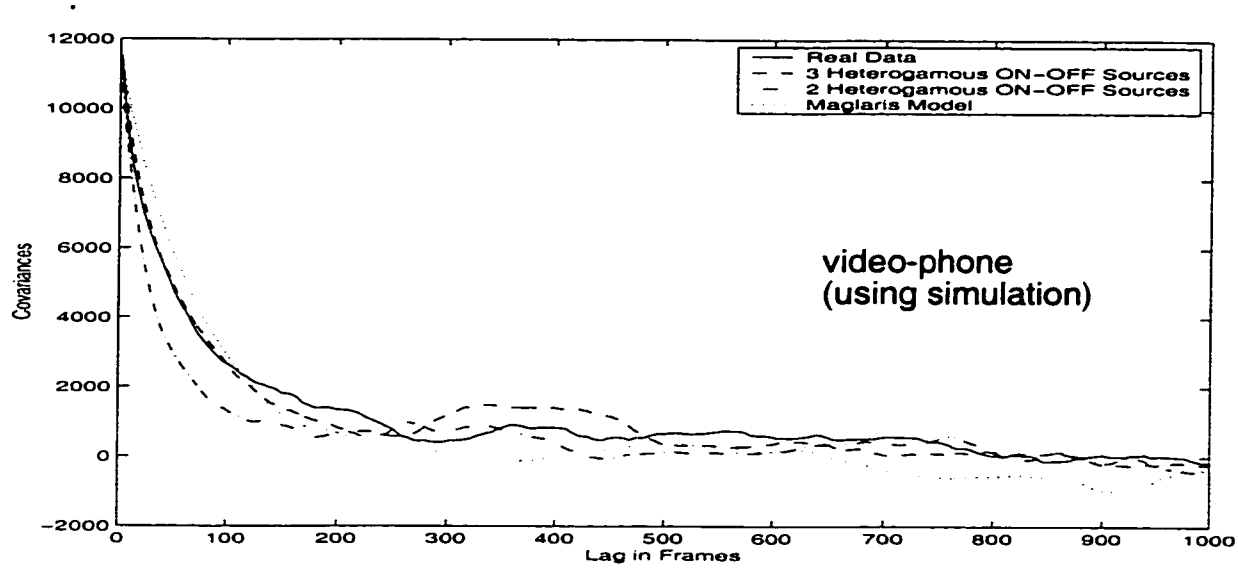

FIGURE.6.7. Covariances of real video-phone data compared with that of Maglaris, and generated 3 and 2 class heterogeneous source model, each class has 1 ON-OFF source.

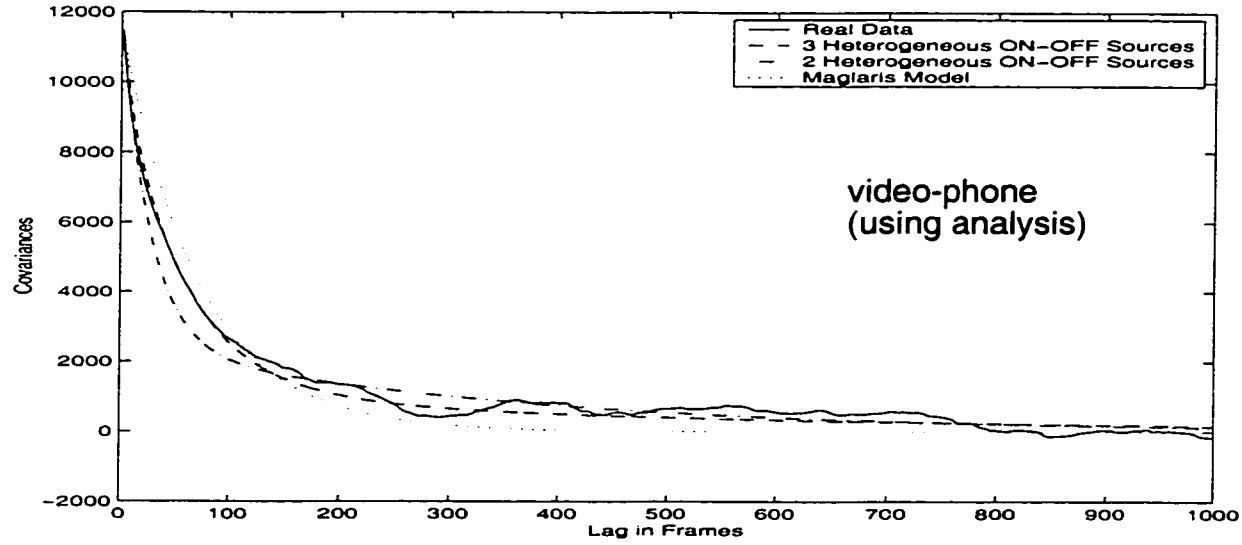

FIGURE.6.8. Comparison of the covariances of real video-phone data, and using formula (6.9) for 3 and 2 class heterogeneous source model, each class has 1 ON-OFF source.

The covariances for the highly correlated traffic real TV series and Movie and that of the 3 and 2 class single ON-OFF source are shown in figure 6.9 and figure 6.11, respectively. Also in the figures we plot the covariance based on Maglaris model. The results shows that our model has better matching of the covariances

than that based on Maglaris model. However, the matching for this kind of entertainment traffic is less accurate than that for the teleconferencing traffic shown in figure 6.5 and figure 6.7. This due to the high value of the correlation index H.

In figure 6.10 and figure 6.12, we plot the covariance of real sequences TV series and Movie, respectively, and that based on the formula given by equation (6.9) and that given by Maglaris of one exponential term, which is given by equation (5) [MAG88]. The matching for both sequences and that based on formula (6.9) is reasonable. The matching of the Maglaris covariance given by equation (5) of [MAG88] and the two real sequences TV series and Movie is reasonable when the lag is not large. As the lag increases the deviation of the covariance of the real data from that of the covariances given by formula (6.9) and that given by Maglaris of one exponential term given by equation (5) [MAG88] becomes clear.

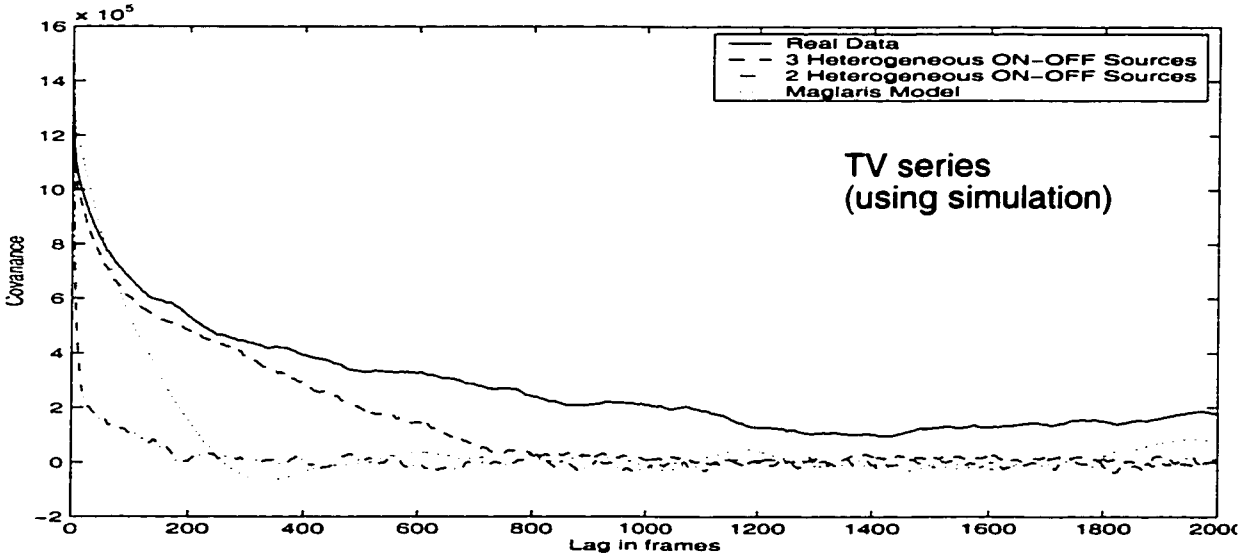

FIGURE.6.9. Covariances of real TV series data compared with that of Maglaris, and generated 3 class heterogeneous source model, each class has 1 ON-OFF source

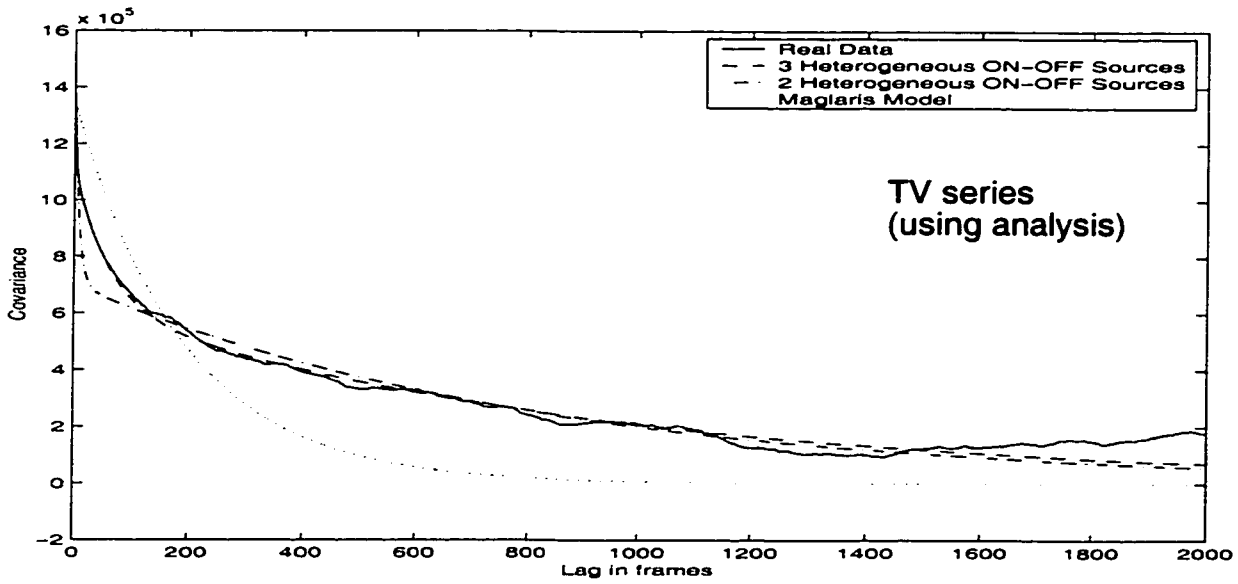

FIGURE.6.10. Comparison of the covariances of real TV series data, and using formula (6.9) for 3 and 2 class heterogeneous source model, each class has 1 ON-OFF source.

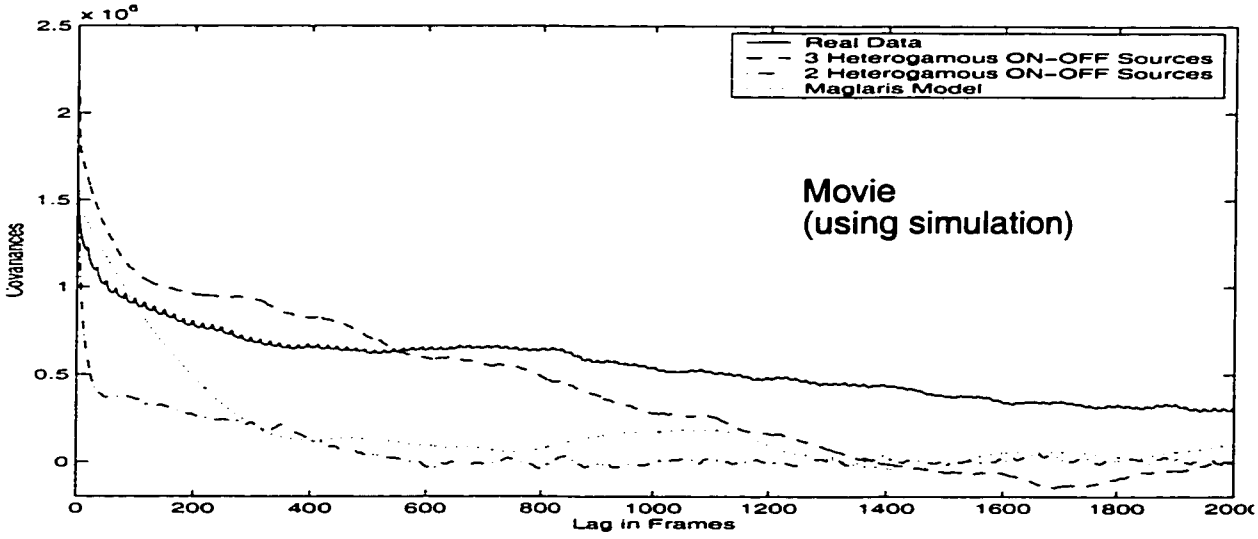

FIGURE.6.11. Covariances of real Movie data compared with that of Maglaris, and generated 3 and 2 class heterogeneous source model, each class has 1 ON-OFF source

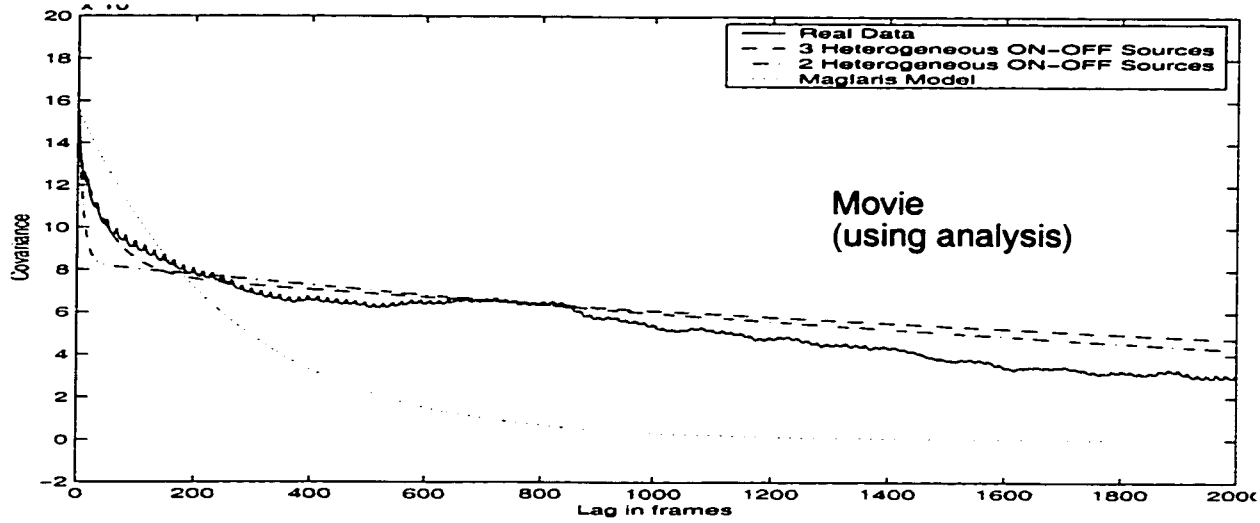

FIGURE.6.12. Comparison of the covariances of real Movie data, and using formula (6.9) for 3 and 2 class heterogeneous source model, each class has 1 ON-OFF source.

The other part of matching the heterogeneous sources to the real data is the IDC. As shown in figure 6.13 and figure 6.14, respectively, the IDC of the synthetic video-conferencing and video-phone traffic matches the IDC of the real data. As for the covariance, increasing the number of heterogeneous sources from 2 to 3 will increase the accuracy of the matching. We also show the IDC based on the Maglaris model.

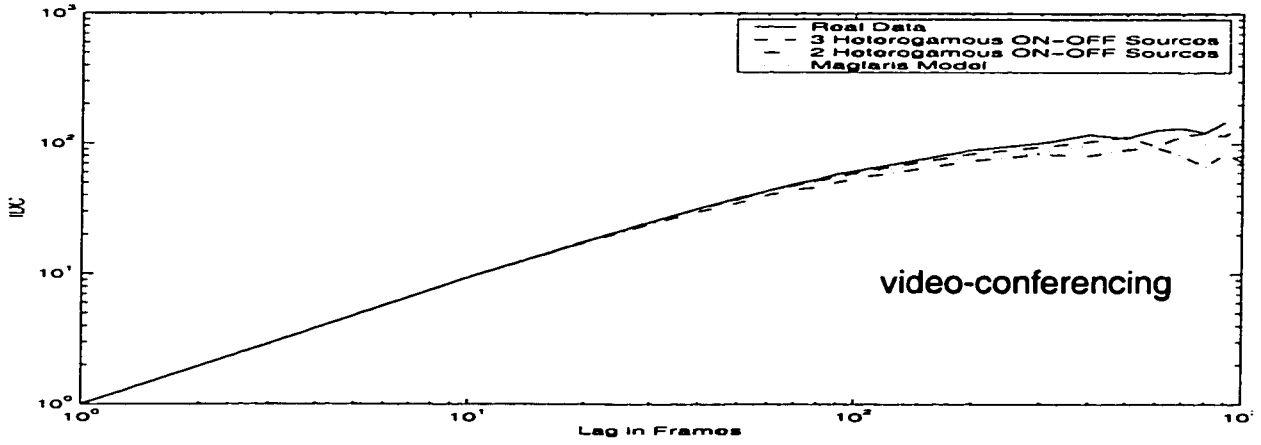

FIGURE.6.13. IDC of real video-conferencing data compared with that of Maglaris and generated 3 and 2class heterogeneous source model, each class has 1 ON-OFF source.

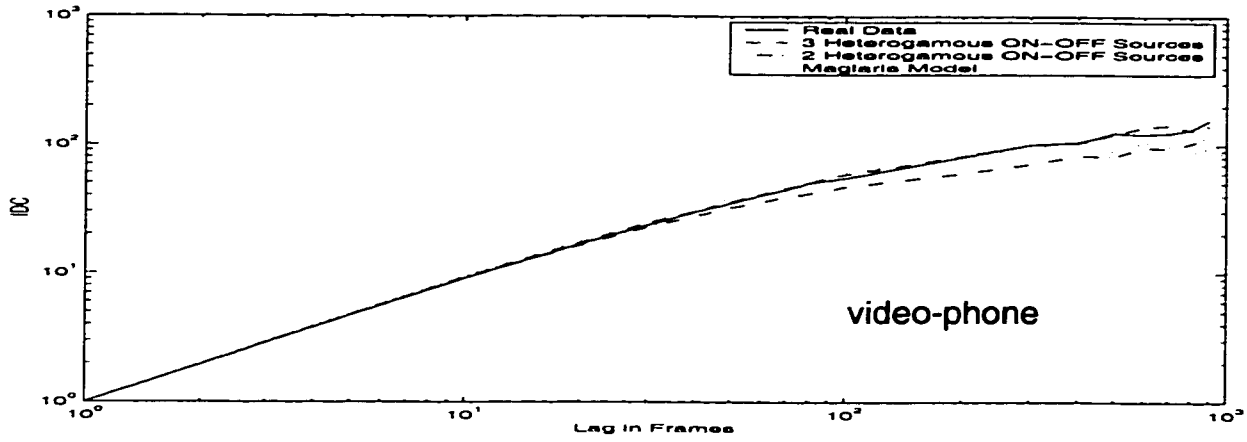

FIGURE.6.14. *IDC* of real video-phone data compared with that of Maglaris, and generated 3 and 2 class heterogeneous source model, each class has 1 ON-OFF source

The *IDC* for the entertainment video data, TV series and Movie, are shown in figure 6.15 and figure 6.16, respectively. The figures show the results for real data, three heterogeneous ON-OFF source model, two heterogeneous ON-OFF source model and that of Maglaris model. As the number of heterogeneous ON-OFF sources increases, the accuracy of the matching becomes more accurate. Moreover, the matching for the three heterogeneous ON-OFF source model performs better than that of the Maglaris model.

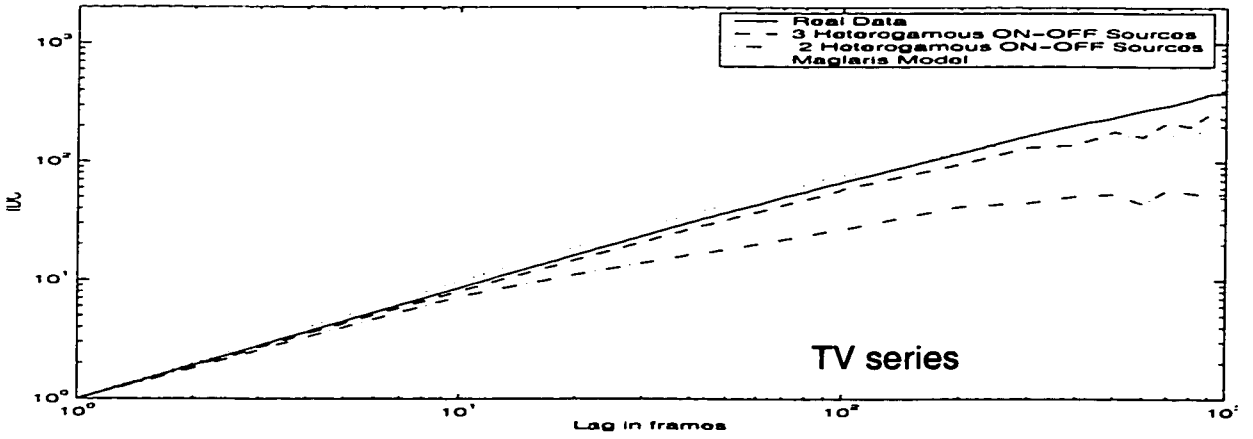

FIGURE.6.15. *IDC* of real TV series data compared with that of Maglaris, and generated 3 and 2 class heterogeneous source model, each class has 1 ON-OFF source

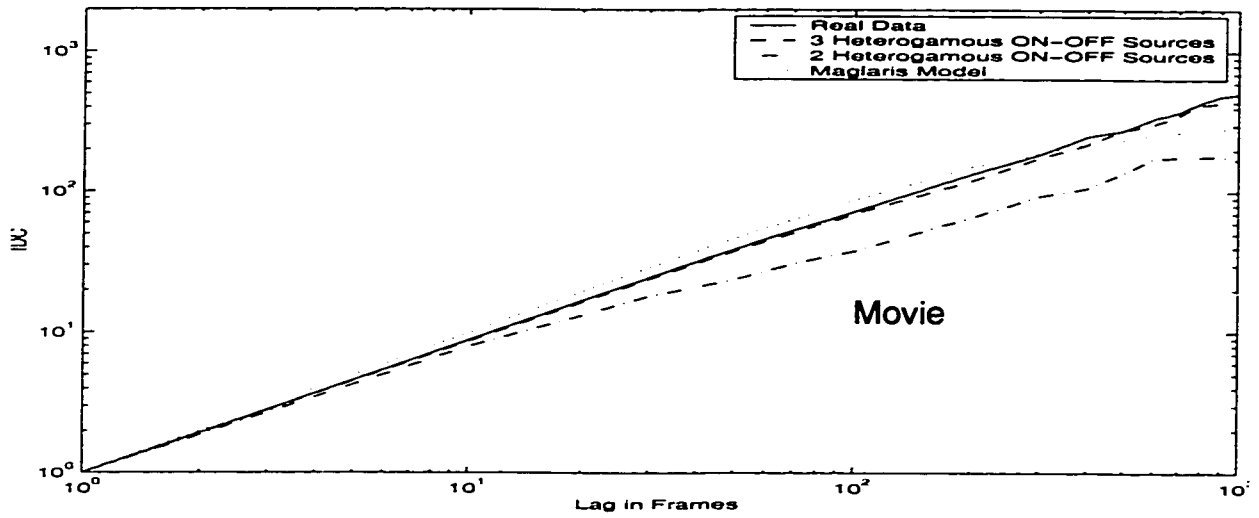

FIGURE.6.16. *IDC* of real Movie data compared with that of Maglaris, and generated 3 and 2 class heterogeneous source model, each class has 1 ON-OFF source

6.4.2 Performance analysis

6.4.2.1 Probability of loss

As can be seen from figure 6.17 and figure 6.18, the probability of loss versus the traffic intensity for buffer capacity of 150 cells for the medium correlated traffic video-conferencing and video-phone data and that obtained from the traffic generated by 3-heterogeneous ON-OFF source model in good agreement. We show also the probability of loss based on Maglaris model of 20 minsources. The 3-heterogeneous ON-OFF source model performs better than the Maglaris model. The prediction becomes more accurate when the number of heterogeneous ON-OFF sources increases. However, going for a larger number of sources will make the analysis more complicated, that is, we will have a larger number of equations to

solve and more parameters to find and the model will lose the advantage of being simple and tractable. The prediction is satisfactory for practical engineering design

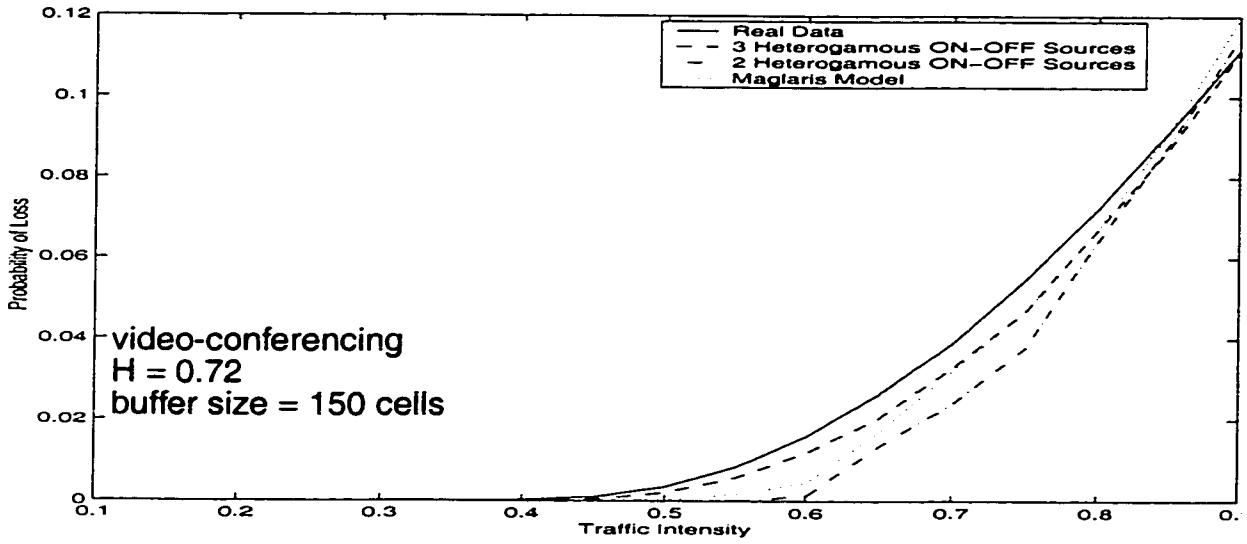

FIGURE.6.17. Probability of loss of real video-conferencing data compared with that of Maglaris and generated 3 and 2 class heterogeneous source model, each class has 1 ON-OFF source.

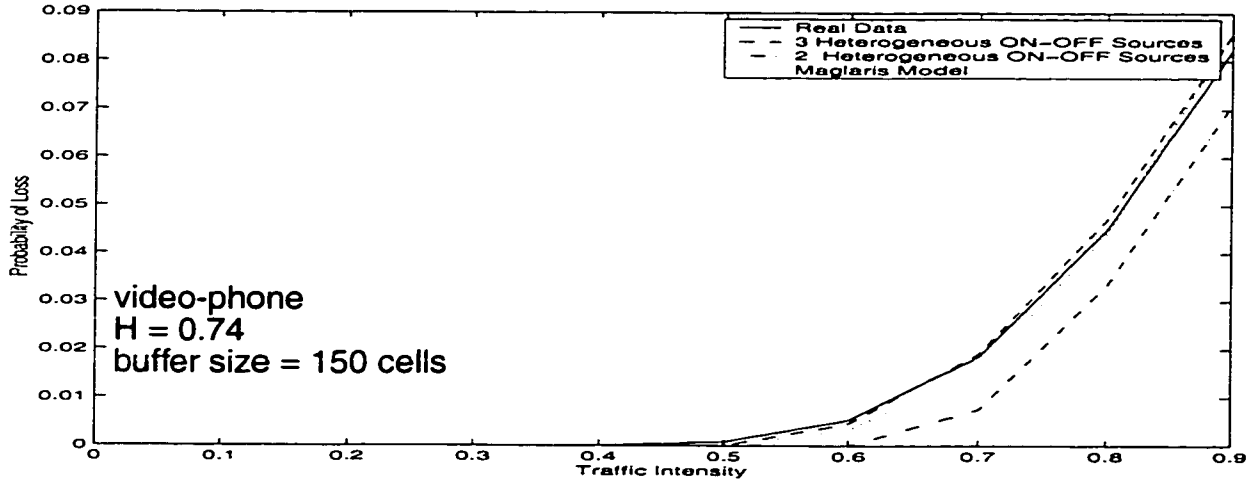

FIGURE.6.18. Probability of loss of real video-phone data compared with that of Maglaris and generated 3 and 2 class heterogeneous source model, each class has 1 ON-OFF source.

The probabilities of loss for the sequences TV series and Movie are shown in figure 6.19 and figure 6.20, respectively. The effect of increasing the number of heterogeneous sources from 2 to 3 is clear. Also, as can be seen, there is an

improvement in the accuracy of the prediction of the probability of loss as compared to that based on the Maglaris model.

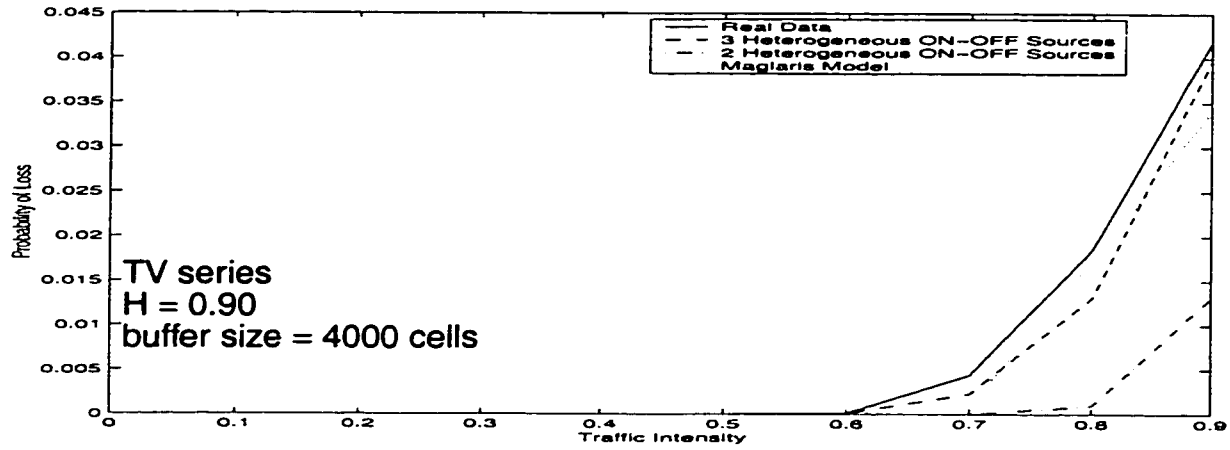

FIGURE.6.19. Probability of loss of real TV series data compared with that of Maglaris and generated 3 class heterogeneous source model, each class has 1 ON-OFF source.

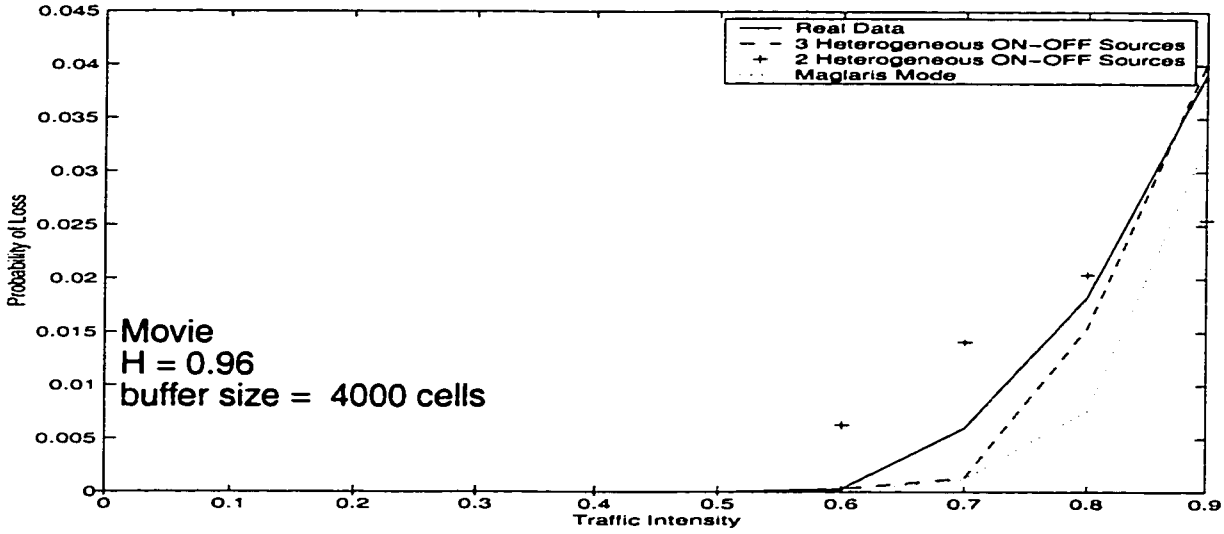

FIGURE.6.20. Probability of loss of real Movie data compared with that of Maglaris and generated 3 class heterogeneous source model, each class has 1 ON-OFF source.

6.4.2.2 Mean queue length

In figure 6.21 and figure 6.22, we show the mean queue length as a function of traffic intensity for video-conferencing and video-phone data and that based on 3 and 2 heterogeneous ON-OFF source model and the Maglaris model. The

buffer capacity is set to a fixed value of 150 cells. The prediction of the mean queue length based on the 3-heterogeneous ON-OFF source model is better than the prediction based on the Maglaris model. As shown, increasing the number of heterogeneous ON-OFF sources from 2 to 3 will increase the accuracy of the prediction.

In figure 6.23 and figure 6.24, the mean queue length as a function of the traffic intensity for the entertainment video data, TV series and Movie, are shown. The prediction is not as good as that for the telconferencing data, however, for practical engineering design the results are satisfactory. Moreover, as for the telconferencing video data, increasing the number of heterogeneous ON-OFF sources will improve the prediction. In comparison with the results of the Maglaris model, there is an improvement in the prediction of the queue length for both video teleconferencing and video entertainment data.

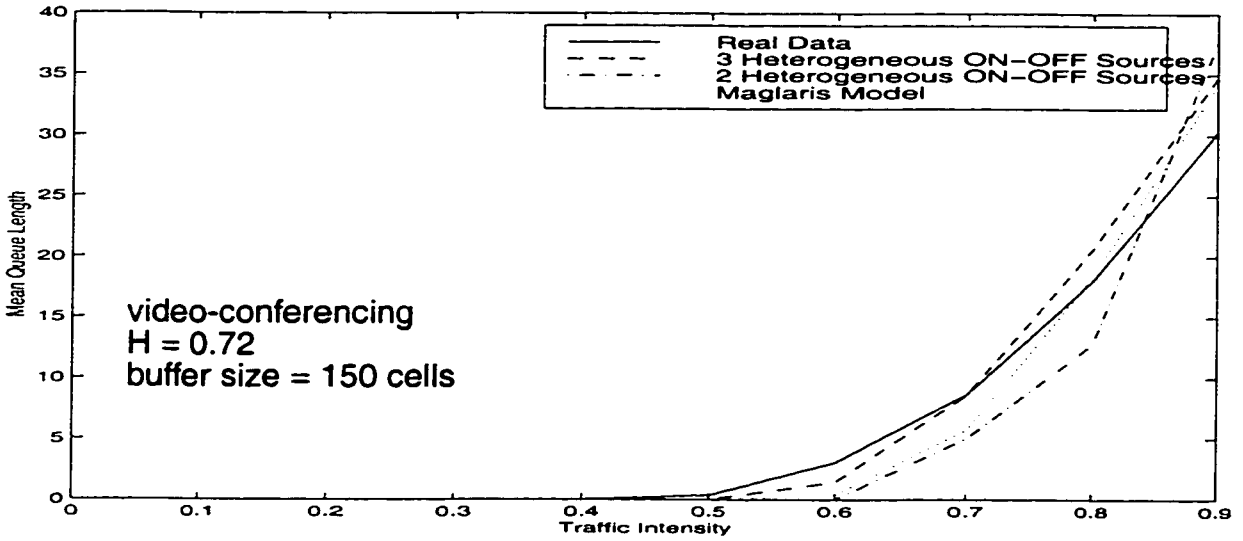

FIGURE.6.21. Mean queue length of real video-conferencing data compared with that of Maglaris and generated 3 and 2 class heterogeneous source model, each class has 1 ON-OFF source.

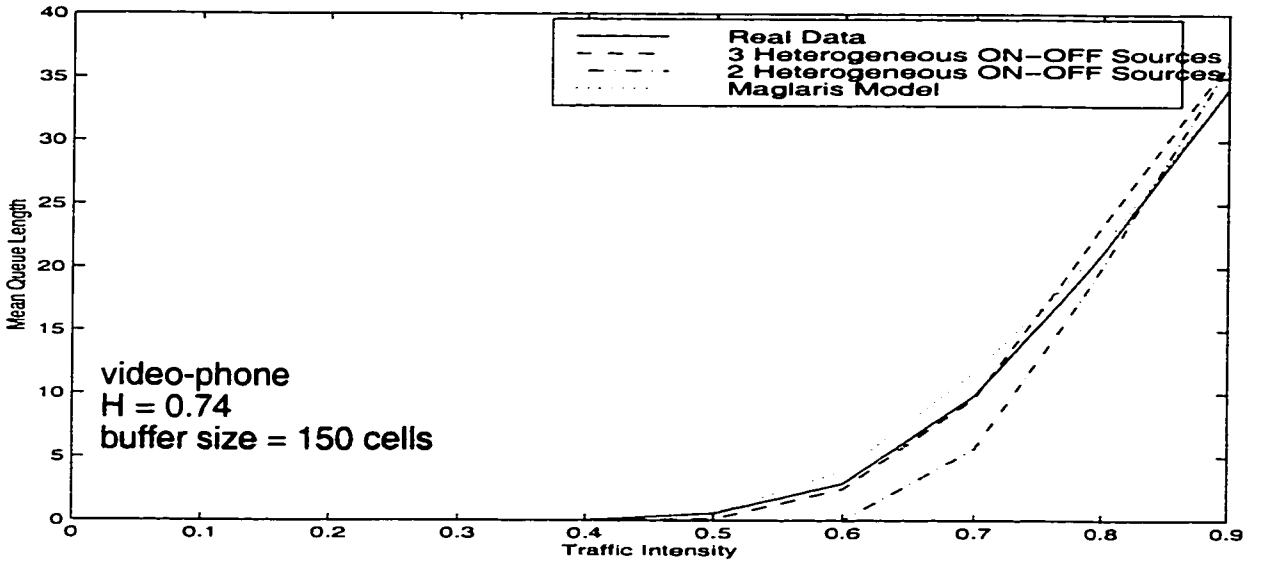

FIGURE.6.22. Mean queue length of real video-phone data compared with that of Maglaris and generated 3 and 2 class heterogeneous source model, each class has 1 ON-OFF source.

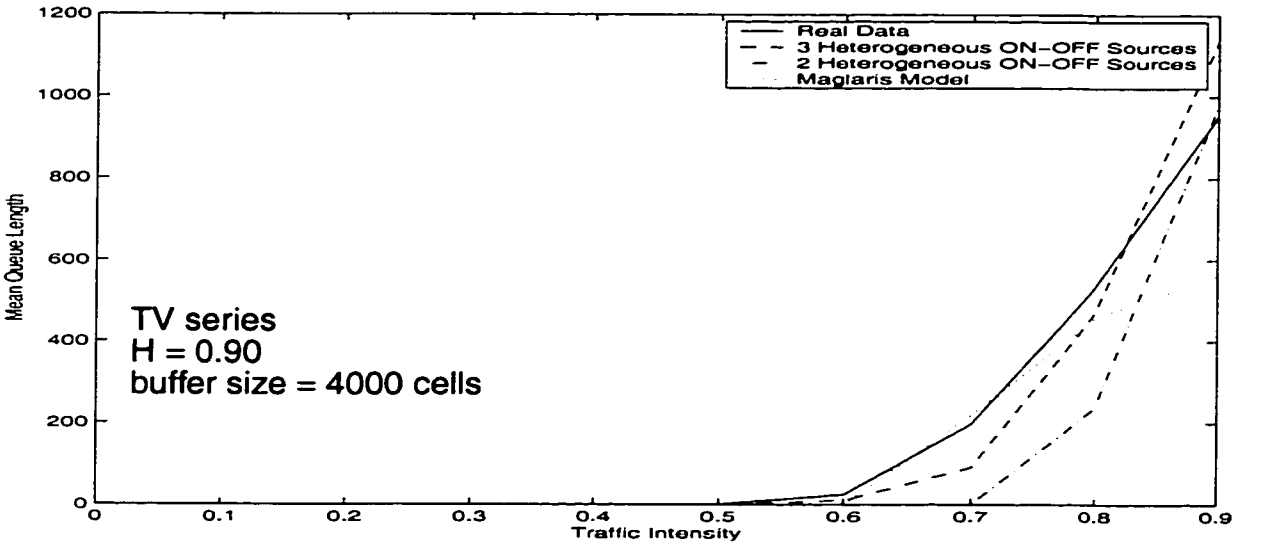

FIGURE.6.23. Mean queue length of TV series data compared with that of Maglaris and generated 3 and 2 class heterogeneous source model, each class has 1 ON-OFF source.

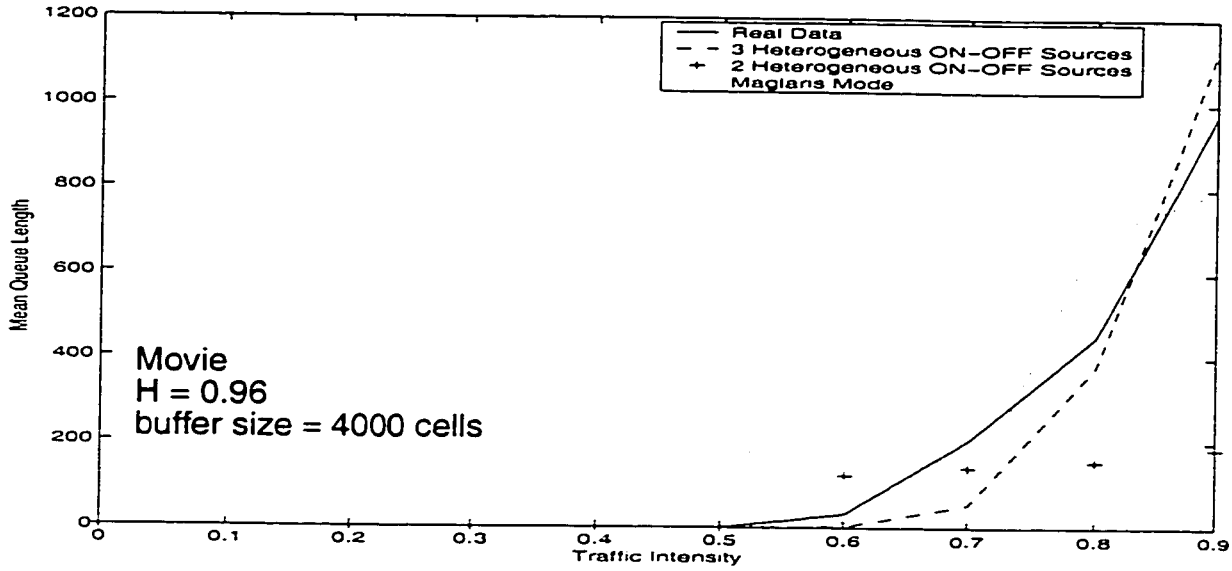

FIGURE.6.24. Mean queue length of real Movie data compared with that of Maglaris and generated 3 and 2 class heterogeneous source model, each class has 1 ON-OFF source.

6.5 Modeling multiplexed sources

As we have presented in the previous chapters, multiplexing several statistically independent and identical highly correlated sources will result in a good prediction and reduction of the probability of loss and mean queue length. With this model, the advantage of multiplexing is due to smoothing the traffic as a result of averaging. We present multiplexing of one medium correlated traffic and another highly correlated traffic; namely video-phone and Movie since the other two video data, video-conferencing and TV series have similar results to those we discuss.

As shown in figure 6.25 for video-phone, there is an improvement in multiplexing a number of these video sources. The matching of the generated 3-heterogeneous ON-OFF source model traffic probability of loss and mean queue length to that of the data when N = 4 sources is quite good as compared to the case when

N =1 source. Also, we can see the effect of multiplexing on the reduction of the probability of loss and the mean queue length for video-phone sequences.

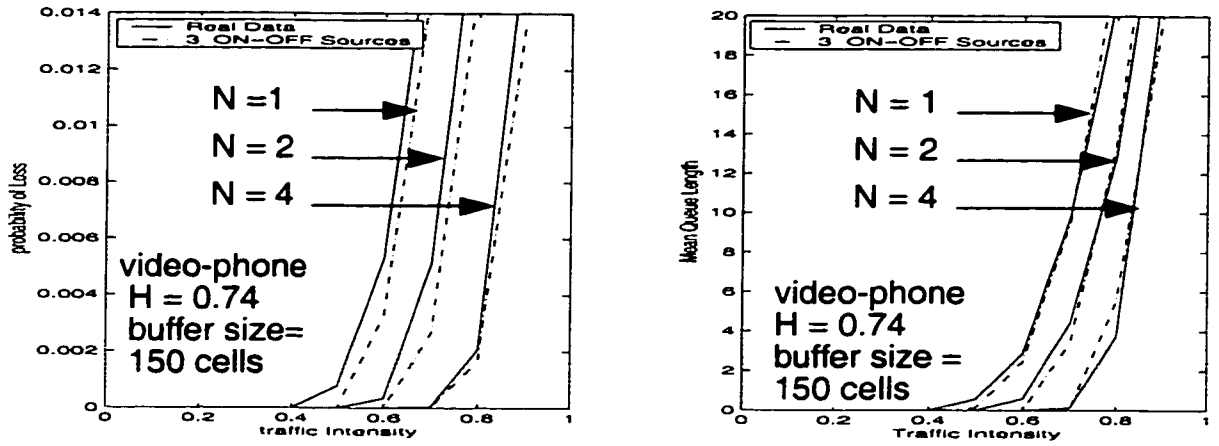

FIGURE.6.25. Comparison of the probability of loss and mean queue length of real video-phone data and that of generated 3 class heterogeneous source model, each class has 1 ON-OFF source, number of multiplexed sources N = 1, 2, 4

In figure 6.26, the multiplexing of several highly correlated traffic, Movie, are shown. The improvement of the multiplexing in the prediction of the probability of loss and mean queue length is more for Movie sequence as compared with the video-phone sequence. Also, there is a reduction in the proability of loss and mean queue length as the number of sources multiplexed increases.

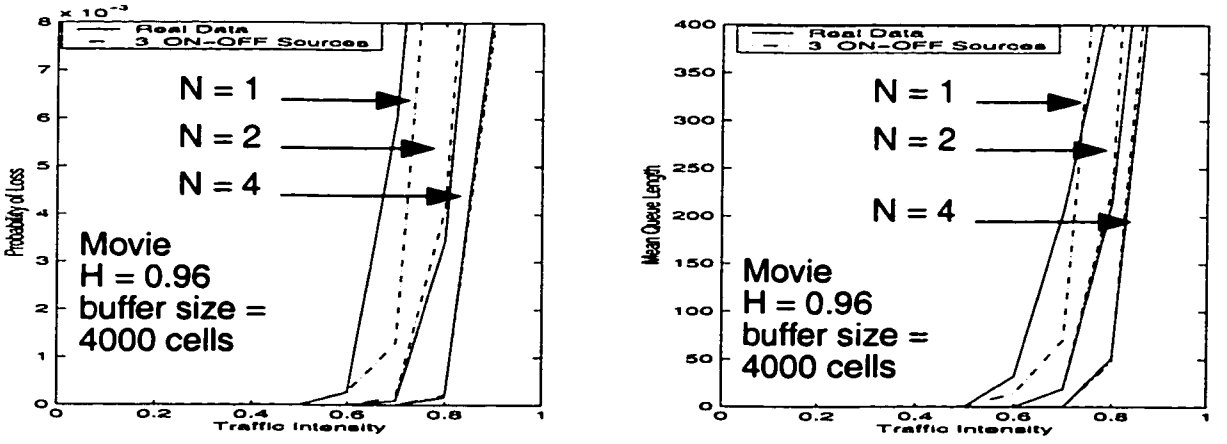

FIGURE.6.26. Comparison of the probability of loss and mean queue length of real Movie data and that of generated 3 class heterogeneous source model, each class has 1 ON-OFF source, number of multiplexed sources N = 1, 2, 4

6.6 Comparison between heterogeneous ON-OFF source models, Maglaris models, Markov chain models and MMPP models

In this section we discuss the different models we used in characterizing and predicting the performance measures of video traffic. We compare the covariance, index of dispersion for counts for the models and the real data. Also we compare the probability of loss, and mean queue length as a function of the traffic intensity and constant buffer capacity for the models and the real video data.

Some models are too inaccurate to be of use, and conclusion based upon them can be very misleading. These models fail because they assume arrivals are uncorrelated, whereas in fact they are correlated. Other models take these correlated arrivals into account and produce more accurate results [HEF86]. It is necessary to look for the probability of loss and the mean queue length, which are the statistics that matter. The choice of the best model is based on the best matching of the probability of loss. Also it must have a small number of fitting parameters to make it analytically tractable and at the same time have the same properties as the actual sources

Before going into the comparison of the models and their prediction of the traffic characteristics and performance measures compared to the real data, we give a brief review of the models that we considered in the previous chapters. All of them are based on the conventional traffic models. They are simple and analytically tractable. They have been used to model self-similar traffic data. They are: Markov-modulated Poisson process, Maglaris model, Markov chain models, and the three heterogeneous binary source model, which we have developed. However, the accuracy of the models in characterizing and predicting the performance measures of the real data depends on how heavily the traffic is correlated.

As we have mentioned, video traffic may be characterized by a MMPP model based on many independent identical ON-OFF minisources. The Maglaris [MAG88] model is a model that consists of a number of ON-OFF sources. Fluid

flow in their queueing analysis. For most cases, one video source is modeled as a sum of 20 minisources. Furthermore, the Markov chain models can be used to approximate the self-similar traffic, even if they do not have the LRD that self-similar traffic exhibits. In [HEY96] it is shown that LRD is not a crucial property in determining acceptable QoS when delay considerations limit the allowable load.

We present a discussion of some results obtained from different models we considered in the previous chapters including our proposed heterogeneous ON-OFF source model. We generated the traffic for heterogeneous ON-OFF source model, Maglaris model, and MMPP using OPNET given the parameters obtained from matching to the real data discussed in the previous chapters. The Markov chain is generated using Matlab software based on estimating the transition probabilities P from an actual sequence and then we use these probabilities to generate the synthetic data.

6.6.1 Comparison of traffic characteristics indices and performance measures of the models

We compare the estimated values of the covariance and the IDC of the real data with that of all the models under consideration. Also, we compare the probability of loss and mean queue length of the real data and that of the models as a function of the traffic intensity given a fixed buffer size.

Figure 6.27 and figure 6.28 shows a comparison of the covariance functions for video-conferencing and video-phone data and that based on 3-heterogeneous ON-OFF source model, Maglaris model, MMPP and the Markov chain models. Although the covariances of all models approximate the real data, the three heterogeneous ON-OFF source model tracks the real data most closely over a long interval of frames.

In figure 6.29 and figure 6.30 we show a comparison of the covariance functions for TV series and Movie and that based on 3-heterogeneous ON-OFF source model, Maglaris model, MMPP and Markov chain. It is clear from the figures that the 3-heterogeneous model preforms better than the other models for this kind of highly correlated traffic

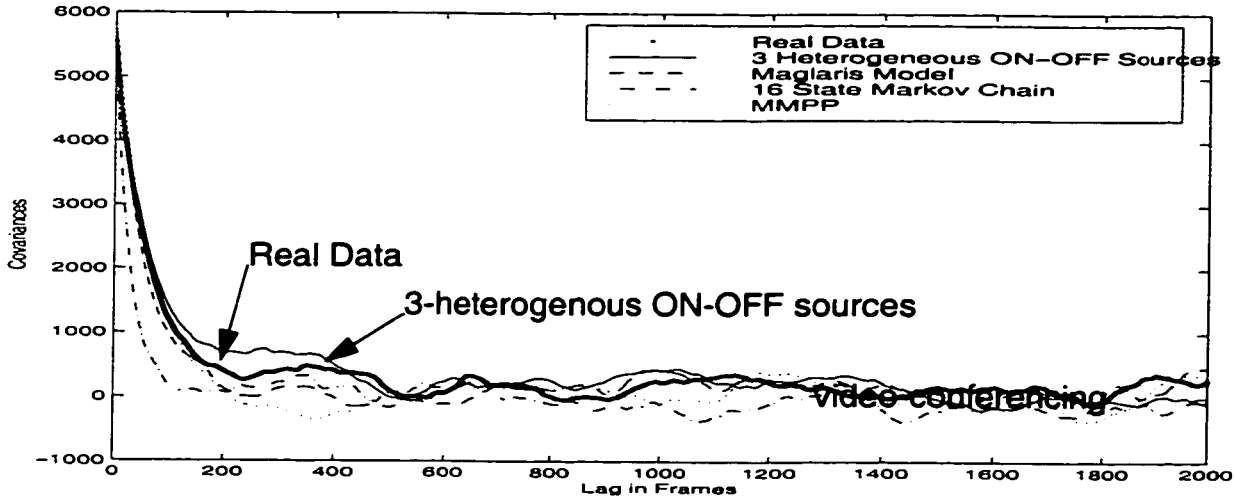

FIGURE.6.27. Comparison of the covariance of video-conferencing data, 3-heterogeneous ON-OFF source model, Maglaris model, 16 state Markov chain and MMPP.

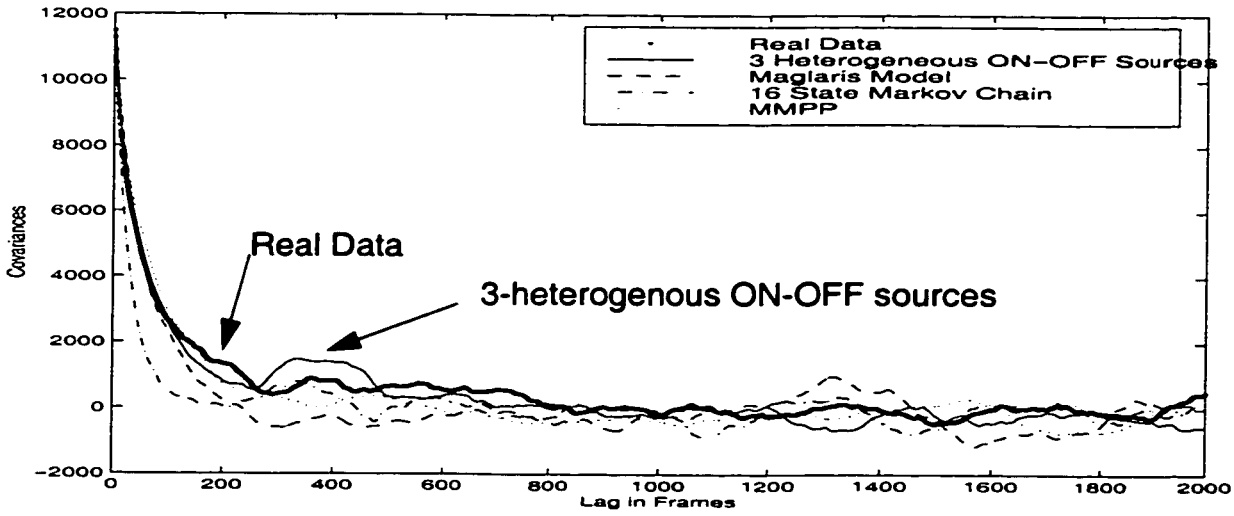

FIGURE.6.28. Comparison of the covariance of video-phone data, 3-heterogeneous ON-OFF source model, Maglaris model, 16 state Markov chain and MMPP.

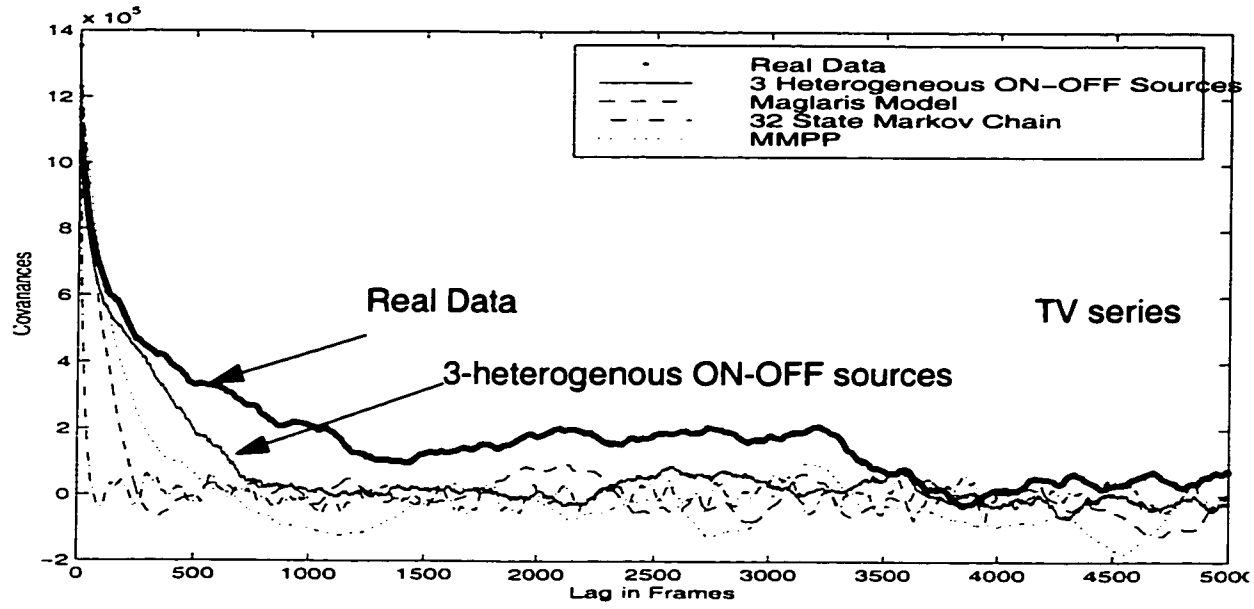

FIGURE.6.29. Comparison of the covariance of TV series data, 3-heterogeneous ON-OFF source model, Maglaris model, 32 state Markov chain and MMPP.

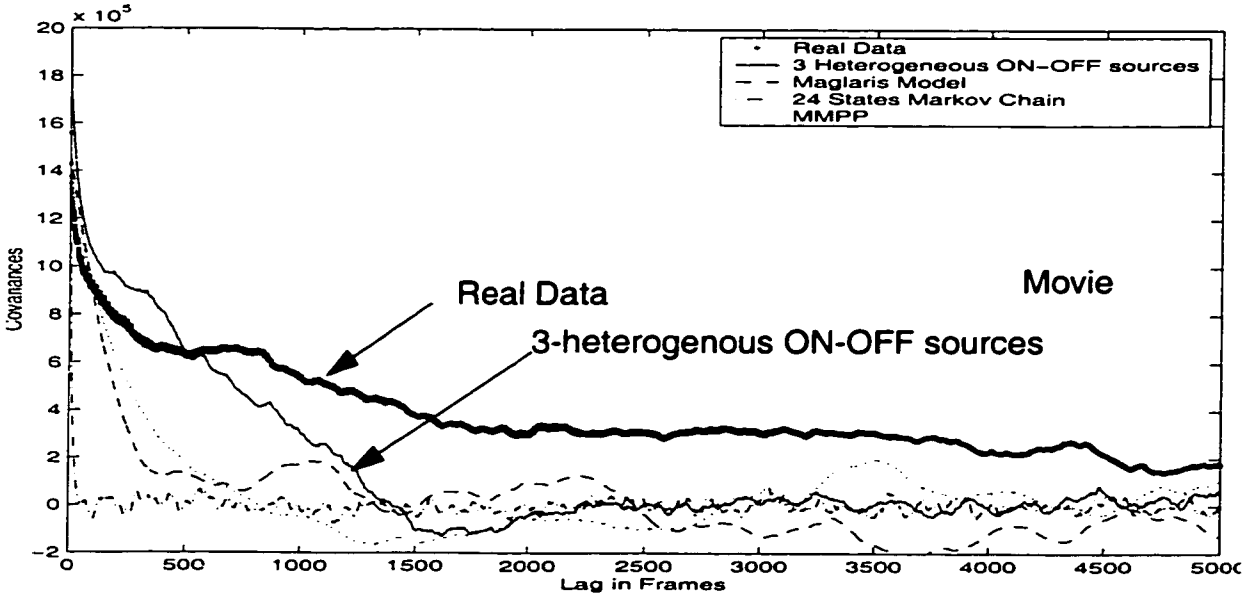

FIGURE.6.30. Comparison of the covariance of Movie series data, 3-heterogeneous ON-OFF source model, Maglaris model, 24 state Markov chain and MMPP.

Figure 6.31 and figure 6.32 show the IDC for video-conferencing and video-phone data and that based on 3-heterogeneous ON-OFF source model, Maglaris

model, MMPP and 16 state Markov Chain. The IDC for 3-heterogeneous ON-OFF source model has the best matching over a long range of lags among the other models that are under consideration. That is, in comparison with the other models used, over the entire range of the lag, the IDC for the 3-heterogeneous source model and the real data are in good agreement.

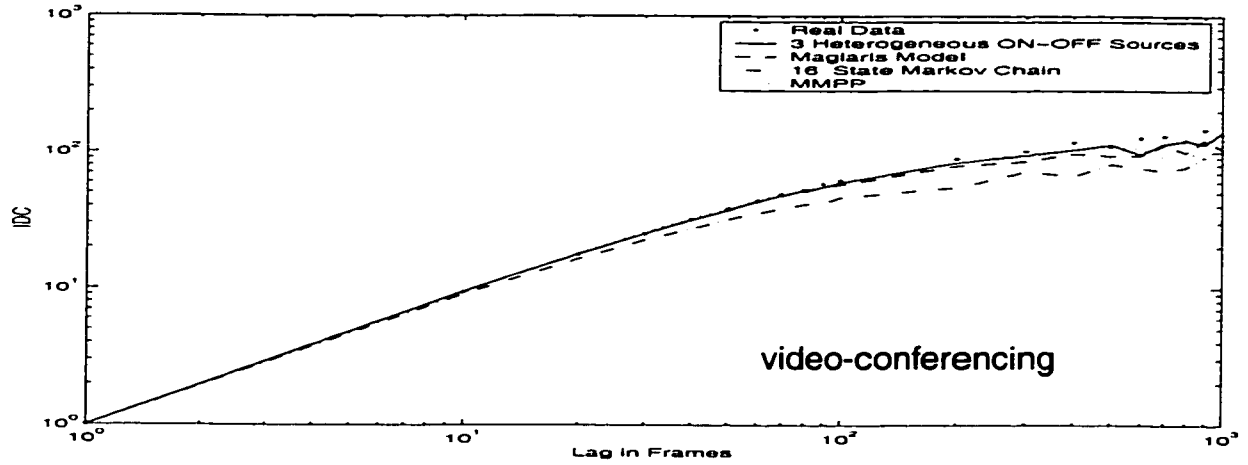

FIGURE.6.31. Comparison of the IDC of video-conferencing data, 3-heterogeneous ON-OFF source model, Maglaris model,16 state Markov chain and MMPP.

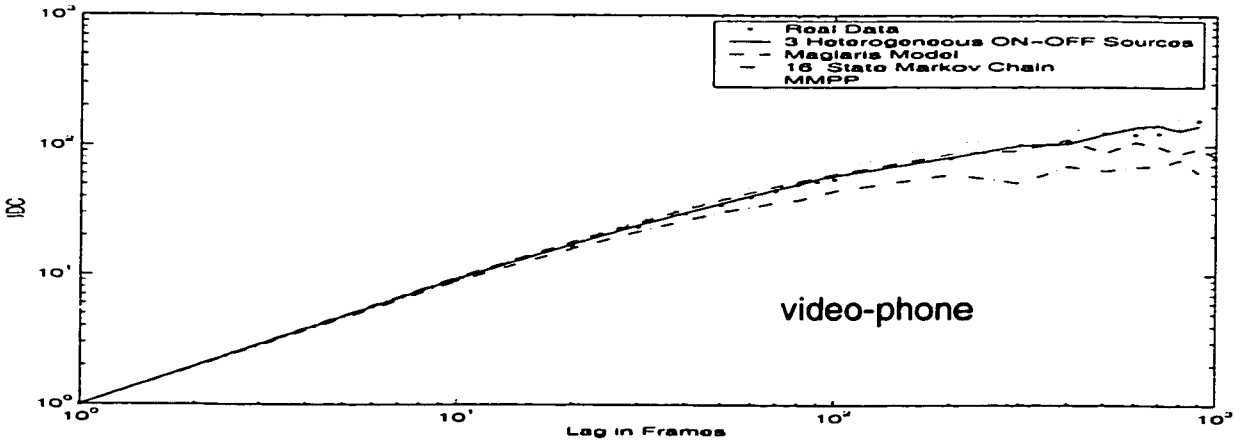

FIGURE.6.32. Comparison of the IDC of video-phone data, 3-heterogeneous ON-OFF source model, Maglaris model,16 state Markov chain and MMPP

In figure 6.33 and figure 6.34, we show the IDC for the entertainment video data, Movie and TV series. The Markov chain does not work for this highly corre-

lated traffic even for a large number of states, as we have discussed in the previous chapter. The 3-heterogeneous ON-OFF source model prediction is good as shown in the figures. Moreover, there is an improvement in the matching using 3-heterogeneous ON-OFF source model over that using Maglaris and MMPP models.

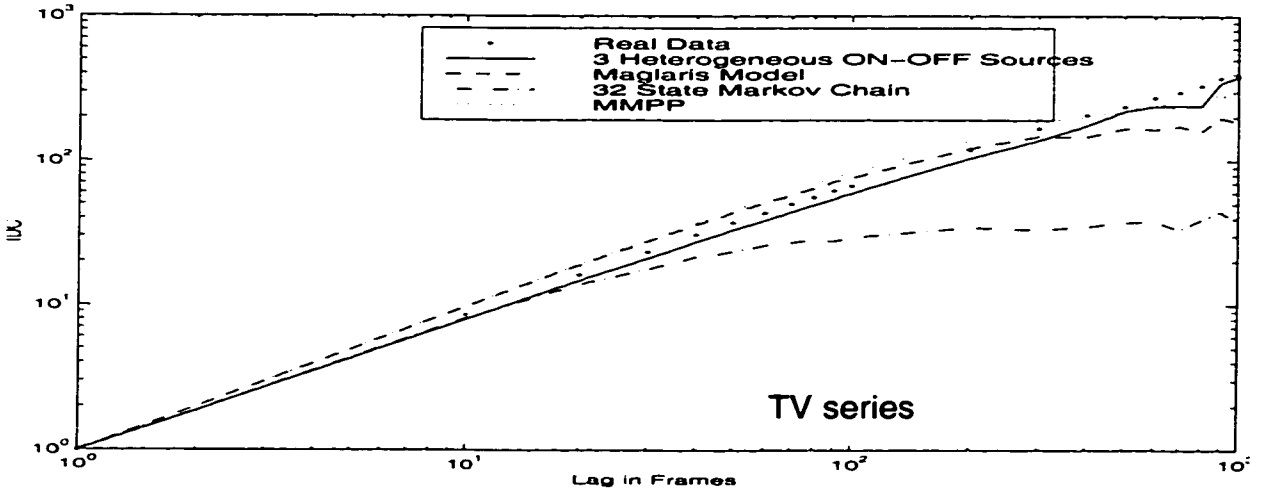

FIGURE.6.33. Comparison of the IDC of TV series data, 3-heterogeneous ON-OFF source model, Maglaris model, 32 state Markov chain and MMPP.

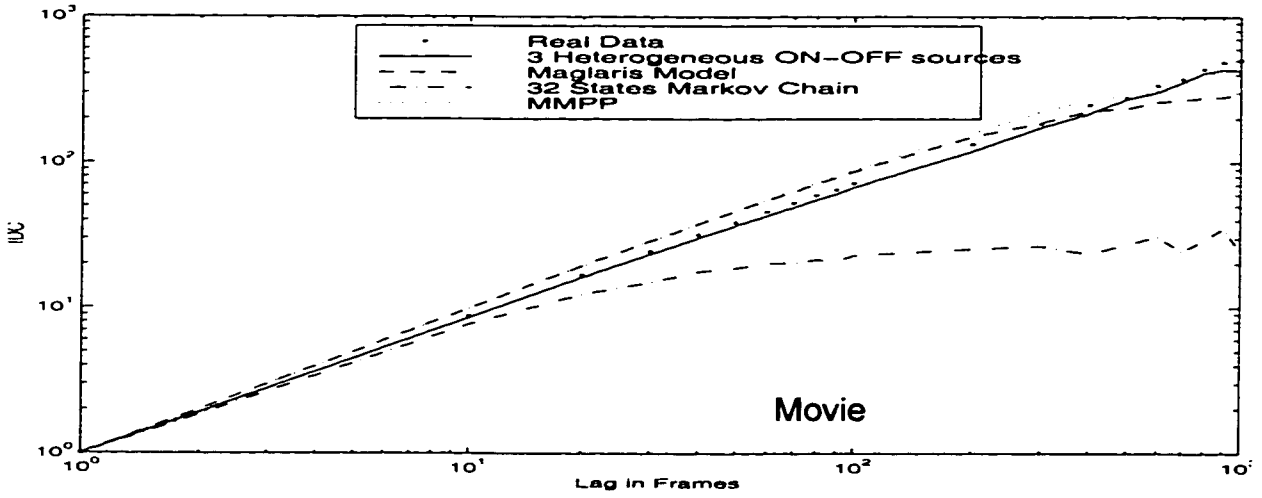

FIGURE.6.34. Comparison of the IDC of Movie data, 3-heterogeneous ON-OFF source model, Maglaris model, 32 state Markov chain and MMPP.

As for the covariance and the IDC, the probability of loss estimation based on the 3-heterogeneous ON-OFF source model and the real data are in good

agreement as compared with the other models. The 3-heterogeneous ON-OFF source model predicts the probability of loss over a substantial large range of traffic intensities and with a buffer capacity of 150 cells. Comparison for the probability of loss for real data, 3-heterogeneous ON-OFF source model, Maglaris model, MMPP and Markov chain are shown in figure 6.35 - 6.38. It is clear that for the telconferencing data that all the models give a reasonable prediction of the probability of loss, however the 3-heterogeneous ON-OFF source model is more accurate. For the entertainment data, Markov chain gives a poor prediction of the probability of loss while the other models are less accurate than the 3-heterogeneous ON-OFF source model.

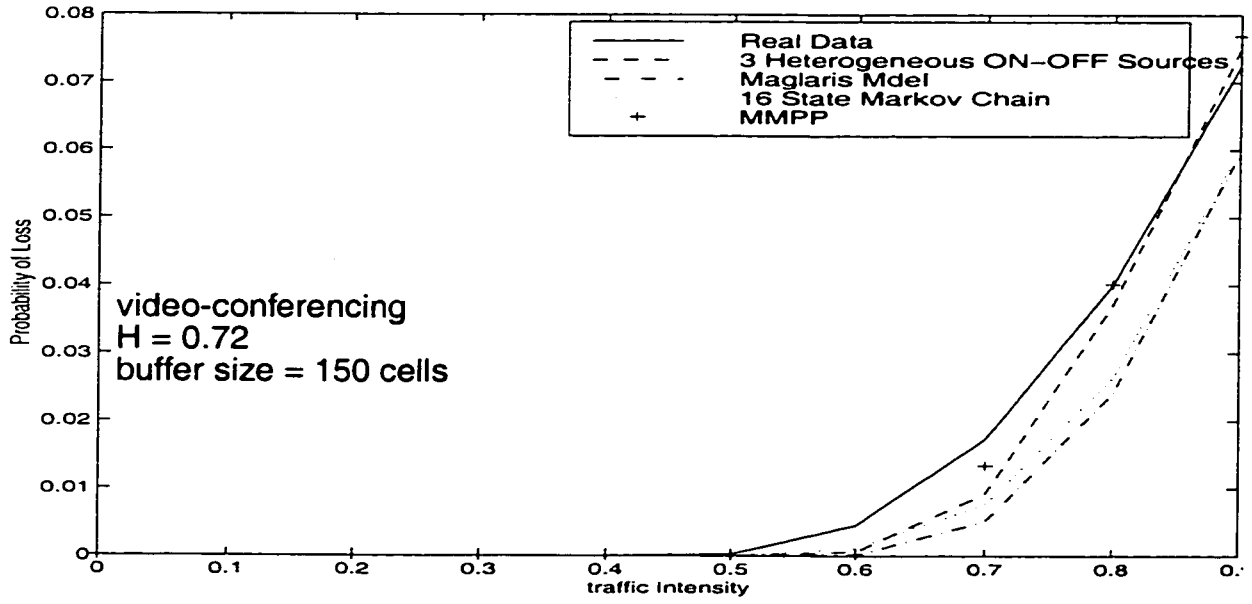

FIGURE.6.35. Probability of loss of real video-conferencing data compared with that of generated 3 ON-OFF heterogeneous source model, Maglaris model, 16 state Markov chain and MMPP.

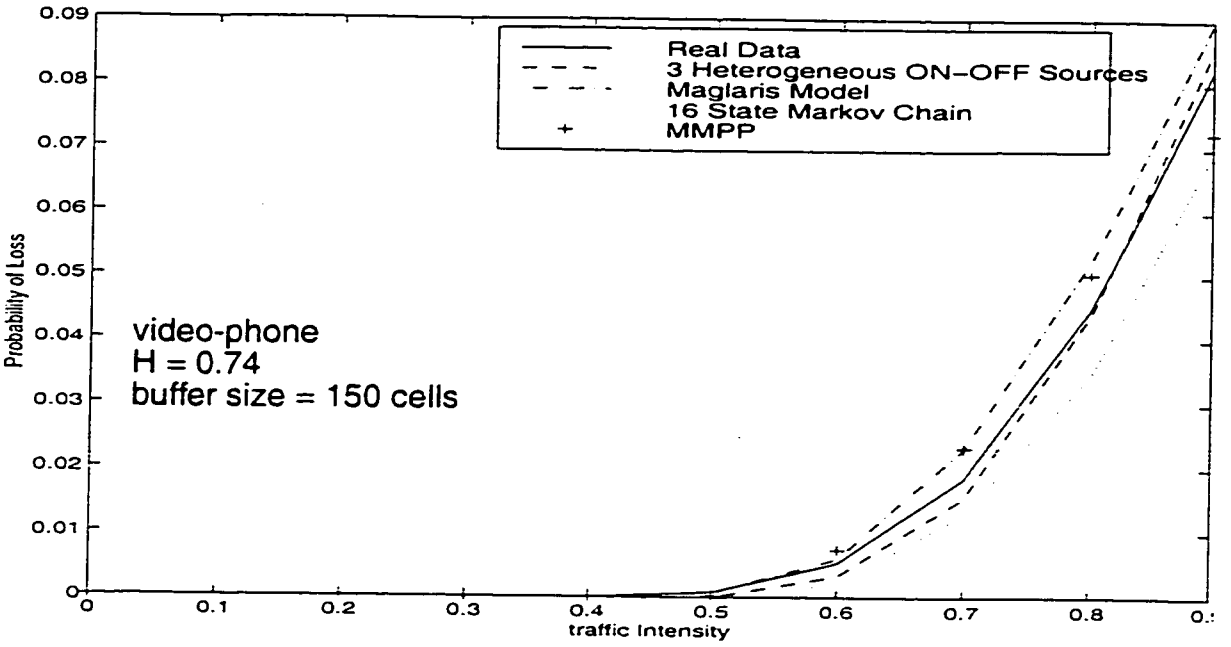

FIGURE.6.36. Probability of loss of real video-phone data compared with that of generated 3 ON-OFF heterogeneous source model, Maglaris model, 16 state Markov chain and MMPP.

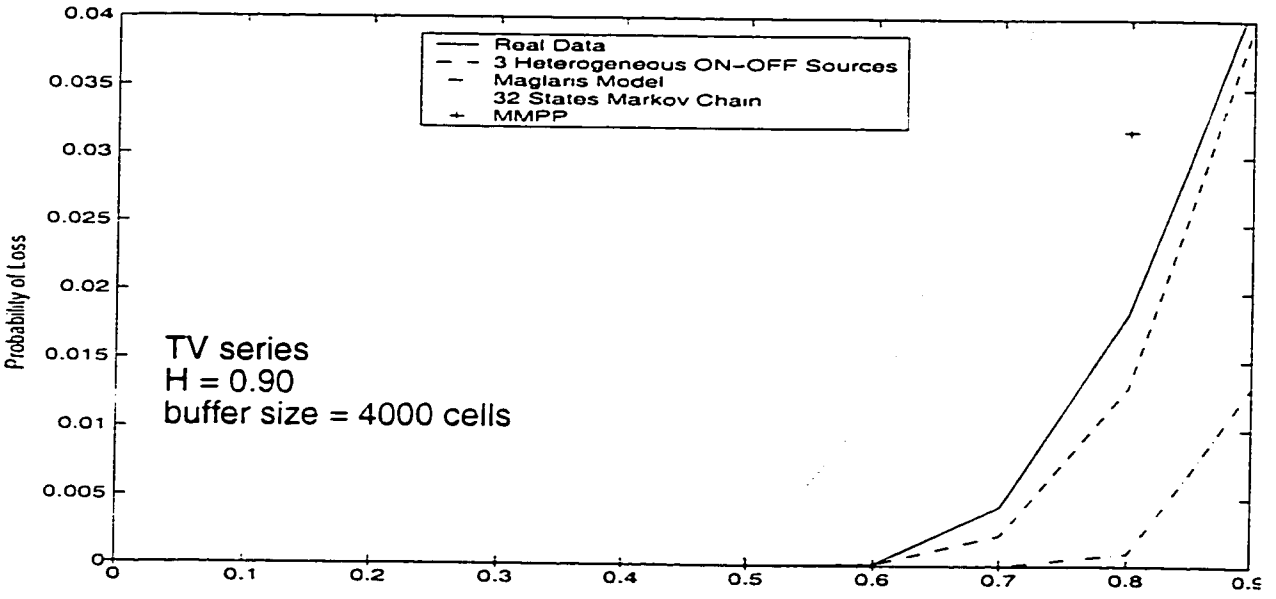

FIGURE.6.37. Probability of loss of TV series compared with that of generated 3 ON-OFF heterogeneous source model, Maglaris model, 32 state Markov chain and MMPP.

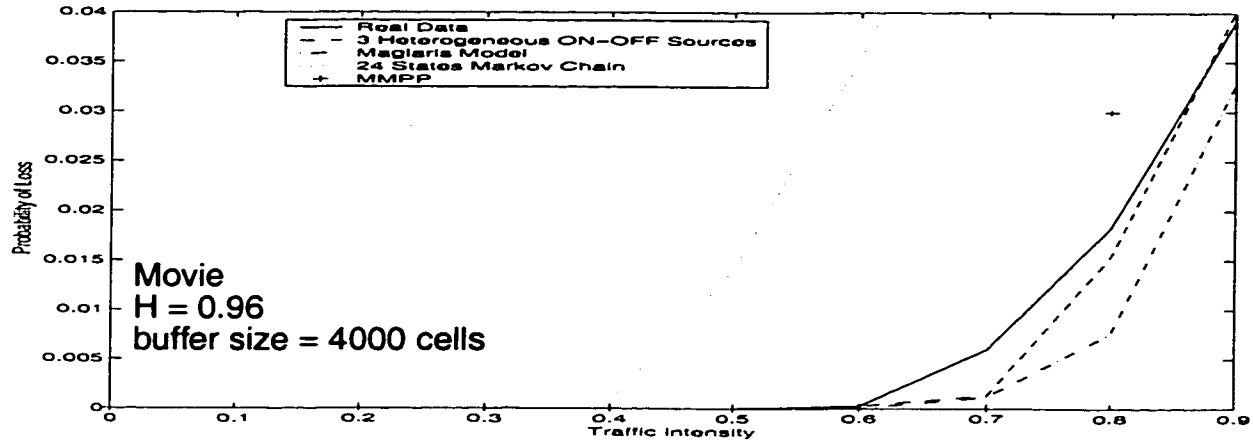

FIGURE.6.38. Probability of loss of Movie compared with that of generated 3 ON-OFF heterogeneous source model, Maglaris model, 24state Markov chain and MMPP.

The mean queue lengths for the teleconferencing video traces are shown in figure 6.39 and figure 6.40. The 3-heterogeneous ON-OFF source model does a very good job of tracking the mean queue length over a large range of traffic intensities among all other models given that the buffer is finite of value 150 cells. Figure 6.41 and figure 6.42 show the mean queue length for the highly correlated entertainment video traces. The 3-heterogeneous model prediction is good for practical engineering design as compared with the other models.

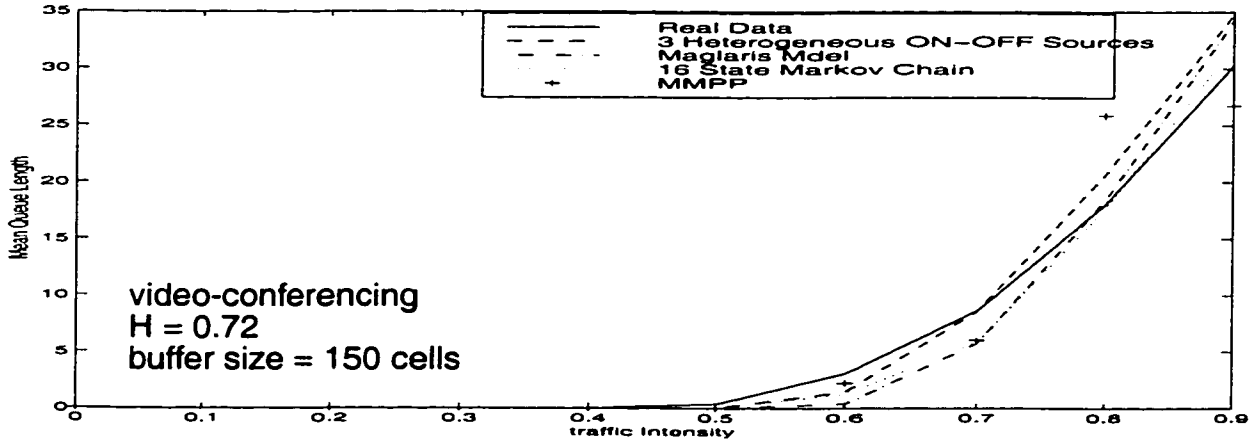

FIGURE.6.39. Mean queue length of real video-conferencing data compared with that of generated 3 ON-OFF heterogeneous source model, Maglaris model, 16 state Markov chain and MMPP.

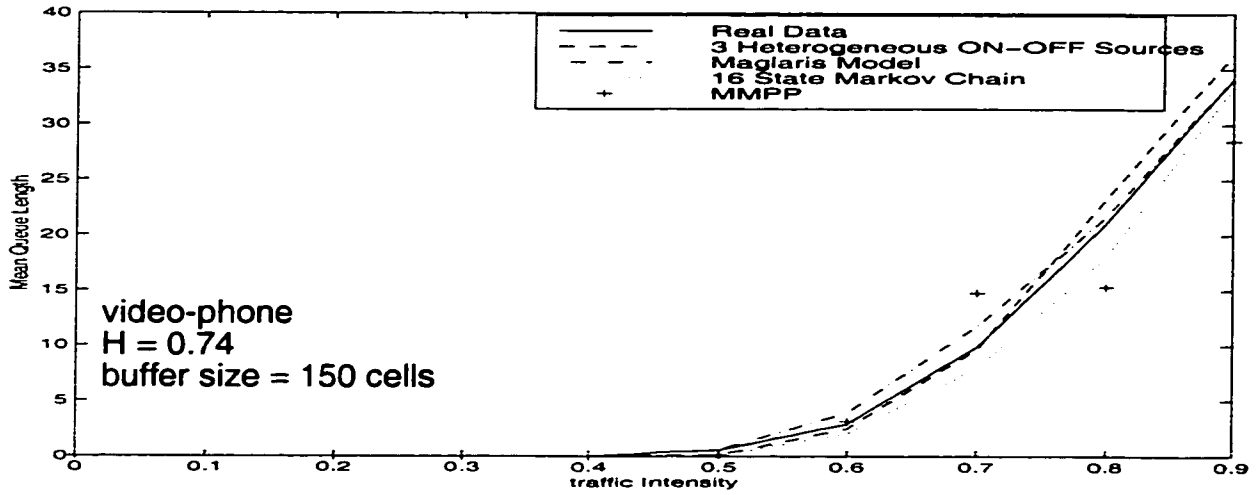

FIGURE.6.40. Mean queue length of real video-phone data compared with that of generated 3 ON-OFF heterogeneous source model, Maglaris model, 16 state Markov chain and MMPP.

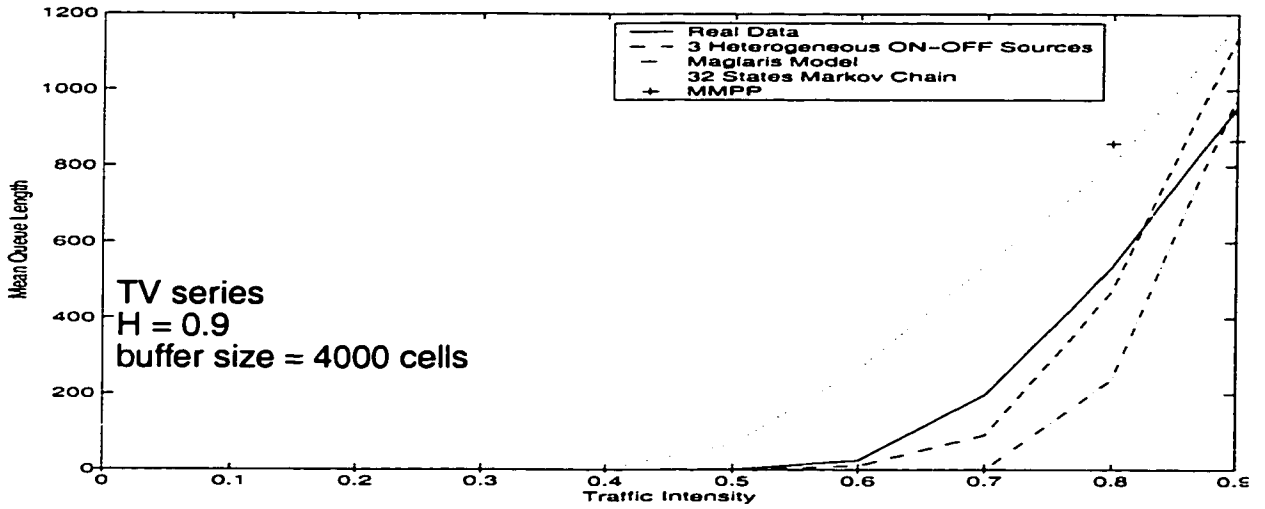

FIGURE.6.41. Mean queue length of TV series data compared with that of generated 3 ON-OFF heterogeneous source model, Maglaris model, 32 state Markov chain and MMPP.

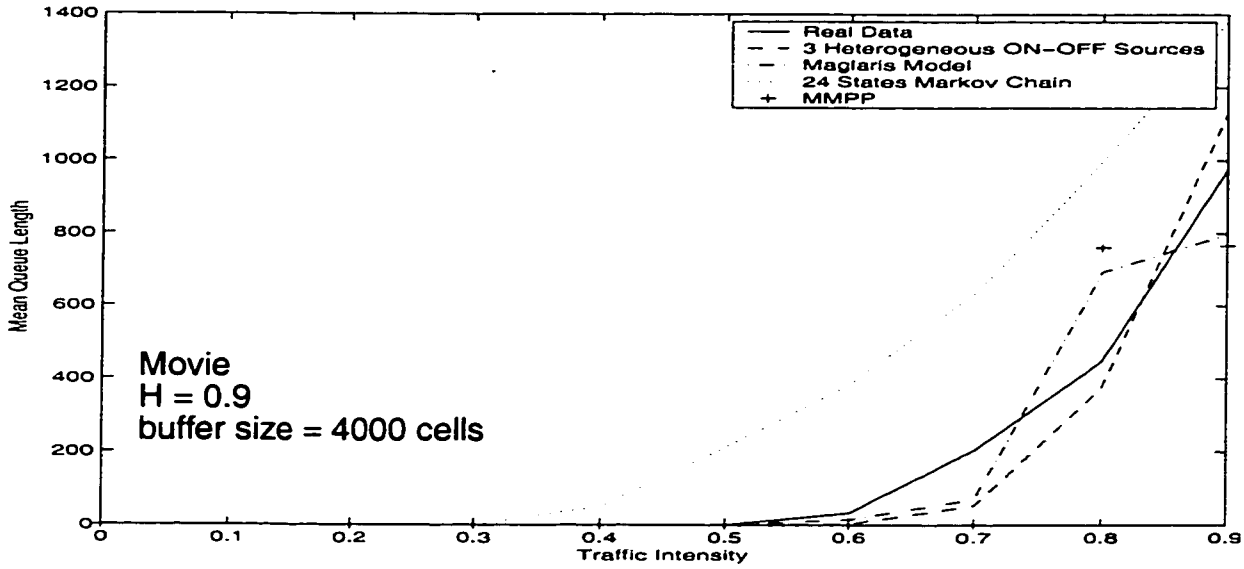

FIGURE.6.42. Mean queue length of Movie data compared with that of generated 3 ON-OFF heterogeneous source model, Maglaris model, 24 state Markov chain and MMPP.

6.7 Discussion

We have proposed a model for characterizing correlated cell arrival of real self-similar video data. Based on a second order statistical analysis, we have used heterogeneous ON-OFF source model to characterize the traffic. The model consist of m classes of ON-OFF sources. Although the ON-OFF periods are exponentially distributed and the number of sources is small, we have a good matching for the covariance and the IDC. It is clear from the results we obtained, as the number of classes and the number of ON-OFF sources per class increases, the accuracy of the model will be increased especially for highly correlated traffic. However, increasing the number of classes and number of sources per class will result in analytical and computational complexities.

We showed that it is possible to predict the probability of loss and mean queue length as a function of the traffic load with buffer capacity of constant value.

As for the covariance and IDC, better results will be achieved if we use a larger number of classes but with complexity trade-off. In a comparison with the Maglaris scheme, the heterogeneous ON-OFF source model QoS may be significantly better.

As we have presented in section 6.6, it is clear that the three heterogeneous ON-OFF source model is the best model among the other models that we considered. We have seen that although some of the models approximately can predict the IDC and covariance, they have less far predictive ability for the probability of loss and mean queue length especially when the correlation index is large. This means that models can predict covariance and IDC accurately but can have different QoS.

Using just the 3-heterogeneous ON-OFF source model gives good results for matching the traffic characteristics indices and prediction of the QoS and at the same time the analysis is simple. Going for a large number of classes and more sources per class will give more accurate results, however, the analysis will be more complicated.

CHAPTER VII

Congestion and Admission Control of Self-Similar Traffic based on Multiple Types ON-OFF Sources

7.1 Introduction

Most traffic sources in ATM networks are bursty, which makes them ideal for statistical multiplexing, which can be better achieved if effective congestion control and traffic management schemes (i.e., the set of policies and mechanisms that allow a network to efficiently satisfy a diverse range of service requests) can be developed. The basic idea behind traffic management is that the users should be able to tell the network what services they require for a connection. Conversely, the network must be able to monitor and control traffic according to agreements with users.

Congestion happens whenever the demand is more than the available capacity:

$$\sum_i \text{Demand}_i > \text{Available Capacity}$$

The success or failure of ATM networks depends on the development of an effective congestion-control framework. Some aspects of ATM networks that complicate the control problem include:

- Traffic characteristics of various types of services are not well understood.

- Different services have different types of QoS requirements at considerably varying levels.

- Various B-ISDN VBR sources generate traffic at significantly different rates from Kbps to Gbps.

- As the transmission speed increases, the ratio of call duration to the cell transmission time increases. Due to large bandwidth-propagation delay product in B-ISDN, there is a very large number of cells in transit at any time in the network. Large propagation delays compared with the transmission times give rise to large periods between the onset and the detection of the congestion conditions by the network control elements.

In our approach there is feedback from the multiplexer to the end system which gives the source the information necessary to respond, by appropriately modifying their submission rates, to changes in the available bandwidth, so that congestion is controlled or even avoided and the available bandwidth is used. It is a rate based flow control under the category of proactive control schemes.

Congestion can occur when the network accepts too many calls, which causes the QoS to deteriorate. If feedback is employed, the throughput could be improved, with improvements depending strongly on how large the feedback delay is. If there is no feedback delay, the node may signal congestion to the sources when transmission capacity is fully utilized. In this case, the rate at the source is reduced and data is buffered at the source. When some feedback delay is present, congestion signaling should occur within a round trip delay period.

Real time services can not tolerate long delays, and therefore buffering for these services must be kept to a minimum; accordingly, bandwidth is the principal network resource that affects their performance, and connection admission con-

trol, which is a set of actions taken by the ATM network to determine if the network has sufficient resources to support a new connection request, can be based on call bandwidth requirements. The characterization of bandwidth requirements can be accomplished in different ways. For the CBR service category, peak bandwidth admission control is applicable and can result in good bandwidth utilization. This is identical to admission control applied to circuit switched networks. Applying peak bandwidth admission control to real time VBR service category can result in a low efficiency of bandwidth utilization. In this case, bandwidth is allocated on an equivalent bandwidth basis defined as the bandwidth required for call-level performance to satisfy the required QoS. The value stands between peak bandwidth and its mean.

As we have presented in chapter 6, it is clear that the heterogeneous ON-OFF source model is the best model among the other models that we considered. Using just 3-heterogeneous ON-OFF sources gives good results for matching the traffic characteristics indices and prediction of the QoS and at the same time the analysis is simple.

Our concern here is with probability of time to overload in a round trip delay time, defined as the probability that the multiplexer input bit rate is greater than the output bit rate, and admission control for a number of heterogeneous sources. The probability distribution for the time to overload in a round trip delay can be found as the first passage time distribution of a multidimensional birth and death process to a set of states which define overload. Specifically, we advocate that the probability distribution to overload in a network round trip time serve as a quantita-

tive measure of congestion imminence. This is in accordance with the fact, that the network round trip time is the reaction time of the system. The probability distribution for the time to overload can be expressed as a summation of exponentially decaying terms of order equal the product of the number of subscribed sources over all source types.

The time scales that are involved, allow us to use the fluid flow approximation technique. Fluid models characterize traffic as a continuous stream. As we have discussed in chapter 3, a fluid model that is normally used to model traffic is the Markov modulated fluid model. In this model, the current state of the underlying Markov chain determines the flow rate. Anik, et. al. [ANI82] used the fluid model to derive the steady state complementary probability distribution for the buffer content. In a paper by [ELW93], they examined the admission control problem for arbitrary, possibly heterogeneous bursty traffic sources. By preventing admission to an excessive number of calls or sources of the multiplexer, call admission balances between grade of services, as determined by delay, cell loss probability, for instance, and efficient use of the network resources. They showed that for general Markovian traffic sources, it is possible to assign an effective bandwidth to each source, which is the maximal real eigenvalue of a matrix derived from the source and the channel characteristics. As expected, the effective bandwidth of a source is shown to be bounded by peak rates and mean rates.

The analysis in [COL96] is similar to that of Elwalid and Mitra [ELW94] where a fluid flow approximation is used. Their scheme involves rate control when a queue threshold is passed. They assumed the existence of a statistical shaper

that can, for instance, take traffic from a source with a given average burst length and produce traffic with a lower average burst length, or perhaps even a longer average burst length, but with a lower cell rate within the burst.

The fluid flow model for sets of homogenous sources was used by Tsingotjidis and Hayes [PER97] to predict the onset of congestion. The problem is formulated as a first passage time to an overload state.The overload state is defined to be input flow greater than the capacity of output lines. Since the traffic is real-time, the role of buffering in smoothing fluctuation is limited. In view of the quantities of flows that are involved, the time delays that are appropriate and the tractability of the problem, we feel this definition of overload state is reasonable. The salient result of this work is that the correlation time of video processes and the round-trip time of links, such as LEO and MEO satellite, are such that a significant advantage can be realized. The video model in this work is that of Maglaris. As we have seen, the heterogeneous ON-OFF source model that we have developed is a considerable improvement on the Maglaris model; accordingly it is applied to the congestion control problem.

7.2 The general mathematical model

Based on a second order statistical analysis [FAR98], we have used heterogeneous ON-OFF source model to characterize the VBR video self-similar traffic. Although the ON-OFF periods are exponentially distributed and the number of sources is small, we have a good matching for the covariance. We also showed that it is possible to predict the probability of loss and mean queue length. In a

comparison with the Maglaris scheme, the heterogeneous ON-OFF source model QoS may be significantly better. The model can be generalized for any number of levels, which we consider below.

In section 6.2 we have presented a model that characterizes heterogeneous ON-OFF source model at equilibrium; i.e., time has no role in this case. We ended for that case with an equation from which the covariance function was extracted and the model parameters are calculated. In the following we use the same model that was presented in section 6.2 having in this case two variables, time and buffer content.

In this section we introduce the time dependency in order to find the probability distribution of time to overload in a round trip delay time. We have a system of m independent classes, let N_i $(i = 1, 2,....m)$ denote the number of sources in class i. Within a class the sources are identical and independent. Let $[n_1 n_2....n_m; t; u]$ be the state with n_i source in class i "ON" $(i = 1, 2,....m)$ and the buffer content does not exceed u at time t and $p_{n_1 n_2.....n_m}(t, u)$ be its probability. The times spent in the ON and OFF states are exponentially distributed with means $1/\beta_i$ and $1/\alpha_i$, $i = 1, 2,....m$, respectively. Also, any source from class i is generating packets at fixed peak rate R_i. Packets are served at a rate of C packet per time unit. The string of indexes $[n_1 n_2....n_m]$ is considered the normal one in one differential equation. The string is indicated by dots ..... while only deviations from the normal string are given explicitly. We utilize the fluid flow approximation [ANI82], which has shown much promise in analysis of ATM networks. The computational complexity of this technique is independent of buffer

size; accordingly, more complex (in terms of Markovian state space) traffic models than other techniques are allowed. Therefore, similar to [ANI82], we have the following set of partial differential equations governing the model:

$$\sum_{i=1}^{m} \frac{\partial}{\partial t} p_{\ldots\ldots}(t,u) + \left(\sum_{i=1}^{m} R_i n_i - C \right) \frac{\partial}{\partial u} p_{\ldots\ldots}(t,u) = \sum_{i=1}^{m} [\alpha_i (N_i - n_i + 1)] p_{\ldots n_i - 1 \ldots}(t,u) - \{\alpha_i (N_i - n_i) + \beta_i n_i\} p_{\ldots\ldots}(t,u) + \beta_i (n_i + 1) p_{\ldots n_i + 1 \ldots}(t,u) \quad (7.1).$$

$$(0 \le n_i \le N_i; \quad i = 1, 2, \ldots\ldots m)$$

where the variables t and u are time and buffer contents, respectively. In the sequel, we shall consider a particular order of states, which is useful in our study.

We express equation (7.1) in the following familiar matrix form,

$$\frac{\partial}{\partial t} \boldsymbol{p}(t,u) + \boldsymbol{D} \frac{\partial}{\partial u} \boldsymbol{p}(t,u) = \boldsymbol{Q} \boldsymbol{p}(t,u) \quad (7.2).$$

$\boldsymbol{p}(t,u)$ is a column vector equal $[p_{00\ldots0}(t,u), \ldots\ldots\ldots\ldots, p_{n_1 n_2 \ldots\ldots n_m}(t,u)]^T$, $\boldsymbol{D}$ is an $(N_1+1) \times (N_2+1) \times \ldots\ldots \times (N_m+1)$ diagonal matrix and $\boldsymbol{Q}$ is an $(N_1+1) \times (N_2+1) \times \ldots\ldots \times (N_m+1)$ infinitesimal generator matrix. Therefore, the dimensionality of the problem is the product of the number of subscribed sources over all source types.

For example, let us consider the bufferless case $u = 0$ and setting $\frac{\partial}{\partial u} \boldsymbol{p}(t,u) = 0$, and assuming we have three classes and each class has one source $(N_1 = N_2 = N_3 = 1)$ as shown in figure 6.3. Let us consider the state $n_1 n_2 n_3 = 000$. In this case, equation(7.1) becomes,

$$\frac{dp_{000}(t)}{dt} = -(\alpha_1 + \alpha_2 + \alpha_3) p_{000}(t) + \beta_3 p_{001}(t) + \beta_2 p_{010}(t) + \beta_1 p_{100}(t)$$

For the state $n_1 n_2 n_3 = 001$, equation (7.1) will be:

$$\frac{dp_{001}(t)}{dt} = \alpha_3 p_{000}(t) - (\alpha_1 + \alpha_2 + \beta_3)p_{001}(t) + \beta_2 p_{011}(t) + \beta_1 p_{101}(t)$$

and so on for the other six states.

The infinitesimal generator matrix Q for the three heterogeneous binary ON-OFF source model is given by,

$$Q = \begin{bmatrix}
-(\alpha_1+\alpha_2+\alpha_3) & \alpha_3 & \alpha_2 & 0 & \alpha_1 & 0 & 0 & 0 \\
\beta_3 & -(\alpha_1+\alpha_2+\beta_3) & 0 & \alpha_2 & 0 & \alpha_1 & 0 & 0 \\
\beta_2 & 0 & -(\alpha_1+\alpha_3+\beta_2) & \alpha_3 & 0 & 0 & \alpha_1 & 0 \\
0 & \beta_2 & \beta_3 & -(\alpha_1+\beta_2+\beta_3) & 0 & 0 & 0 & \alpha_1 \\
\beta_1 & 0 & 0 & 0 & -(\beta_1+\alpha_2+\alpha_3) & \alpha_3 & \alpha_2 & 0 \\
0 & \beta_1 & 0 & 0 & \beta_3 & -(\beta_1+\alpha_2+\beta_3) & 0 & \alpha_2 \\
0 & 0 & \beta_1 & 0 & \beta_2 & 0 & -(\beta_1+\beta_2+\alpha_3) & \alpha_3 \\
0 & 0 & 0 & \beta_1 & 0 & \beta_2 & \beta_3 & -(\beta_1+\beta_2+\beta_3)
\end{bmatrix}$$

In the homogenous case, the problem is a simple birth-death process. For our case we have a multidimensional birth and death process. The dimensionality of the problem is the product of the number of subscribed sources over all source types. For N sources three classes system, let $q(k, l, p;x, y, z)$ =transition rates from state k, l, p to state x, y, z. The local state transition diagram for a three types of source model is shown in figure 7.1.

The birth-death rates with $k, l, p = 0, 1, ...N$, are given by:

$$g(k, l, p;k-1, l, p) = k\beta_1 \qquad (7.3).$$

$$g(k, l, p;k, l-1, p) = l\beta_2 \qquad (7.4).$$

$$g(k, l, p;k, l, p-1) = p\beta_3 \qquad (7.5).$$

$$g(k, l, p;k+1, l, p) = (N-k)\alpha_1 \qquad (7.6).$$

$$g(k, l, p;k, l+1, p) = (N-l)\alpha_2 \qquad (7.7).$$

$$g(k, l, p;k, l, p+1) = (N-p)\alpha_3 \tag{7.8}$$

$$g(k, l, p;k, l, p) = -\sum_{k=0,\, l=0,\, p=0}^{N} g(k, l, p;k-1, l, p) + g(k, l, p;k, l-1, p) + g(k, l, p;k, l, p-1) + g(k, l, p;k+1, l, p) + g(k, l, p;k, l+1, p) + g(k, l, p;k, l, p+1) \tag{7.9}$$

equations (7.3) - (7.9) will be used in the next section to calculate the infinitesimal generator matrix Q of dimensionality $(N+1)^m \mathrm{x} (N+1)^m$.

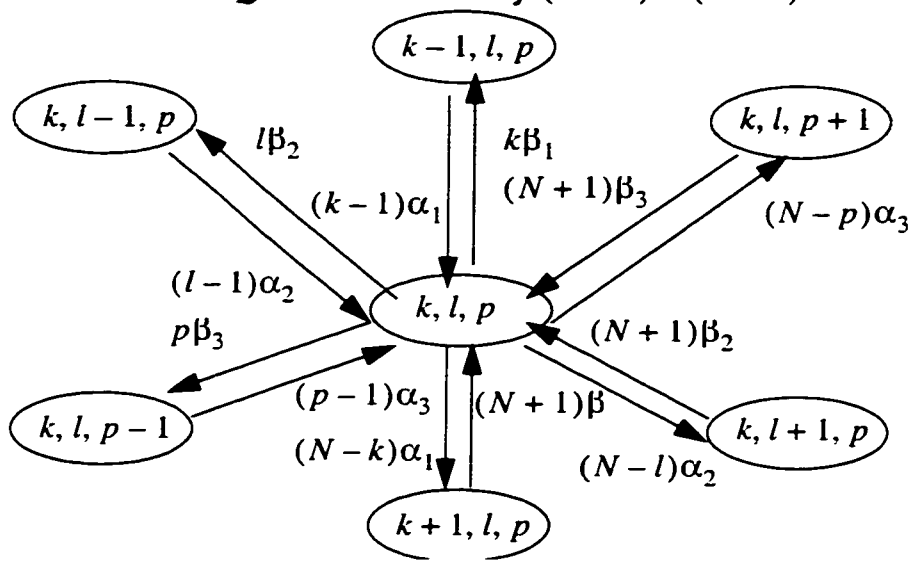

FIGURE.7.1. local state transition diagram for a three types of source model

7.3 Prediction of probability to overload in a round trip delay time for a bufferless multiple class resources

The issue here is that we have real time traffic multiplexed on a channel of capacity C and we want to find a way to prevent the system from going to an overload state in a round trip delay time. That is, we want to keep our system in the underload region. Overload occurs when the sources input rates from all active source are greater than the available output capacity C.

In section 7.3.1, we find the matrix form that governs the bufferless multiple class sources case. In section 7.3.2, we show how to arrange the matrix Q according to increasing values of the rate. In section 7.3.3, we find the probability of the system in reaching the overload in a round trip delay using the heterogeneous ON-OFF source model or the multiclass ON-OFF source case. That is, we treat the problem of multiplexer overload and give the probability for the time to overload as a solution to a set of differential equations. In section 7.3.4, we find the safe operating region that satisfies a certain probability of overload in a round trip delay.

7.3.1 Matrix form that governs the bufferless multiple class resources case

There are m classes and within class the sources are identical. There can be any number of sources within a class, however, for our case we assume there is only three non-identical classes and one source per class. We have multidimensional birth and death processes. Sources of the ith type are characterized by transition probability rates α_i and β_i, and peak transmission rate during active periods R_i. Also, we assume the system has no buffer and the sources are fed directly to the multiplexer with output channel capacity denoted by C.

Therefore, for the bufferless case, equation (7.1) becomes,

$$\sum_{i=1}^{m} \frac{dp_{\dots\dots}(t)}{dt} = \sum_{i=1}^{m} [\alpha_i(N_i - n_i + 1)]p_{\dots n_i - 1 \dots}(t) - \{\alpha_i(N_i - n_i) + \beta_i n_i\}p_{\dots\dots}(t) + \beta_i(n_i + 1)p_{\dots n_i + 1 \dots}(t) \qquad (7.10).$$

$$(0 \le n_i \le N_i; \quad i = 1, 2, \dots\dots m)$$

which can have the following matrix form,

$$\frac{d\boldsymbol{p}(t)}{dt} = \boldsymbol{Q}\boldsymbol{p}(t) \qquad (7.11).$$

where $\boldsymbol{p}(t)$ is a column vector equal to $[p_{00\ldots0}(t), \ldots\ldots\ldots\ldots, p_{n_1 n_2 \ldots\ldots n_m}(t)]^T$ and $\boldsymbol{Q}$ is the infinitesimal generator matrix.

In the heterogeneous case, the matrix $\boldsymbol{Q}$ comes out in non-ordered form. Each state is a combination of a number of sources that are active, which corresponds to a unique total rate. We need to find a way to enumerate or order the states based on numbering by rate. We may now speak about "row-numbering" or "column-numbering".

7.3.2 Row-column numbering of the matrix $\boldsymbol{Q}$

Assuming that there are N sources, each described by three heterogeneous binary ON-OFF source models. Let n_i, $i = 1, 2, \ldots . m$ indicate the number of type i sources that are active, which have values of $0, 1, 2, \ldots .., N$. Also, let the binary sources peak rates have values of $R_1, R_2, R_3, \ldots . ., R_m$. Moreover, let $R_1 < R_2 < R_3 \ldots\ldots . < R_{m-1} < R_m$. Let the number of active sources of each type describe the state of our system. The state of the system S is given by,

$$S = (n_1 + 1) + n_2(N+1) + n_3(N+1)^2 + n_4(N+1)^3 + \ldots\ldots\ldots + n_m(N+1)^{m-1} \qquad (7.12).$$

The total rate R_T is,

$$R_T = n_1 R_1 + n_2 R_2 + n_3 R_3 + \ldots\ldots\ldots + n_m R_m \qquad (7.13).$$

For the three class system $(m = 3)$, there are $(N+1)^3$ states and rates.

For example, let $N = 1$, $m = 3$, and $R_1 = 1.3, R_2 = 2.2, R_3 = 3.1 (R_1 < R_2 < R_3)$. Then, state number S, number of type $i = 1, 2, 3$ sources that are active n_1, n_2, n_3 and the total rate R_T are given below in Table 7.1.

state number S	n_1	n_2	n_3	total rate R_T
0	0	0	0	0.0
4	0	0	1	3.1
2	0	1	0	2.2
1	0	1	1	5.3
6	1	0	0	1.3
5	1	0	1	4.4
3	1	1	0	3.5
7	1	1	1	6.6

TABLE 7. 1

In general, for N source three class system $(m = 3)$, state number S, the number of type $i = 1, 2, 3$ sources that are active n_1, n_2, n_3 and the total rates R_T are given in Table 7.2.

State number S	n_1	n_2	n_3	Total rate R_T
0	0	0	0	0
1	0	0	1	R_3
2	0	0	2	$2R_3$
.	.	.	.	.
.	.	.	.	.
$2N$	0	1	N	$R_2 + NR_3$
$2N + 1$	0	1	0	R_2
$2N + 2$	0	1	1	$R_2 + R_3$
.	.	.	.	.
.	.	.	.	.
$(N + 1)(N + 1) + 1$	1	0	0	R_1
$(N + 1)(N + 1) + 2$	1	0	1	$R_1 + R_3$
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
$(N + 1)(N + 1)(N + 1)$	N	N	N	$NR_1 + NR_2 + NR_3$

TABLE 7. 2

As can be seen from Table 7.2, in contrast to the homegenous case, the rates are not in increasing order. This complicates the task of calculating the probability distribution function for the time to overload in a round trip delay time since there are multiple boundaries between overload and underload states. Our solution is simply to reorder the states so that increasing state number implies non-decreasing rate.That is, all overload states are within the boundary. Now, in order to proceed, we need a way to enumerate the states of our system so that the total rate R_T comes out in increasing order. To illustrate this, let us go back to the example for the case when N=1 and show the enumerating states S and ordering of the total rate R_T. This is shown in Table 7.3. As shown, state 7 will go to overload first, then state 3, then state 5 and so on until state 0 which is the last to go to overload.

State number S	n_1	n_2	n_3	Total rate R_T
0	0	0	0	0
1	1	0	0	1.3
2	0	1	0	2.2
3	0	0	1	3.1
4	1	1	0	3.5
5	1	0	1	4.4
6	0	1	1	5.3
7	1	1	1	6.6

TABLE 7. 3

The steps of evaluating the infinitesimal generator matrix in ordered form are as follow,

a) Input: number of sources N, channel capacity C, and parameters α_i, β_i, R_i for $i = 1, 2,....m$.

b) Compute the state number S using equation (7.12).

c) Compute the total rate R_T using equation (7.13).

d) Compute the infinitesimal generator matrix Q using equations (7.3) - (7.9) respectively. It has dimensionality of $(N+1)^m \mathrm{x} (N+1)^m$.

e) Sort the rates such that $R_{T_j} < R_{T_{j+1}}$, $j = 0, 1, 2, ...(N+1)^m$. Denote the sorted matrix by Q^s.

f) Enumerate the states of the new matrix Q^s.

7.3.3 Probability distribution function for the time to overload in a round trip time

The probability distribution for the time to overload is used as a characterization of congestion imminence in a network round-trip delay. The probability distribution for the time to overload evaluated at the network round-trip delay can be expressed in terms of the system parameters α_i, β_i, R_i, $i = 1, 2, ...m$ and the number of sources N. Equation (7.11) is a set of first order linear heterogeneous differential equations; its solution has the following form,

$$\boldsymbol{P}(\tau) = e^{\boldsymbol{Q}\tau}\boldsymbol{P}(0) \qquad (7.14).$$

where $\boldsymbol{P}(0)$ is the initial condition.

Substituting for the sorted infinitesimal generator matrix Q^s and the round trip delay τ_{rt} in equation (7.14), we have,

$$\boldsymbol{P}(\tau_{rt}) = e^{\boldsymbol{Q}^s\tau_{rt}}\boldsymbol{P}(0) \qquad (7.15).$$

$e^{\boldsymbol{Q}^s\tau}$ can be expressed as a sum involving the eigenvalues and eigenvectors of matrix $\boldsymbol{Q}^s$ [KEI64], [PER97]. Let z_j, $j = 1, 2, ...w+1$ be the eigenvalues of $\boldsymbol{Q}^s$, V_j and U_j its right and left eigenvectors respectively associated with the jth eigenvalue, normalized as

$$U_j^T V_k = \begin{cases} 1 & j = k \\ 0 & j \neq k \end{cases} \qquad (7.16).$$

Equation (7.15) now has the following solution,

$$\boldsymbol{P}(\tau_{rt}) = \sum_{j=1}^{w+1} e^{z_j \tau_{rt}} \boldsymbol{V}_j \boldsymbol{U}_j^T \boldsymbol{P}(0) \qquad (7.17).$$

Equation (7.17) depends on the round trip delay of the network τ_{rt}, on the eigenvalues and eigenvectors of matrix $\boldsymbol{Q}^s$; and therefore on the number of sources N, the channel capacity C, the number of underload states w, and on the source parameters α_i, β_i, R_i, $i = 1, 2, ...m$. It gives a measure of the probability to overload in less than a round trip delay time τ_{rt}, which in turn is a measure of congestion.

Of course, in order for the system to be in overload, it is necessary that the arrival rate from all active sources be greater than the channel capacity. To find the probability distribution for the time to overload in a round trip delay, we use the following steps,

a) Use steps *a-f* of section 7.3.2.

b) If total arrival rate from all active sources > channel capacity, then number of underload states = w.

d) Combine all other states above w into one state and call it the overload state. Total number of states = $w+1$. Call the new matrix $Q^s{}_{w+1}$.

f) Find probability distribution of time to overload in a round trip delay using equation (7.17) with the matrix equal to $Q^s{}_{w+1}$.

g) Repeat for the calculation for the probability distribution of time to overload for different round trip delays.

Including the overload state, the matrix Q^s_{w+1} has $w+1$ column and w rows. To make the matrix square, we augment a row of all zero elements at the $w+1$ row. The computational complexity depends on the size of the matrix $Q^s{}_{w+1}$, which depends on the number of sources N, the underload states w and the channel capacity C. As N increases, the size of the matrix $Q^s{}_{w+1}$ increases. Increasing C will increase w and therefore the size of the matrix $Q^s{}_{w+1}$. In the sequel, we find the probability distribution for the time to overload for several VBR video data traces.

Numerical example

Let us consider the application of equation (7.17) to several real video traffic data traces. The data available to us is four VBR video traces that are of different scene types, video-conferencing, video-phone, popular TV series, and Movie. The video traces are generated at a rate of 25 frames/sec. If C, the channel capacity is expressed in bits/s, then mean (peak) rate must be converted from cells per frame by multiplying it by number of cells generated per frame, number of bytes per cell and number of bits per byte. As we have seen in section 2.3.1, video-con-

ferencing and video-phone have medium Hurst parameters 0.72 and 0.74 respectively and low mean and peak rates. The other two, TV series and Movie have high Hurst parameters of 0.9 and 0.96 respectively and large mean and peak rates. For more detail, see Table 2.3 and Table 2.4.

To have an idea of the self-similarity effect of the video data, we reproduce figure 2.14 of chapter 2 as in figure 7.2,. We show the autocorrelation function for the four video data traces versus the lag. The autocorrelation function for Movie has the lowest decay rate while video-conferencing has the highest decay rate. Accordingly, Movie is the most correlated and video-conferencing is the least correlated in our video data traces.

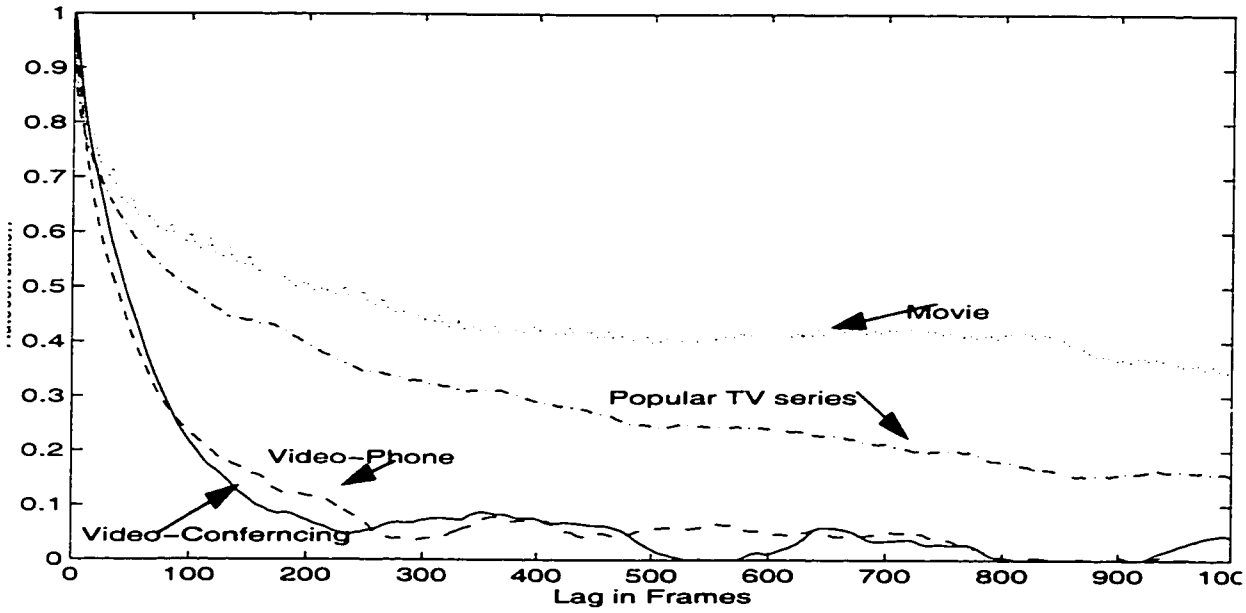

FIGURE.7.2. Autocorrelation functions for four different video data

The two highly correlated traces, popular TV series and Movie, have larger peak rate, mean rate and Hurst parameter as compared to the less correlated traffic video-conferencing and video-phone. As can be seen from Table 7.4, the peak (mean) rate for the highly correlated traffic such as Movie is more than 4 (10)

times bigger than the peak rate of the medium correlated traffic such as video-conferencing. So, because of their different scene types and statistics, comparison between teleconferencing and entertainment data is not useful in getting information about the behavior of the multiplexed traffic such as comparing the probability distribution to overload in a round trip delay time. We compare those traces that have comparable peak (mean) rates and comparable Hurst parameters; we compare video-conferencing with the video-phone and the popular TV series with the Movie.

VBR Video sequence	**mean number of cells/frame [Mbits/s]**	**peak number of cells/frame [Mbits/s]**
video-conferencing	130.29 [1.67]	629.0 [8.05]
video-phone	170.61 [2.18]	897.0 [11.48]
TV series	5336.4 [14.94]	11801.0 [33.04]
Movie	5948.4 [16.7]	13325.0 [37.31]

TABLE 7. 4 Mean and peak number of cells per frame of VBR video data

We consider multiplexing a number of video sources over a channel of capacity $C = 10$ Mbits/s for video-conferencing, video-phone and on a OC-1 link of channel capacity $C = 51.84$ Mbits/s for popular TV series and Movie. The choice of the channel was based on the bit rate of the data; video-conferencing and video-phone have low bit rates as compared to TV series and Movie. Each video source can be modeled by three heterogeneous ON-OFF source model with parameters $\alpha_i, \beta_i, R_i,\ i = 1, 2, 3$ [FAR98]. Figure 7.3 gives the probability distribution for the time to overload for video-conferencing and video-phone data for different values of $n_1,\ n_2,\ n_3$, the number of active sources of each class with $N = 6$ and $C = 10$ Mbits/s. Probability distribution for the time to overload for TV

series and Movie data for different values of n_1, n_2, n_3 are shown in figure 7.4 with $N = 5$ multiplexed on a OC-1 link of channel capacity $C = 51.84$ Mbits/s.

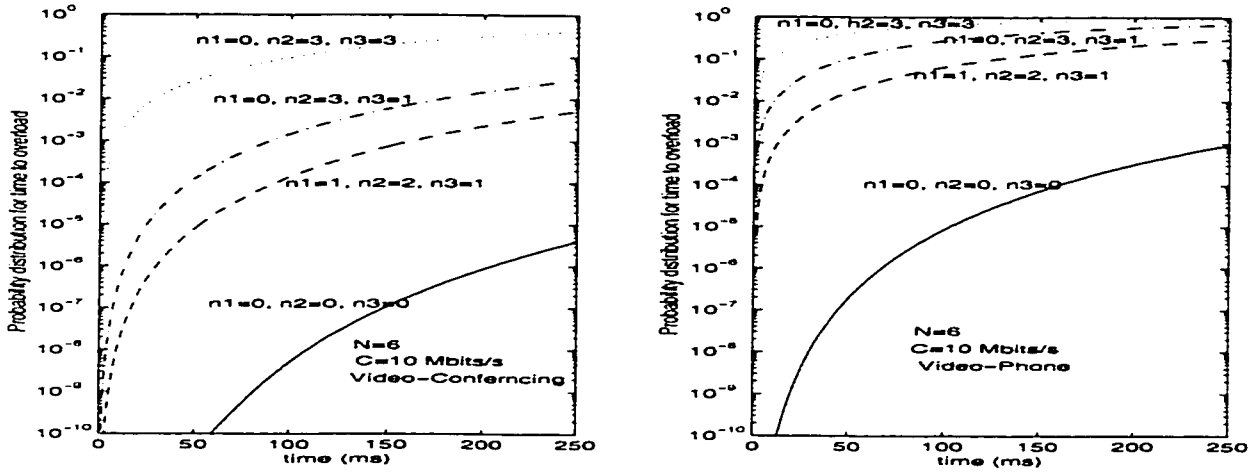

FIGURE.7.3. Probability distribution for the time to overload in a round trip delay for video-conferencing and video-phone data for different values of n1, n2, n3; number of sources $N = 6$, channel capacity $C = 10$ Mbits/s.

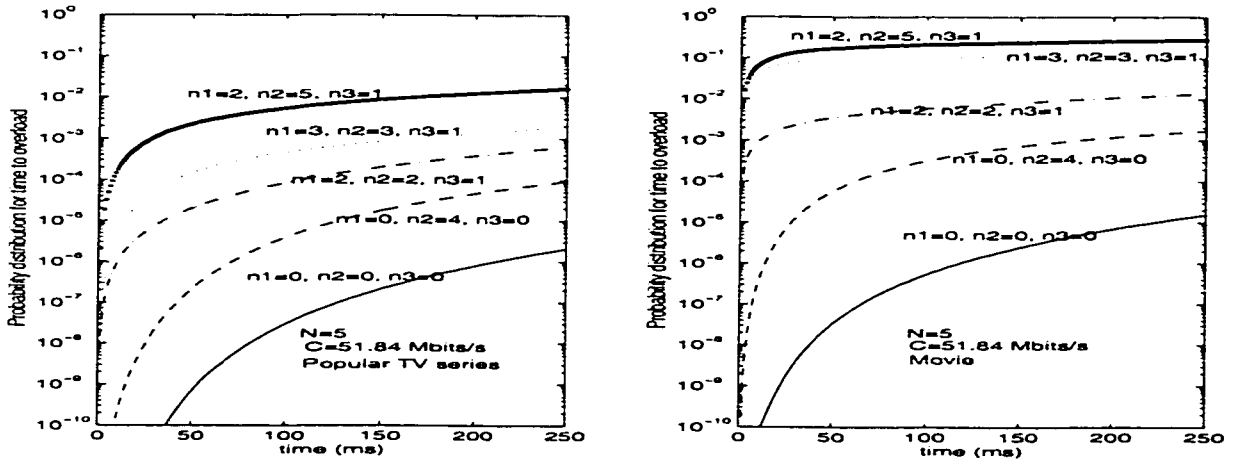

FIGURE.7.4. Probability distribution for the time to overload in a round trip delay for popular TV series, and Movie data for different values of n1, n2, n3; number of sources $N = 5$, channel capacity $C = 51.84$ Mbits/s

There is a relation between round trip delay and time until overload. If round trip delay is less than time until overload, then there is no problem and network can operate safely. However, if round trip delay is larger than time until overload,

congestion can not be avoided and some correction action must be taken in order to avoid congestion.

For networks with short round trip delays, low values for the probability to overload are possible. As the round trip delay increases, the probability distribution for the time to overload increases. It is clear for the heterogeneous case that, the probability distribution for the time to overload depends heavily on the mix of the active sources of the different classes. Moreover, the round trip delay τ_{rt} is compatible with the time constant τ_c (see Table 4.4 for more details on the time constants of the VBR data); i.e, if the delay is within the time constant of the traffic, then low values of probability to overload are possible.

It is interesting to see that the probability distribution to overload in round trip delay increases for more correlated traffic; for the same probability to overload in a round trip delay, highly correlated traffic goes to overload faster than low correlated when multiplexed over the same channel capacity. This is can be seen from figure 7.3, when we compare video-conferencing with video-phone (multiplexed over 10 Mbits/s channel) or from figure 7.4 when we compare TV series with Movie (multiplexed over 51.84 Mbits/s channel). So, there is a difficulty in multiplexing highly correlated traffic compared to low correlated video data.

7.4 Admission control

The objective of admission control is to limit the number of admitted sources into the network so the probability to overload in a round trip delay can be kept below a specified threshold. To determine the threshold where service to an addi-

tional call must be denied, resource characterization is required. The admission control problem is to characterize sources for which the admission criterion $p(t_{rt}) \leq \varepsilon$ is satisfied where $p(t_{rt})$ denotes the probability to overload in less than a network round trip time t_{rt} and ε is a very small value. The problem here is to characterize sources for which the buffer overflow probability $p(\varepsilon)$ does not exceed ε. When the total input rate from the active sources to the network is below the threshold there is no need to send feedback signals to the sources to slow down since congestion is not immanent. However, above the threshold, signals must be sent back to the sources to slow down.

To find the safe operating region that satisfies a certain probability to overload in a round trip delay time ε, we use steps that were used in section 7.3.3 with minor modification and we end up with the following,

a) Use steps *a-d* of section 7.3.3.

b) Search for the rate that gives us probability to overload in less than a network round trip time that satisfies the QoS criteria ε.

d) Repeat for the calculation for the total rate that satisfies the QoS criteria for different round trip delays.

Numerical example

We consider the same VBR video data used in the example presented in section 7.3.3. We consider two cases for the channel capacity. For video traffic that has low bit rates such as video-conferencing and video-phone, we use channel of capacity 10 Mbits/s, while for traffic with high bit rates such as TV series and

Movie, we use OC-1 link of channel capacity $C = 51.84$ Mbits/s. Comparison for other channel capacity is possible. Also, we consider two cases for the performance index ε; $\varepsilon = 10^{-2}$ and $\varepsilon = 10^{-5}$.

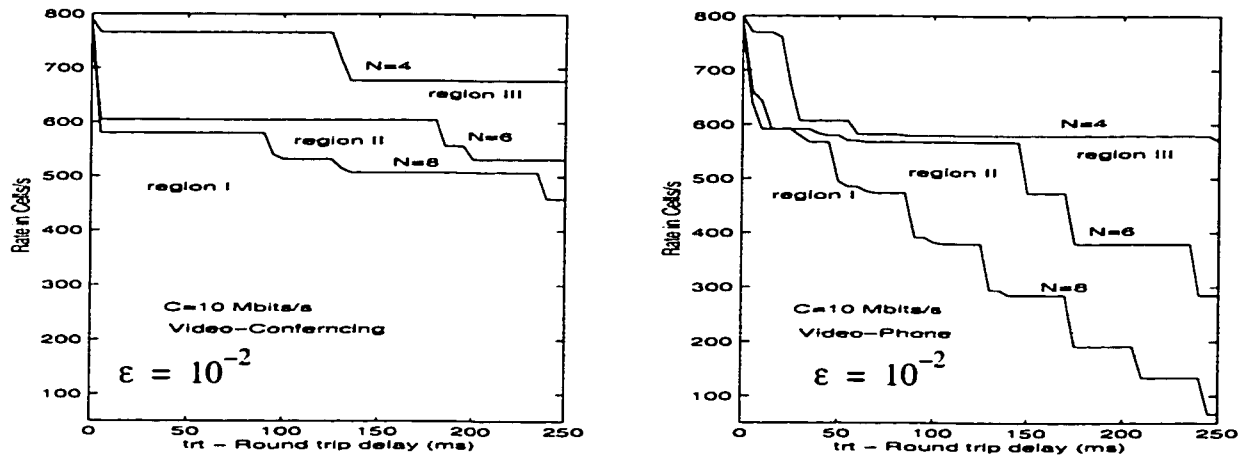

FIGURE.7.5. Operating regions for video-conferencing and video-phone sources with parameters, N the number of multiplexing video sources; channel capacity $C = 10$ Mbits/s; probability to overload in a round trip delay time of value $\varepsilon = 0.01$

Figure 7.5 depicts the safe operating region for VBR video-conferencing and video-phone as a function of the number of the input source N, $N = 4, 6, 8$ to the network of channel capacity C=10 Mbits/s and the probability to overload in a round trip delay time $\varepsilon = 10^{-2}$. As shown, increasing the number of admitted sources N to the network, results in shrinking the safe operating region. Also, as the round trip delay decreases, the number of sources that can be admitted to the network increases.

For a tighter criterion for the probability to overload, e.g; $\varepsilon = 10^{-5}$, the operating region shrinks and in this case correction measures are needed in a great extent. Figure 7.6 depicts safe operating region for VBR video-conferencing and video-phone as a function of the number of the input sources N, $N = 4, 6, 8$ to

the network of channel capacity C=10 Mbits/s and the probability to overload in a round trip delay time $\varepsilon = 10^{-5}$.

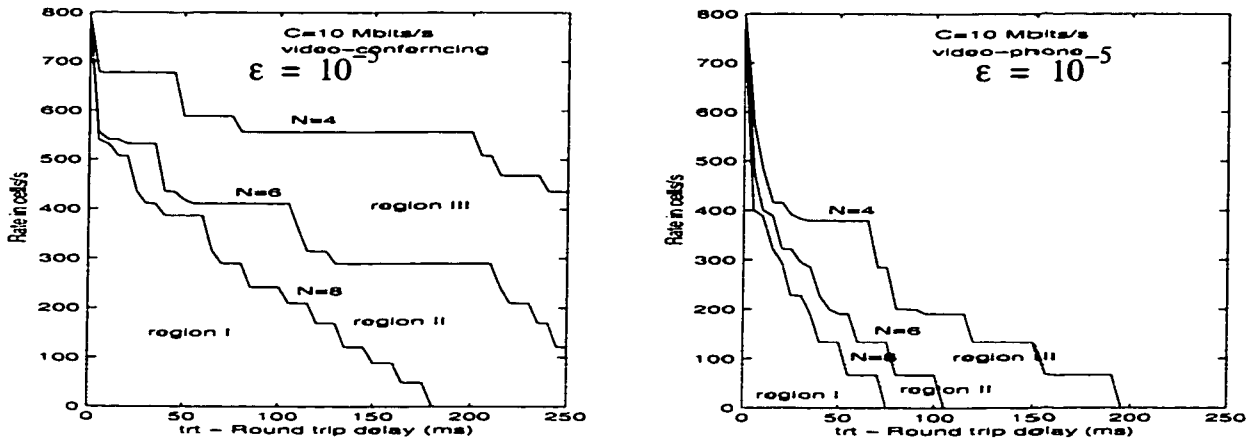

FIGURE.7.6. Operating regions for video-conferencing and video-phone sources with parameters, N the number of multiplexing video sources; channel capacity C = 10 Mbits/s; probability to overload in a round trip delay time of value ε = 0.00001

In figure 7.7, we show the safe operating region for VBR TV series and Movie as a function of the number of the input sources N, $N = 4, 6, 8$ to the network multiplexed on OC-1 link of channel capacity C = 51.84 Mbits/s and the probability to overload in a round trip delay time $\varepsilon = 10^{-5}$. Figure 7.8 depicts the safe operating region for VBR TV series and Movie as a function of the number of the input sources N, $N = 4, 6, 8$ to the network multiplexed on channel capacity C=20 Mbits/s and the probability to overload in a round trip delay time $\varepsilon = 10^{-2}$. Comparison between TV series and Movie shows that, for the same channel capacity and same number of admitted sources, the safe operating region for Movie is less than that for TV series. This is mainly due to the fact that Movie bit rates are larger than TV series. Also, this may be due to higher correlation of Movie traffic as compared to TV series traffic. The same explanation mentioned

above applies for the other two video data traces, video-conferencing and video-phone.

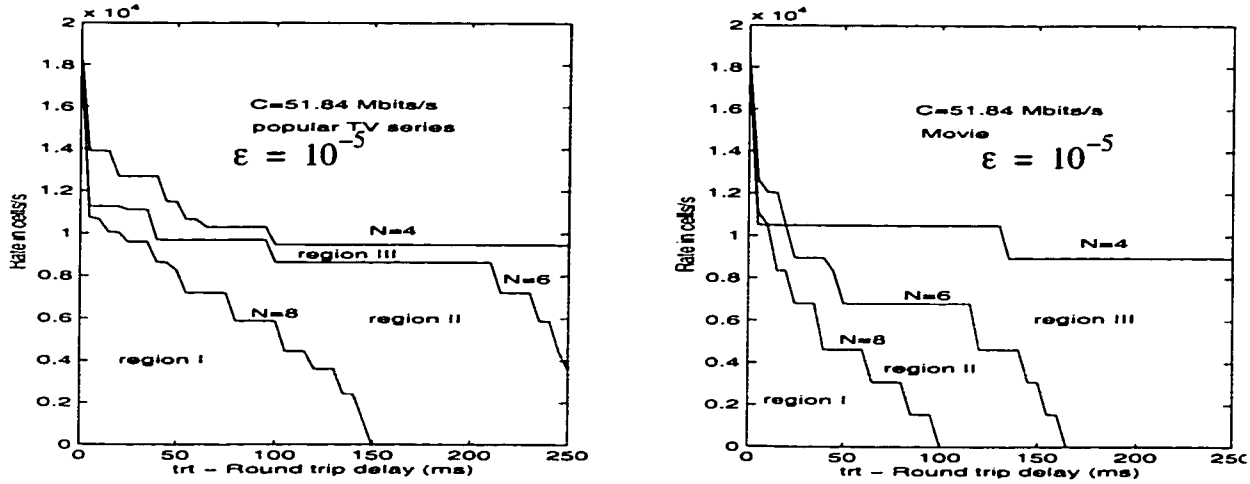

FIGURE.7.7. Operating regions for popular TV series and Movie sources with parameters, N the number of multiplexing video sources; channel capacity C = 51.84 Mbits/s; probability to overload in a round trip delay time of value $\varepsilon = 0.00001$

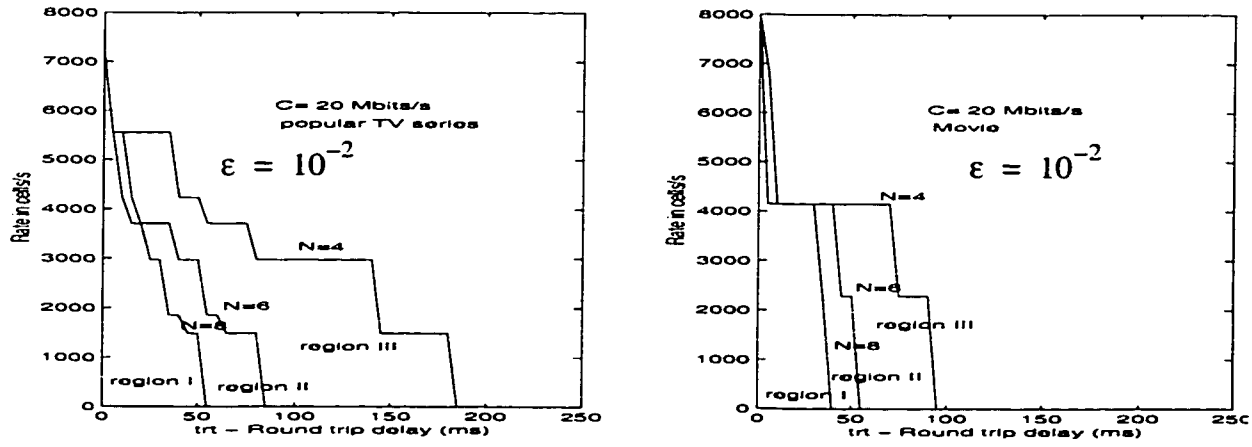

FIGURE.7.8. Operating regions for popular TV series and Movie sources with parameters, N the number of multiplexing video sources; channel capacity C = 20 Mbits/s; probability to overload in a round trip delay time of value $\varepsilon = 0.01$

For the same number of input sources N and same QoS criteria ε, increasing the channel capacity C will increase the safe operating region. Therefore, it is possible to have more sources admitted to the network if we increase the channel capacity C to satisfy the same QoS criteria ε. In figure 7.9, for $N = 4$ and

$\varepsilon = 10^{-5}$ safe operating regions are shown for the channel capacity C = 5 Mbits/s, 7.5 Mbits/s and 10 Mbits/s.

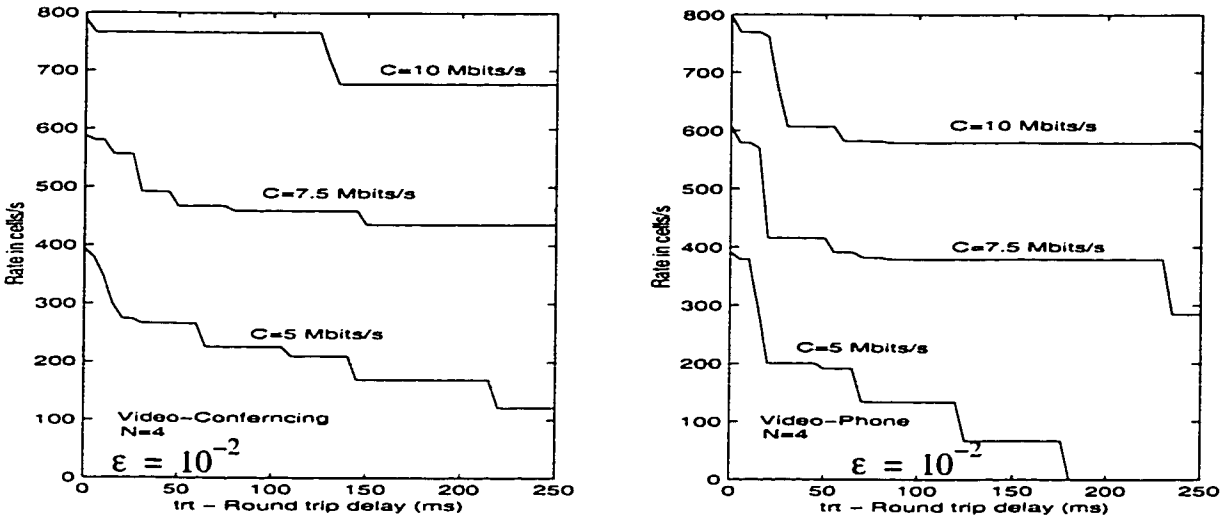

FIGURE.7.9. Operating regions for video-conferencing and video-phone sources for N = 4 sources; channel capacity C = 5 Mbits/s, 7.5 Mbits/s,10 Mbits/s; probability to overload in a round trip delay time of value $\varepsilon = 0.01$

For a network with a given QoS criteria ε, there is a trade-off between the number of admitted sources and the round trip delay. Given ε, if round trip delay increases which affects the safe operating region, then number of sources admitted has to be decreased such that the system meets the desired QoS criteria, and vice versa. So, if we need to operate in a specific safe operating region to satisfy a certain probability of overflow for a given network round trip delay, we need to control the number of admitted sources to the network by sending some corrective actions back to the input of the multiplexer. How much earlier the control action should be taken depends mainly on propagation delay, traffic burstiness and the rate that is in effect. Finally, as we have mentioned in section 7.3 when we dis-

cussed the probability to overload in a round trip delay, for admission control the round trip delay τ_{rt} is compatible with the time constant τ_c.

7.5 Discussion

We considered multiplexing a number of video source over a channel of capacity $C = 10$ Mbits/s for video-conferencing, video-phone and on a OC-1 link of channel capacity $C = 51.84$ Mbits/s for popular TV series and Movie. The probabilities distributions for the time to overload in a round trip delay for different VBR video data using a heterogeneous ON-OFF source model are found. For the same round trip delay, highly correlated traffic have more probability to overload than low correlated traffic when multiplexed over the same channel capacity. Admission control is also considered. We considered two cases for the performance index ε; $\varepsilon = 10^{-2}$ and $\varepsilon = 10^{-5}$. For a network with a given QoS criteria ε, there is a trade-off between the number of admitted sources and the round trip delay. The number of admitted sources are adjusted in order to have a certain QoS criteria.

CHAPTER VIII

Conclusions and Further Research

In this thesis, we have presented modelling and simulation schemes for self-similar traffic. We have generated synthetic self-similar traffic based on simulation and analytical models of the real Ethernet and VBR video data that are available to us. We presented the conventional traffic models in order to characterize and predict the performance measures for real Ethernet and VBR video data. We have extensively investigated the statistical properties of real Ethernet and video data. We proposed two procedures to fit the PMPP model to the real Ethernet data and MMPP model to the video data. The fitting of simulation models such as FGN, FBM and F-ARIMA to the real Ethernet data is excellent and provides good prediction for the probability of loss and mean queue length. Markov chain has good fitting and performance prediction for teleconferencing video data, however, it is not good for entertainment video data which has high correlation index. Multiplexing a number of sources will improve the fitting and the prediction of the performance measures. Moreover, we have proposed a new model, the heterogeneous ON-OFF source model, to model VBR video. The proposed model has shown a better characterization and better prediction of the QoS for the correlated video traffic that we discussed, than the other conventional traffic models available in the literature. The proposed model is very simple and tractable. The basic idea of the 3-class model is that there are three time frames for transitions: short term, medium term and long term, respectively. The transition rates are such that $\alpha_1 \gg \alpha_2 \gg \ldots\ldots \gg \alpha_m$ and $\beta_1 \gg \beta_2 \gg \ldots\ldots \gg \beta_m$, where for our model we have $m = 3$, so that the shorter the time frame, the more rapid the transition. Our analysis is based on second order statistics. We approximated the non-exponential covari-

ance function of the correlated traffic by exponential function. The idea of approximating equation (6.3) by equation (6.9), because performance models with component long-tail distributions tend to be difficult to analyze. The number of parameters that are needed for the model is not large. The small number of parameters makes the analysis (finding the covariance and the parameters of the sources) simple. The heterogeneous ON-OFF source model that we have developed is a considerable improvement on the Maglaris model; accordingly it is applied to the congestion control problem. We have calculated the probability distributions for the time to overload in a round trip delay time. We also considered admission control where for a network the number of admitted sources is adjusted in order to have a certain QoS.

The results that we obtained in chapter 2 through chapter 7 from applying the different statistical and analytical models to the real Etherenet and real VBR video data can conclude the following:

- Self -similar traffic can be characterized based on second order statistics, the covariance and the IDC. If the matching of the covariance and the IDC to the data is good then we will have a good prediction of the probability of loss and mean queue length.

- Simulation models such as FGN, FBM and F-ARIMA models can accurately characterize Ethernet data. Moreover, PMPP can be used to characterize the data and give reasonable results for practical engineering design, however, it is not as good as FGN, FBM and F-ARIMA

- Multiplexing several statistically independent and identical highly correlated sources will result in a reduction of the probability of loss and mean queue length. The advantage of multiplexing is due to smoothing the peaks and valleys of traffic as a result of averaging.

- The choice of the best model is based on the best matching of the probability of loss. Also it must have a small number of fitting parameters to make it analytically tractable and at the same time have the same properties of the actual sources

- The best analytical model that is effective for video data is the heterogeneous ON-OFF source model. This model can characterize traffic that has different Hurst parameter as compared with the other analytical models such as MMPP and Markov chain models. The heterogeneous ON-OFF source model is analytically tractable and simple. However, MMPP and Markov chain models can preform good prediction when the index of correlation is not high and the traffic is not heavy.

This work can be explored in many directions and possible extensions of the results obtained so far including the following:

- Working with real data what are the parameters that count in calculation of performance? Two different self-similar traffic models with the same Hurst parameter H may result in very different tail asymptotic and very different probability of loss and queue length.This indicates that we have to look at the impacts of other factors, such as H combined with higher-order statistics.

- Given the self-similar traffic as an input, the concern is to look how it behaves with the leaky Bucket. Further, studying leaky Bucket performance, including characterizing its output process and deriving the cell loss probabilities in the network.

- The existence of self-similar behaviour may require an examination of current flow control mechanisms. As the amount of network traffic increases

and data networks become even more and more common it is important that the problems of packet delivery and error correction using flow control are addressed.

- Network protocol behaviour is extremely complicated in real life because of the complex interaction within the operating system. A simple model of a network and protocols which produces self-similar behaviour would be a valuable tool in understanding what is happening.

- An obvious extension of this work will be to analyze more video and Ethernet data to determine the consistency and generality of these results. Also, it would be important to develop the admission control techniques more fully by studying their effectiveness using simulation.

Appendix A

Glossary

ABR Available Bit Rate
ATM Asynchronous Transfer Mode
AAL ATM Adaptation Layer
BELLCORE Bell Communication Research
B-ISDN Broadband Integrated Services Digital Networks
BMAP Batch Markov Arrival Processes
CAC Connection Admission Control
CBR Constant Bit Rate
COV Covariance of independent m class ON-OFF sources
CO Covariance of the number of packets of a long-tail process
CCITT International Consultative Committee for Telephone and Telegraph
CLP Cell Loss Priority
FARIMA Fractional Autoregressive Moving Average
FBM Fractional Brownian Motion
FGN Fractional Gaussian Noise
GEO Geosynchronous Earth Orbit
GFC Generic Flow Control
Gbit/s Gigabit per second
H Hurst parameter
HEC Header Error Check
IDC Index of Dispersion for counts
IDI Index of Dispersion for intervals
IID Independent Identically Distributed
ISDN Integrated Services Digital Networks
Kbit/s kilobit per second
LAN Local Area Network
LEO Low Earth Orbit
LRD Long Range Dependence
MAP Markov Arrival Processes
MEO Medium Earth Orbit
MMPP Markov Modulated Poisson Processes
Mbit/s Megabit per second
N-ISDN Narrowband Integrated Services Digital Networks

NNI Network-Node Interface
OC Optical Carrier
OPNET Optimized Network Engineering Tools
PMPP Pareto Modulated Poisson Process
PMR Peak to Mean Ratio
PR Priority field
PT Payload Type
QoS Quality of Service
RES Reserved
SCOV Squared Coefficient of Variation
SDH Synchronous Digital Hierarchy
SONET Synchronous Optical Network
SRD Short Range Dependence
TDM Time Division Multiplexing
UNI User-Network Interface
Var Variance of independent m class ON-OFF sources
VBR Variable Bit Rate
VC Virtual Channel
VCI Virtual Channel Identifier
VPI Virtual Path Identifier
WAN Wide Area Networks

Appendix B

List of Symbols

a_i: Coefficient in the spectral expansion solution
A_n: Poisson interarrival times
C: Channel capacity
$C(n)$: Discrete covariance
$\boldsymbol{D}$: Fluid flow diagonal matrix
$F_i(u)$: Equilibrium probabilities that i source are ON and buffer size $\leq u$
$F(u)$: Equilibrium probability distribution for the queue length
f: Number of frames generated per second
$G(u)$: Probability of buffer overflow beyond $\leq u$
H: Hurst parameter
I_k: IDC at lag k
I_∞: IDC at infinity
J_k: IDI at lag k
k: Lag
L: Number of quantization levels
$\boldsymbol{M}$: $N \times N$ transition matrix
m: Number of classes
N: Number of ON-OFF sources
N_i: Number of ON-OFF sources in class i
n: Samples size
$n_1 n_2 \ldots . n_m$: State with n_i source in class i ON
$N(t)$: Number of arrivals in $(0, t)$
P: Transition probabilities matrix
p_{ij}: Number of transition from i to j / number of transitions out of i
$p_i(t, u)$: Probability that, the queue length does not exceed u and i sources are active at time t
p: Order of the autoregression
$p(t_{rt})$: Probability to overload in less than a network round trip time t_{rt}
$\boldsymbol{Q}$: Infinitesimal generator matrix
$\boldsymbol{Q}^s$: Sorted Infinitesimal generator matrix
$q(k, l, p; x, y, z)$: Transition rates from state k, l, p to state x, y, z.
R: Data rate when source is in the ON state
R_T: Total Rate

$r(k)$: Autocorrelation at lag k
S_T: State of the system
t: Time
U_j: Left eigenvectors
u: Buffer size
V_j: Right eigenvectors
w: Number of underload states
X: WSS process
$X^{(n)}$: Aggregated WSS process
z_i: Eigenvalues
ν: Variance
μ: Mean value
μ_3: Third moment
φ_i: Eigenvectors
ρ_k : Covariance
σ: Standard deviation
$C(n)$: Discrete covariance
λ: Poisson arrival rate
Λ: Markov chain rate matrix
δ: Skewness
τ_c: Time constant
τ_{rt}: Round trip delay
π: Steady-state probability
σ_i: Transition rate of going out of state i
Δt: Incremental time interval
α: Duration of ON state
β: Duration of OFF state
Δ: Quantization step
ε: Very small value

References

[AME91] J. Amenyo, A.A. Lazar and G. Pacifici,"Proactive cooperative scheduling and buffer management for multimedia networks," CTR Technical report #240-91-21, New York, Aug. 1991.

[AND97] A.T. Andersen and B. Nielsen, "An application of superpositions of two state Markovian sources to the modelling of self-similar behavior," IEEE INFOCOM 97, April1997, Kobe, Japan.

[ANI82] D.Anick, D Mitra and M. M. Sondhi, "Stochastic theory of a data handling system with multiple sources," Bell Syest. Tech. J. 61 (1982), 1871-1849.

[ARN85] B. C. Arnold, "Pareto distribution," in Encyclopedia of statistical sciences, S. Kotz and N. Johnson Eds. New York: Wiley, 1985, vol. 6, pp. 569-574.

[BAE91] J. J. Bae and T. Suda, "Survey of traffic control schemes and protocols in ATM networks," Proceedings of the IEEE, vol. 79, no. 2, pp. 170-189, Feb. 1991.

[BAI91] A. Baiocchi, N. B. Melazzi, M. Listanti, A. Roveri, R. Winkler, "Loss performance analysis of an ATM multiplexer loaded with high speed ON-OFF sources," IEEE JSAC. Comm. vol. 9, no. 3, pp. 388-393, 1991

[BAI92] A. Baiocchi, N. B. Melazzi, A. R. Salvatore, "Stochastic fluid analysis of an ATM multiplexer loaded with heterogeneous ON-OFF sources, an efficient computational approach," INFOCOM 92, Florence, Italy, 3C.3.1-10, 1992.

[BER90] A. W. Berger, "Performance analysis of a rate control throttle where tokens and jobs queue," in Proc. IEEE INFOCOM'90, pp. 30-38, 1990.

[BER95] J. Beran, R. Sherman, M. S. Taqqu and W. Willinger, "Long-range dependence in variable-bit video traffic, "IEEE trans. Commun. (accepted for publication subject to revision), 1994.

[BRA69] P. Brady, "A model for generating ON-OFF speech in two-way conversations," Bell Syst. J., vol. 48, Sept. 1969.

[CID88] I. Cidon and I. S. Gopal, "PARIS: An approach to integrated high speed private networks", Int. J. Digital and Analog Cabled Systems, vol. 1., pp. 77-86, Apr.-June 1988.

[COL96] B.R. Collier and H. S. Kim, "Traffic rate and shape control with queue threshold congestion recognition", ICC 96, pp. 746-750, June 23-27, 1996

[COX84] D. R. Cox, "Long-range dependence: A review", in statistics: An appraisal, Proceedings 50th Anniversary Conference, Iowa State Statistical Library, H. A. David and H. T. David, editors, Iowa State University Press, pp. 55-74, 1984.

[DUF94] D. Duffy, A. McIntosh, M. Rosenstein, and W. Willinger, "Statistical analysis of CCSN/SS7 traffic data from working CCS subnetworks", IEEE JSAC, 12(3), pp. 544-551, April, 1994.

[DAI86] J. Daigle and J. Langeford, "Models for analysis of packet voice communication systems," IEEE J. Sel Areas Commun., Sept. 1986.

[DEP95] M. Deprycker, "Asynchronous Transfer Mode," Chichester, U.K.: Elis Horwood, simon and Schuster Int'l, 1995.

[DUF93] N. G. Duffield, N. O'Connell, "Large deviations and overflow probabilities for the general single-server queue, with applications", IEE Eleventh UK Teletraffic Symposium Performance Engin. in Telecom. Networks. pp 30 / 1-8, 1993.

[ELW94] A.I. Elwalid and D. Mitra, "Effective bandwidth of general Markovian traffic sources and admission control of high speed networks," IEEE/ACM tran. Networking, vol. 1, no. 3, pp. 329-343, June 1993.

[ELW94] A.I. Elwalid and D. Mitra, "Statistical multiplexing with loss priorities in rate-based congestion control of high speed networks," IEEE Trans. on Comm., vol. 42, no. 11, pp 2989-3002, November 1994.

[FAR96] R. Faraj, J. F. Hayes, "The characterization of self-similar traffic in ATM network", 18th Biennial Symposium on Communications, Kingston, Ontario, June 2 - 5, 1996.

[FAR98] R. Faraj, J.F. hayes, "An application of heterogeneous ON-OFF sources to the modeling of self-similar traffic," 19th Symposium'98 on Comm., Kingston, Ontario, pp48-51., May 31 to June 3 1998.

[FAR99] R. Faraj, J.F. Hayes,"Congestion and admission control of self-similar traffic based on multiple types on-off sources"IEEE CCECE, May 9-12, 1999, Edmonton, Alberta, Canada

[FEL97] A. Feldmann and W. Whitt, "Fitting mixtures of exponentials to long-tail distributions to analyze network performance models," IEEE INFOCOM'97, April1997, Kobe, Japan.

[FOW91] H. J. Fowler and W.E. Leland, "Local area network traffic characteristics, with implications for broadband networks congestion management", IEEE JSAC. Comm. 9, pp. 1139-1149, 1991.

[GAL89] G. Galassi, G. Rigolio, and L. Fratta, "ATM: Bandwidth assignment and bandwidth enforcement policies," in Proc. IEEE GLOBECOM '89, pp. 49.61-49.6.6.

[GAL90] G. Galassi, G. Rigolio, and L. Verri, "Resource management and dimensioning in ATM Networks," IEEE Network Mag., vol. 4, N0. 3,pp. 8-17, May 1990.

[GUS91] R. Gusella, "Characterizing the variability of arrivals process with indexes of dispersion," IEEE JSAC, 9(2),pp. 203-211, Apr. 1991.

[HAY84] J.F. Hayes, "Modelling and analysis of computer communications networks", New York: Plenum Publishing Co., 1984.

[HAN89] R. Handel, "Evolution of ISDN towards broadband ISDN," IEEE Network Magazine, pp. 7-13, Jan. 1989.

[[HEF80] H. Heffes, "A class of data traffic processes-Covariance function characterization and related queueing results," BSTJ, Vol. 59, No. 6, July-August 1980, pp. 897-929.

[HEF86] H. Heffes, and D. M. Lucantoni, "A Markov modulated characterization of packetized voice and data traffic and related statistical multiplexer performance", IEEE JSAC, vol. SAC-4, no 6, pp. 856-868, Sep. 1986.

[HEY92] D. Heyman, A. Tabatabai, and T. V. Lakshman, "Statistical analysis and simulation study of Video Teletraffic in ATM networks," IEEE Trans. Circuits and Systems for Video Technology, vol. 2, 1992, pp. 49-59

[HEY96] D. Heyman, and T. V. Lakshman, "Source models for broadcast video traffic," IEEE/ACM Trans. Networking, vol. 4, no. 1, Feb 1996.

[HEYM96] D. Heyman, and T. V. Lakshman, "What are the implications of long-range dependence for VBR-video traffic Engineering ?," IEEE/ACM Trans. Networking, vol. 4, no. 3, June 1996.

[HOL92] R. Holter, "Managing SONET equipment," IEEE Network Magazine, pp. 36-41, january 19992.

[HON91] D. Hong and T.Suda, "Congestion control and prevention in ATM networks," IEEE Network Magazine, pp. 10-16, July 1991.

[HOS84] J. Hosking, "Modeling persistence in hydrological time series using fractional differencing", Water Resources Research, 20 (12), 1984.

[HU95] X. Hu, J. F. Lambert and A. Pitsillides, "Fast backward predictive congestion notification for ATM networks with significant propagation delays," in Proc. IEEE Globcom'95, Boston, MA, pp.275-279.

[HUA95] C. Huang, M. Devetsikiotis I. Lambadaris and A. Kaye, "Self-similar modeling of Variable Bit Rate compressed video: A unified approach", Submitted to ACM SIGCOMM'95.

[HUA95] J. Huang, and J.F. Hayes, "A study of the matrix analytical method and its application in performance evaluation of broadband and related system," ISORA'95, Beijing.

[HUAN95] J. Huang, Tho Le-Ngoc and J.F. Hayes, "Performance of a broadband satellite communications systems", Canadian Conference on ECE., Sept. 1995.

[HUR51] H. E. Hurst, "Long term storage capacity of reservoirs", Trans. Amer. Soc. Civil Engineers, pp. 770-779, 1951.

[KEI64] J. Keilson, "A review of transient behavior in regular diffusion and birth-death processes," J.Appl. Prob., pp.247-266, 1964.

[KOB78] H. Kobayashi, "Modeling and analysis: An introduction to system performance evaluation methodology", New York: Addison-Wesley Publishing Company, Inc. 1978.

[LAW72]A.J. Lawrence, "Some models for stationary series univariate events", Stochastic Point processes: Statistical Analysis Theory and applications, pp. 199-256, 1972

[LEL91] W. E. Leland, D. V. Wilson, "High time-resolution measurement and analysis of LAN traffic: Implications for LAN interconnection," in Proc. IEEE INFOCOM'91, Bal Harbour, FL, Apr. 1991, pp. 1360-1366.

[[LEL93] W. E. Leland, M. S. Taqqu, W. Willinger and D. V. Wilison, "On the self-similar nature of Ethernet traffic," Proc. of the ACM/SIGCOMM'93, San Francisco, CA, pp. 183-193, 1993.

[LEL94] W. E. Leland, M. S. Taqqu, W. Willinger and D. V. Wilison, "On the self-similar nature of Ethernet traffic (extended version)", IEEE/ACM Transaction on Networking, pp. 1-15, February 1994.

[LUC91] D. Lucantoni, "new results on the single server queue with a batch Markovian Arrival Processes," Commmun.Statist. -Stochastic Models, 7(1):1-46, 1991.

[MAN68]B. B. Mandelbrot and J. W. Van Ness, "Fractional Brownian Motions, Fractional Noises, and applications", SIAM review 10, pp. 422-437, 1968.

[MAG88] B. Maglaris, D. Anastassiou, P. Sen, G. Karlsson, J. Robbins, "Performance Models of statistical multiplexing in packet video communication," IEEE Trans. Commun., April 1988.

[MIL31] OPNET modeling manual vol.1, OPNET version 3.5, MIL 3 inc., 1997.

[MIT88] D. Mitra, "Stochastic theory of a fluid model of producers and consumers coupled by a buffer," Adv. Appl. Prob. 20 (1988), 646-676.

[NOR94] I. Norros, "A storage model with self-similar input", Queueing Systems Theory and Application. vol:16, Iss: 3-4, pp. 387-96, 1994.

[NEU79] M. F. Neuts, "A versatile Markovian point process", J.Appl. Prob., (16), pp 764-79, 1979.

[ONV94] R. O. Onvural, "Asynchronous Transfer Mode Network: Performance Issues," Artech House, 1994.

[PAP84] A. Papoulis, "Probability, Randon Variables, and Stochastics processes", McGraw-Hill International Book Company, 1984

[PAR96] M. Parulekar and A. M. Makowski, "Tail probabilities for a multiplexer with self-similar traffic", Proc. of IEEE INFOCOM'96, pp. 1452-1459, March 1996.

[PER97] P. Tsingotjidis, "A feasibility study for a proactive congestion control for broadband networks," Ph.D. Thesis, Concordia University, August 1997.

[PRU95] P. Pruthi and A. Erramilli, "Heavy-tailed ON/OFF sources behavior and self-similar traffic," ICC 95.

[SAI91] H. Saito and K. Shiomoto, "Dynamic call admission control in ATM networks," IEEE Jornal on Selected Areas in Communications, vol. 9, pp. 982-989, Sep. 1991

[SEN89] P. Sen, B. Maglaris, N. Rikli, and D. Anastassiou, "Models for packet switching of variable-bit-rate video sources," IEEE Journal on Selected Areas in Communications, vol. 7, no 5, pp. 865-869, June 1989.

[SUB95] S. Subramanian, and T. Le-Ngoc, "Modeling of aggregate multimedia traffic," Technical report, Concordia University, No.4, 1995.

[SRI86] K.Sriram and W. Whitt, "Characterizing superposition arrival processes in packet multiplexers of voice and data," IEEE J. Sel Areas Commun., Sept. 1986.

[TSY97] B. Tsybakov, and Nicoals D.Georganas,"On self-similar traffic in ATM queues definitions, overflow probability bound, and cell delay distribution", IEEE/ACM Transactions on Networking Volume 5 , Issue 3 (1997).

[TUC88] R. C.F. Tucker, "Accurate method for analysis of a packet-speech multiplexer with limited delay", IEEE Trans. Comm., vol.36, no.4, April 88.

[VEI93] D. Veitch, "Novel models of broadband traffic", Proc. IEEE Globelcom'93, Houston, Tx, Dec. 1993.

[WI95] W. Lau, A. Erramilli, J. L. Wang, W. Willinger, "Self-similar traffic generation: The random midpoint displacement algorithm and its properties," Proceeding of ICC, pp. 466-472,Seattle, June 1995.

[WO90] G. M. Woodruff and R. Kositpaiboon, "Multimedia traffic management principles of guaranteed ATM network performance," IEEE Journal on Selected Areas in Communications, vol. 8, pp. 437-446, April 1990