

- Lib. Copy

# INVESTIGATION OF MULTIPLEXING BEHAVIOR OF MPEG ENCODED VIDEO DATA STREAMS

Dissertation submitted to  
**JAWAHARLAL NEHRU UNIVERSITY**  
in partial fulfillment of requirements  
for the award of the degree of  
**Master of Technology**  
in  
**Computer Science**

By

**G.V.N. PRASAD YADAV**



School of Computer & Systems Sciences  
**JAWAHARLAL NEHRU UNIVERSITY**  
NEW DELHI - 110 067

January 1999

1718

**INVESTIGATION OF MPEG MULTIPLEXING BEHAVIOR  
OF MPEG AND VIDEO DATA STREAMS**

**UNIVERSITY**

Dissertation submitted to  
**JAWAHARLAL NEHRU UNIVERSITY**  
in partial fulfillment of requirements  
for the award of the degree of  
**Master of Technology**  
in  
**Computer Science**

By

**G.V.N. PRASAD YADAV**

71P+fig.




**School of Computer & Systems Sciences  
JAWAHARLAL NEHRU UNIVERSITY  
NEW DELHI - 110 067**

**January 1999**


## CERTIFICATE

This is to certify that the dissertation entitled "*Investigation of multiplexing behavior of MPEG encoded video data streams*" which is being submitted by **Mr. G.V.N. PRASAD YADAV** to the School of Computer & System Sciences, Jawaharlal Nehru University, for the award of **Master of Technology in Computer Science**, is a record of bonafide work carried out by him.

This work is original and has not been submitted in part or full to any university or institution for the award of any degree.

  
Prof. P.C. Saxena 8/1/99

(Dean SC&SS)

  
Dr. R.C. Phoha

(Supervisor)

## DECLARATION

This is to certify that the dissertation entitled "*Investigation of multiplexing behavior of MPEG encoded video data streams*" which is being submitted to the School of Computer & System Sciences, Jawaharlal Nehru University, for the award of **Master of Technology in Computer Science**, is a record of bonafide work carried out by me.

This work is original and has not been submitted in part or full to any university or institution for the award of any degree.

**G.V.N. PRASAD YADAV**

## ACKNOWLEDGEMENTS

I am very glad to express my sincere gratitude to my supervisor Dr. R.C. Phoha, School of Computer & Systems Sciences, Jawaharlal Nehru University for his valuable guidance in completing this dissertation successfully. He provided me important literature and books related to this dissertation. He always inspired me. I feel it is a great privilege to have had the opportunity to work under his prestigious supervision. His constant encouragement and support helped me immensely.

I would like to express my sincere thanks to Prof. P. C. Saxena, Dean, School of Computer & Systems Sciences, Jawaharlal Nehru University for providing the necessary facilities in the centre for the successful completion of this Dissertation.

I take this opportunity to thank all the members of SC&SS and friends for their help and suggestions during the course of my project work.

**G.V.N. PRASAD YADAV**

..... *dedicated to*  
*my beloved Parents*

# Contents

	<b>Abstract</b>	<b>1</b>
<b>1.</b>	<b>Introduction</b>	<b>2</b>
	1.1 ATM Over View	2
	1.2 ATM Cell	4
	1.3 ATM Applications	6
	1.4 The Service Classes Defined by ATM	7
	1.5 Multiplexing	11
<b>2.</b>	<b>Video Coding</b>	<b>15</b>
	2.1 General Over View of Video Coding	17
	2.2 Basic Coding Control	20
	2.3 Real Time Video Services	22
	2.4 Coding Control Strategy for SBR	23
	2.5 Transmission Policies	32
	2.6 SBR-DBR Comparison	37
<b>3.</b>	<b>MPEG Encoding</b>	<b>42</b>
	3.1 MPEG Encoder	42
	3.2 The Distribution of a MPEG Source	45
	3.3 AR Modeling of MPEG	47
<b>4.</b>	<b>Modeling and Simulation Results</b>	<b>49</b>
	4.1 Modeling of VBR MPEG Video Source	49
	4.2 Modeling of Multiplexer	53
	4.3 Convergence of Simulation Results	53
	4.4 ATM Multiplexer Performance Evaluation	59
<b>5.</b>	<b>Conclusion</b>	<b>68</b>
	<b>References</b>	<b>70</b>

# Abstract

The ability of the network elements to apply statistical multiplexing while providing a guaranteed quality of service to the users is depending on the statistical properties of the ATM traffic streams. Thus it is essential to investigate the multiplexing behavior of different ATM traffic streams with respect to the ATM QoS parameters cell loss, cell delay and cell delay variation using realistic source models. Since compressed video will have a major share in future broadband traffic and the MPEG coding algorithm up to now has achieved wide popularity, we will concentrate on VBR MPEG video data streams.

The main aim of this dissertation is two fold. First the consequences of the usage of LRD traffic streams in discrete event simulations in terms of the necessary simulation duration and the convergence to a steady state behavior are investigated. Therefore, we will investigate how the convergence rate of the confidence interval size of the empirical mean value of discrete stochastic processes depends on the Hurst parameter.

Second, it will be shown that both, the LRD and periodic traffic components have a significant impact on the multiplexer performance in terms of the cell loss probability and the cell delay but mainly act on different time scales.

The actually noticeable influence on the Quality of service, that the VBR MPEG video data streams perceive, depends on the multiplexer buffer size and the stochastic properties of the video data streams at the different time scales.



# Chapter-1

## Introduction

---

### *1.1 Asynchronous Transfer Mode Overview*

Asynchronous Transfer Mode (ATM) is a form of data transmission that allows voice, video and data to be sent along the same network. In the past, voice, video and data were transferred using separate networks: voice traffic over the phone, video over cable networks and data over networks.

ATM is a cell-based, connection-oriented, switching and multiplexing technology designed to be a fast, general-purpose transfer mode for multiple services. ATM is a technology-defined by protocol standard created by the ITU-T, ANSI, ETSI and the ATM Forum. It is asynchronous because cells are not transferred periodically. Cells are given time slots on demand. ATM is also intended to be used across LANs and WANs to provide seamless connectivity ( i.e. one transfer method will be used for both LANs and WANs).

ATM is a layer in the Broadband Integrated Services Digital Network(B-ISDN). B-ISDN is a protocol model similar to the Open Systems Interconnection Reference Model (OSIRM) which is the seven layer protocol model used widely in today's networks. OSIRM defines data communications in a multiple architecture and vendor environment.

The three key layers of B-ISDN are the Physical Layer, the ATM Layer and the ATM Adaptation Layers ( AALs ).The Physical Layer is responsible for the electrical or optical transmission and receipt along the physical media between two devices. The

physical layer is responsible for supporting different physical media and different media interface rates. The physical layer maps the cells into time-division multiplexed frame which are then sent over the physical medium. The ATM layer handles multiplexing and switching and maintains the Quality of Service (QoS). Its functions are performed by processing the information in the cell header. Each time an ATM connection is made, a traffic contract is established between the network and the user. The contract establishes the characteristics of the user data. By determining the type of data, the network is able to allocate a certain amount of resources (bandwidth) to the user to ensure the traffic can be supported. The user communicates such details as the average and burst traffic rates and the acceptable loss and delay levels. This information is determined by the source of the traffic (i.e. database transfer requires a different bandwidth allocation than a video-conferencing call). These parameters combined are known as the QoS. The ATM layer is responsible for maintaining the QoS for each connection. The QoS is maintained by such functions as flow control, queuing priority and cell loss priority. The ATM Adaptation Layers adapt user traffic to a cell format. Data of various types, having different characteristics is chopped into 48-byte cells by the AALs. At the receiving end the cells are reassembled into the original form by the AAL. Different AALs exist for different types of traffic. The AALs are divided into the Convergence Sublayer (CS) and the Segmentation and Reassembly (SAR) sublayers. The CS is further divided into the Common Part (CPCS) and the Service Specific (SSCS). The function of the CS layer is to describe the method in which non-ATM traffic is converted into ATM traffic. The SAR inserts data into ATM cells and adds header information. There are four types of AALs to handle different traffic types. AAL 1 is used for constant bit-rate traffic and circuit

emulation. AAL 2 is not defined yet but will be used for connection-oriented, variable-bit rate traffic. AAL3/4 and AAL 5 are used for connection-oriented or connectionless traffic of variable bit-rate traffic.

## ***1.2 ATM Cell***

The ATM cell is the fixed length data unit used to transmit data. The data is encapsulated into a 48 byte payload and is preceded by a 5-byte header. In an ATM network, all data is switched and multiplexed in these cells. The header includes information about the contents of the payload and about the method of transmission. The sections in the header are a series of bits, which are recognized and processed by the ATM layer. Sections included in the header are Generic Flow Control (GFC), Cell Loss Priority (CLP), Payload Type, Header Error Control, the Virtual Path Identifier and the Virtual Channel Identifier. The GFC section provides information that is used in the multiplexing process. The GFC is intended to support flow control. The CLP bit indicates the loss priority of an individual cell. Cells are assigned a binary code indicate either high or low priority. A cell loss priority value of zero indicates that the cell contents are of high priority. High priority cells are least likely to be discarded during periods of congestion. Those cells with a high priority will only be discarded after all low priority cells have been discarded. Cell loss is more detrimental to data transmission than it is to voice or video transmission. Cell loss in data transmission results in corrupted files.

The Payload Type section describes the contents of the payload. The Payload Type section contains three bits, which indicate whether the payload contains user data or layer management information. User data is data of any traffic type that has been packaged into

an ATM cell. An example of management information is information involved in call set-up. This section also notes whether the cell experienced congestion. The Header Error Control field consists of error checking bits. The Header Error Control field is an 8-bit Cyclic Redundancy Code to check for single bit and some multi-bit errors. This field provides error checking only for the header field, not the payload. The Virtual Path Identifier and Virtual Channel Identifier provide information on the path that the cell will take in its transmission. The path is divided into channels. The choice of the 48-byte payload was made as a compromise to accommodate multiple forms of traffic. The two candidate payload sizes were initially 32 and 64 bytes. The size of the cell has an effect on both transmission efficiency and packetization delay. A long payload is more efficient than a small payload since, with a large payload, more data can be transmitted per cell with the same amount of overhead (header). For data transmission alone, a large payload is desirable. The longer the payload is, however, the more time is spent packaging. Certain traffic types are sensitive to time such as voice. If packaging time is too long, and the cells are not sent off quickly, the quality of the voice transmission will decrease. The 48 byte payload size was the result of a compromise that had to be reached between the 64 byte payload which would provide efficient data transfer but poor quality voice and the 32 byte payload which could transmit voice without echo but provided inefficient data transfer. The 48-byte payload size allows ATM to carry multiple forms of traffic. Both time-sensitive traffic (voice) and time-insensitive traffic can be carried with the best possible balance between efficiency and packetization delay. The cell header comes in two forms: the User-Network Interface (UNI) header and the network Node Interface (NNI) header. The UNI is described as the point where the user enters the network. The NNI is the

interface between networks. The difference between the two header types is that the UNI header has a Generic Flow Control section. The GFC area is not used for the NNI cell. The GFC section contains information used for multiplexing. Each cell is multiplexed based on this information. At the NNI, stage the information is already multiplexed and no additional user data will be added. The GFC area for the NNI cell is used for transmission path information (VPI/VCI).

### ***1.3 ATM Applications:***

The following is a sample of some applications made possible or optimized by ATM:

- working at home
- home shopping using voice, video and on-line databases
- video on demand for popular movies
- interactive multimedia games and applications
- distance learning
- on-line video libraries
- videoconferencing
- medical imaging
- remote database access

ATM is a connection-oriented transfer mode. Connection-oriented transfer requires that a connection between points be made prior to transfer. The path between communicating devices must also be established prior to transfer. Since there is an

established connection in this form of transmission, a confirmation from the receiver is not required. Frame relay is another example of connection-oriented transmission.

ATM is theoretically capable of supporting connectionless traffic, such as IP. The details of IP over ATM are being defined by the IETF. The IETF is developing protocols for encapsulation, multicasting, addressing, address resolution, neighbor discovery, use of published ATM Forum UNI signaling for IP call set up and connection negotiation, and network management, as appropriate, to allow the operation of inter network protocols over an ATM network

#### ***1.4 The Service classes defined by ATM***

The service classes defined by ATM forum traffic management group are CBR, VBR, ABR, UBR. Each class is defined as follows:

##### **1. CBR (constant bit rate)**

The CBR service class is intended for real-time applications, i.e. those requiring tightly constrained delay and delay variation, as would be appropriate for voice and video applications. The consistent availability of a fixed quantity of bandwidth is considered appropriate for CBR service. Cells, which are delayed beyond the value, specified by CTD (cell transfer delay) are assumed to be significantly less value to the application.

For CBR, the following ATM attributes are specified:

- PCR/CDVT(peak cell rate/cell delay variation tolerance)
- Cell Loss Rate

- CTD/CDV
- CLR may be unspecified for CLP=1.

## **2. Real time VBR**

The real time VBR service class is intended for real-time applications, i.e., those requiring tightly constrained delay and delay variation, as would be appropriate for voice and video applications. Sources are expected to transmit at a rate, which varies with time. Equivalently the source can be described "bursty". Cells, which are delayed beyond the value specified by CTD, are assumed to be of significantly less value to the application. Real-time VBR service may support statistical multiplexing of real-time sources, or may provide a consistently guaranteed QoS.

For real time VBR, the following ATM attributes are specified:

- PCR/CDVT
- CLR
- CTD/CDV
- SCR and BT(sustainable cell rate and burst tolerance)

## **3. Non-real time VBR**

The non-real time VBR service class is intended for non-real time applications, which have 'bursty', traffic characteristics and which can be characterized in terms of a GCRA. For those cells, which are transferred, it expects a bound on the cell transfer delay. Non-real time VBR service supports statistical multiplexing of connections.

For non-real time VBR, the following attributes are supported:

- PCR/CDVT
- CLR

- CTD
- SCR and BT

#### **4. UBR (unspecified bit rate)**

The UBR service class is intended for delay-tolerant or non-real-time applications, i.e., those which do not require tightly constrained delay and delay variation, such as traditional computer communications applications. Sources are expected to transmit non-continuous bursts of cells. UBR service supports a high degree of statistical multiplexing among sources. UBR service includes no notion of a per-VC allocated bandwidth resource. Transport of cells in UBR service is not necessarily guaranteed by mechanisms operating at the cell level. However it is expected that resources will be provisioned for UBR service in such a way as to make it usable for some set of applications. UBR service may be considered as interpretation of the common term "best effort service".

#### **5. ABR (available bit rate)**

Many applications have the ability to reduce their information transfer rate if the network requires them to do so. Likewise, they may wish to increase their information transfer rate if there is extra bandwidth available within the network. There may not be deterministic parameters because the users are willing to live with unreserved bandwidth. To support traffic from such sources in an ATM network will require facilities different from those for Peak Cell Rate of Sustainable Cell Rate traffic. The ABR service is designed to fill this need.



The Broadband Integrated Services Digital Network (B-ISDN) will be based on the Asynchronous Transfer Mode (ATM) that uses packets with a constant length of 53 bytes called cells. It allows statistical multiplexing of variable bit rate (VBR) sources to make efficient use of the network resources. The ability of the network elements to apply statistical multiplexing while providing a guaranteed quality of service (QoS) to the users is, besides the total load offered, decisively depending on the statistical properties of the ATM traffic streams. Thus it is essential to investigate the multiplexing behaviour of different ATM traffic streams with respect to the ATM QoS parameters cell loss, cell delay and cell delay variation using realistic source models.

Quality of service is an important issue for ATM networks, because they are used for real time traffic, such as audio and video. When a virtual circuit is established, both the transport layer and the ATM network layer must agree on a contract defining the service. This contract may have legal implications.

The contract between the customer and the network has three parts.

1. The traffic to be offered.
2. The service agreed upon.
3. The compliance agreed upon.

The contract may be different for each direction. For a video-on-demand application, the required bandwidth from the user's remote control to the video server might be 1200bps. In the other direction it might be 5mbps. The virtual circuit will not be setup if the customer and the carrier cannot agree on terms or the carrier is unable to provide the service desired. The first part of the contract is the traffic descriptor. It

characterizes the load to be offered. The second part of the contract specifies the quality of service desired by the customer and accepted by the carrier. Both the load and the service must be formulated in terms of measurable quantities, so compliance can be objectively determined. Merely saying "moderate load" or "good service" will not do.

The ATM standard defines a number of quality of service parameters in order to make it possible to have concrete traffic contracts. The customer and carrier can negotiate the values of these parameters.

For each Quality of service parameter, the worst case performance of each parameter is specified, and the carrier is required to meet or exceed it. In some cases, the parameter is minimum and in others it is a maximum. The Quality of service of parameter, here again, is specified separately for each direction.

### ***1.5 Multiplexing***

Multiplexing is the sharing of one physical transmission medium by more than one data stream. In ATM, cells containing different forms of data are multiplexed over the same bandwidth. Each time slot may contain cells from voice, video or data traffic types. The traffic type with the most throughputs required will take up most of the bandwidth. For example, a multiplexed multimedia application may have five out of eight cells for video, two for sound and one for data. Multiplexing improves efficiency by maximizing resources.

ATM uses statistical multiplex gain to improve efficiency. This process involves dynamically assigned time slots only to users who need them. Time slots are not reserved for individual users and they are not sent if no data needs to be transmitted.

Video data will have a major share in future broadband traffic due to the introduction of video on demand services, conferencing services and as a main constituent of multimedia applications. Since the beginning of the MPEG (Motion Pictures Experts Group) standardization efforts [MPEG-1, MPEG-2], the MPEG coding technique has achieved wide popularity. Therefore, we will investigate the multiplexing behaviour of MPEG encoded video data streams. Due to the different structure and complexity of the consecutive pictures of a video, the compression of the video data inherently results in a VBR video data stream if a constant picture quality shall be attained. To investigate the impact of VBR MPEG encoded video data streams on the network elements, stochastic models are used to characterize the behaviour of the video sources and the associated ATM traffic streams.

Generally, the behaviour of a video source and thus the stochastic characteristics of the associated ATM VBR video data stream depend on the coding technique, the application and the video contents itself. These influences act on different time scales. At least three levels can be distinguished, namely the cell level, the picture level and the scene level. The ATM packetization process dominates the behaviour at the cell level. The encoder algorithms and parameters essentially determine the picture level. Finally, the scene level is mainly governed by the fluctuations in the amount of information in

consecutive pictures. The amount of information depends on the contents and type of the video material and is independent of the coding algorithm.

Recent research results indicate that VBR video traffic exhibits a property called long range dependence (LRD) or persistence that can be characterized by the Hurst parameter  $H$ . Long-range dependent traffic streams are highly correlated and in the case of video data even pictures a long time span apart cannot be considered independent of each other. The LRD property mainly results from the different complexity of consecutive pictures. Therefore, the Hurst parameter  $H$  is depending on the movie contents and varies over a broad range of values, but it is almost independent of the video-coding algorithm. The offered traffic volume per picture is characterized by highly asymmetrical empirical probability distribution functions (PDFS) that can be approximated by log normal PDF's, VBR MPEG video data streams additionally reveal periodic components due to the application of three different picture types with different compression ratios. The that VBR MPEG video data streams exhibit LRD and periodic properties has a significant impact on the ATM multiplexer performance, but also affects its evaluation using discrete event simulations. Thus the focus of this dissertation is twofold to quantify these influences.

The ATM multiplexer performance results that will be presented in this dissertation based on discrete event simulations conducted using the hierarchical VBR MPEG source model presented. It uses three log normal distribution functions to model the size of the I-, P- and B-pictures at the picture level. These distribution functions are derived from a

fast fractional Gaussian noise (ffGn) random process via a transformation that takes MPEG GOP (group of pictures) pattern into account. The ffGn random process models the long-range dependent fluctuations of the picture information contents at the scene level. It is an approximation of the fractional Gaussian noise process suited very well for computer simulations.

The remaining of the dissertation is organized as follows: First, we will briefly describe the main characteristics of the MPEG encoding algorithm that are reflected in the source model and restate the VBR MPEG source model as far as necessary for the following sections. Then, the impact of the use of LRD VBR traffic streams on the convergence of discrete event simulation results is investigated. Next, the performance of an ATM multiplexer that is fed by a number of VBR MPEG video data streams is evaluated.

## Chapter – 2

# VIDEO CODING

---

ATM multiplexing will be the way to integrate services on a common network. Video services are expected to attract the common public to the broad band. However, these services have special requirements, probably the most stringent concerning transmission. The high data rates involved the need of synchronisation, and interactive services the real time constraints are the main characteristics, which make video services so demanding in terms of network resources. Video services include video telephony or video conference, that have strong real time constraints, and digital television, information retrieval or others that are non real time services or have light constraints. Video coders generate data at a rate, which varies constantly depending mainly on what is called image complexity. This means that traffic generated by the coder is essentially variable bit rate. When transmitting this data over ATM, a choice between several ATM transfer capabilities is available for establishing the connection.

The first option is deterministic bit rate, where a fixed bandwidth is guaranteed. This bandwidth can be set to the maximum bit rate generated by the coder, so data can be fed directly into the network. In this case, most of the available bandwidth would be misused since the maximum rate occurs rarely. In order to use network resources efficiently, constant bit rate should be produced. Storing the coding data into a smoothing buffer and therefore introducing delay to this. If good video

quality is expected then this buffer should be large enough to accommodate image transients, thus allowing for large transmission delays which could turn unacceptable for real time applications.

The use of SBR connection seems more promising. A mean bit rate is guaranteed, but the transmission of bursts is allowed. In our context, emission of bursts could be used to transmit image transients with low delay. A first problem to solve in this case is to find out good connection parameters for transmitting different video services while satisfying some quality standards. These parameters are the sustainable cell rate that determines the maximum mean data rate, the peak cell rate and the maximum burst size. Available bit rate or other reservation protocols are not considered in this work since they require a user to network dialogue, which is incompatible with real time constraints.

In an SBR connection, cells are tested for conformance at the usage parameter control by means of the generic cell rate algorithm. Letting the network discard non conforming cells would result in severe degradation of the reproduced image. On the other hand, reduction of data flow by an increase in quantisation error at the coder gives a much tolerable result. This is why only conforming traffic is considered as fed into the network. So only quantisation errors will be present at the decoding images.

The traffic emitted depends on the generation control algorithm and on the transmission policy. The former determines the data volume, while the latter decides how exactly this data is delivered.

## 2.1 GENERAL OVERVIEW OF VIDEO CODING

Video coding is based on the removal of spatio temporal redundancy from the original signal. The MPEG compression standards are widely accepted. For this work, a software version of a MPEG compatible coder was implemented. We now describe the basic ideas of MPEG and how the generated data rate can be controlled. The input signal to the coder is a sequence of frames in a certain format. A frame is considered as the basic unit for coding and presentation. Frames are logically divided into logically slices and macro blocks. A macro block is a small portion of the image whose size is generally 16/16 pixel, itself being formed by 8/8 pixel blocks. A slice is a horizontal strip of the frame, one macroblock wide.

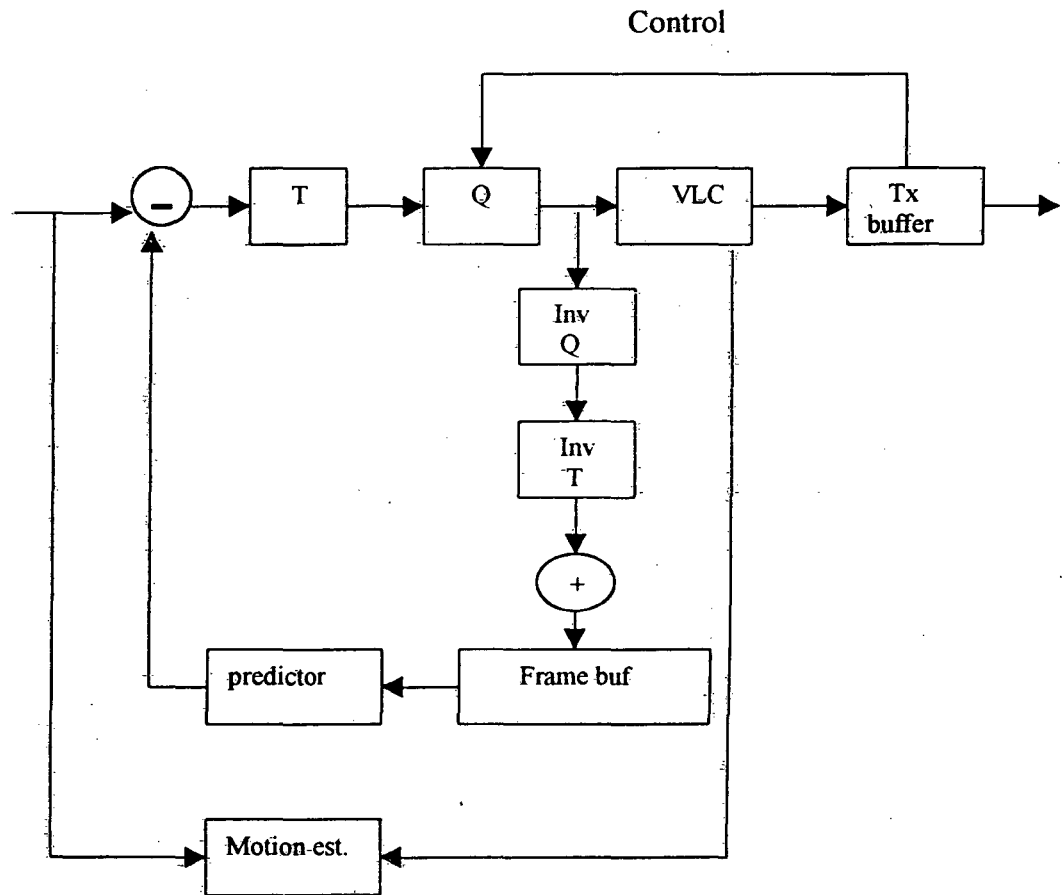


Fig.1 Diagram of typical video coder



Fig. 1 shows a simplified diagram of controlled MPEG coder. The inner loop is a predicative loop, where temporal redundancy is removed. The frame to be coded is compared to a predictive frame. This predictive frame made up from the last (decoded) frames where each macro block is affected by local motion estimation, a technique known as " motion compensation". The difference frame, which normally contains mostly null values then, goes through extraction of spatial redundancy. Each block goes through a discrete cosine transform in order to get the spectral coefficients. These coefficients are then quantised with granularity given by an external parameter  $qp$  (quantisation parameter). A low  $qp$  produces a large amount of data out of the coder, and a high quality decoded image. A high  $qp$  produces little data, so the coded image is noisy and the prediction loop takes several frames to stabilize. Data for coefficients, quantisation and motion compensation are then statistically compressed by using variable length coding.

Temporal prediction can be used in three ways in MPEG.

Intra (I) frames use no temporal prediction. I frames are normally located periodically in the data stream for editing purposes and for limiting transmission error propagation. These frames use only spatial compression, so they produce a large amount of data. In this case of real time services, I frames would require large transmission buffers for storage, which would result in an unacceptable delay. Small I frames could be generated by using a high  $qp$ , but this would result in a severe image degradation on the decoded image.

Predictive (P) frames are based on preceding P or I frames.

Bi-directional or interpolated (b) frames depend both on past and future i or p frames, thus resulting in a compression rate. The use of B frames requires a several frame delay for coding-decoding, which becomes unacceptable in the case of real time services. Therefore, when delay is a major restriction, only p frames are used. This strategy is part of the "low delay profile". The effect of transmission errors, if they occur, is extinguished by a gradual refresh mechanism. As all frames are depictive, there is no generation pattern, as in the group of pictures of the MPEG standard. Variations in the generation reflect changes in the video input, not in the coding mode. The refreshment methods that avoid error performance are assumed to produce only a slight overhead. The outer feedback loop shown in the diagram is needed if some restrictions on the generated data should be met. The data generated by each frame depends both on  $qp$  and on the image complexity. Frames difficult to predict are complex images. These occur during scene transients, the best example being a scene cut, or when the scene contains lots of local movements. In the case of scene cut, if  $qp$  is kept low, an acceptable quantisation noise level. If a whole portion of the sequence becomes complex, a new equilibrium important volume of data will be generated for this frame leading to a good decoded image. If  $qp$  is high, no such burst will be produced., but it will take several frames for the prediction loop to establish around an acceptable quantisation noise level. If a whole portion of the sequence becomes complex, a new equilibrium between data generated and image quality will be reached. Taking for example a CBR transmission and a real time service, a certain delay is assigned to the smoothing buffer. This imposes a maximum size for this buffer. If the coder generates data at a rate higher than the buffer's output

rate, the buffer level will grow. A *qp* controller must be devised for the buffer never to overflow nor empty. This controller must calculate the *qp* to be used for the next frame according to the current buffer level. In VBR operation, a generation control is still needed in order to honor the connection contract, even if restrictions are looser.

## **2.2 BASIC CODING CONTROL**

In general, the data rate produced by a coder can be adjusted by changing the frame rate, or altering the quantisation parameter (*qp*). The latter is the most usual technique. This section gives the *qp* control algorithm used for simulations. The controller must give a good trade off between quality consistency and generation stability, if the controller is too reactive to changes in the buffer level, *qp* will vary constantly thus lowering quality consistency, an effect considered annoying. If, on the other hand, *qp* is made too stable, then there is a high risk for the buffer to overflow or become empty due to unpredictable changes in image complexity. The variation of *qp* is inevitable in the case of a CBR transmission. For VBR, quality consistency may be improved as a burst can be emitted avoiding the filling of the transmission buffer.

The actual algorithm is based on the use of a fixed size buffer emptied at a constant rate (as is the case of CBR). It will be shown in section 4 that this algorithm ensures that the generated data can be transmitted over an SBR connection. The algorithm will observe a virtual buffer level instead of the actual transmission buffer level. All references to the 'transmission buffer' will be substituted with 'virtual buffer' for the case of SBR conforming transmission.

In the test models, the slice is used as the action time for the generation controller, choosing a  $qp$  for each slice. However, previous work has shown that a per frame control gives a more consistent quality. This has been adopted as the basic operation for our control, only altering  $qp$  at slice or macroblock level after the triggering of alarm mechanisms which avoid buffer overflow. In the following, these two operation forms are described.

### Normal Algorithm

1. Observe the buffer level and fix a desired frame generation volume. The transmission buffer (or virtual buffer in SBR) has a fixed target level (e.g.  $\frac{1}{4}$  of the buffer size). The deviation from this target level is used to calculate the amount of data that should be generated by each of the following  $nf$  frames in order to reach the target level. A low value for the parameter  $nf$  results in a very reactive control strategy, whereas a high  $nf$  results in a looser control over the buffer level and a higher quality consistency.
2. Choose a  $qp$  for coding the next frame. Given the desired generation for the next frame, an adequate  $qp$  must be found. Due to real-time constraints, this cannot be done iteratively, but before the actual coding. Generation as a function of  $qp$  is not known a priori, so it is estimated based on previous frames as follows: the data generation and the corresponding  $qp$  for the last two frames are stored ( $G1$ ,  $G2$ ,  $qp1$  and  $qp2$ ). If the desired generation lies between  $G1$  and  $G2$ , then a linear interpolation is used to find  $qp$ . Otherwise, a hyperbolic extrapolation is made,

TM-7660



using only the nearest point (this comes from the fact that the product  $qp$  generation is approximately constant).

### **Alarm Mechanisms**

A sudden increase in image complexity can cause the buffer to overflow. An alarm level is defined as a percentage of the buffer capacity (e.g. 95%). Two actions are taken:

1. At slice level, whenever the slope of the buffer level is high enough for the buffer to reach the alarm level in the current frame,  $qp$  is increased by 1 unit.
2. At macroblock level, if the buffer is full beyond the alarm level,  $qp$  is increased proportionally to the excess level.

The controlled system is stable. Alarm is triggered only in exceptional cases such as scene-cuts.

### ***2.3 REAL TIME VIDEO SERVICES***

Video services with real time constraints are those where a low end to end delay must be satisfied. Examples of this type of services are videotelephony or videoconference. These are iterative services where acquisition to presentation delays must be kept below 150ms in order not to confuse the users. With higher levels of delay, the fluidity of a conversation becomes affected.

When these services are carried over a data network, several factors add to total delay: acquisition, coding, transmission and reception buffering, propagation, switching and decoding. Therefore, a delay of 80 to 100 ms is assigned for buffering

purposes. This delay is possible since only predictive (P) frames are coded (refer to section 1 for a discussion). Switching delays should be negligible since cell rates throughout the network are much higher than the application's cell rate. Propagation delays are low in the case of nation wide communications.

The image quality needed depends on the specific service. In real time video services the mean quality requirement is not critical. Users are interested in the expression of the speaker and not in having a clean, high definition image. Moreover, screen size is small in the case of video telephony so coding rate facts are less apparent. Therefore, there is a quality threshold above that no further improvement is perceived.

The video signals quality signal quality has two components: mean quality and response to transients. For a fixed mean quality, the user expects the overall quality to be consistent.

The data rates for this kind of services is rather low, in the range 128 to 512 kbps. More complex signals (TV quality, for example) would require the default rate for MPEG1 (1.15 Mbps).

## ***2.4 CODING CONTROL STRATEGY FOR SBR.***

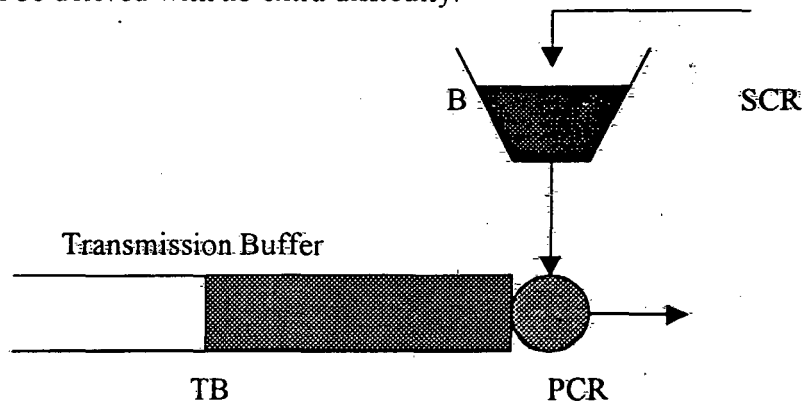
### ***WHAT PARAMETERS DOES VIDEO NEED?***

When establishing an SBR connection, three parameters are specified: sustainable cell rate (SCR), peak cell rate (PCR) and maximum burst size (MBS).

PCR is the maximum rate at which cell s can be emitted by the source. SCR is the maximum mean cell rate and MBS is the maximum size (in cells) a burst emitted at PCR can have.

Cell conformance is tested at the UPC by means of the generic cell rate algorithm (GCRA) with parameters  $1/SCR$  for the interval and  $B$  for the limit.  $B$  is deduced from the other parameter as follows:  $B=MBS (1-SCR/PCR)$ .

In order to obtain simpler and more intuitive results, a fluid model will be used throughout the dissertation for the traffic and the conformance algorithms. Exact results can be derived with no extra difficulty.



**Fig 2. SBR Transmission**

For the fluid model, conformance at the UPC is verified through a leaky bucket algorithm, where the bucket depth is  $B$  and the mean rate  $SCR$ , or equivalently, input traffic should be of  $(\sigma, \rho)$ -type with  $\sigma=B$  and  $\rho=SCR$ .  $B$  and  $SCR$  are normally expressed in cells and cells per second respectively,  $\sigma$  in bits and  $\rho$  in bits per second.

From the source's point of view, traffic can only be emitted if enough credits are available. The source can be thought of as having a credit pool which is filled at rate SCR, has a maximum of B credits and can be emptied at a maximum rate PCR. For each unit of data emitted, a credit unit has to be extracted from the credit pool. With this mechanism, the only restriction on the traffic produced by the source is that it is conforming at the UPC. Since the credit pool is very related to the bucket in the leaky bucket algorithm, the term 'bucket' will be used from now on to refer to the credit pool at the source.

Non conforming traffic should be avoided in video transmission. Indeed, if the network discards a single cell, the decoder will have to discard all data until a synchronisation mark is received. This is because data transmitted is compressed being impossible to resynchronise at any point in the bitstream.

This SBR traffic generation scheme can be applied as a server for the video coder's transmission buffer (recall that for CBR transmission a constant rate server has to be used). As a basic rule, the server will be assumed to be conservative: no credits will be lost if there is data waiting for transmission. In other words, letting the bucket overflow while there is data to transmit will waste no bandwidth.

Given an amount of data to be transmitted and a number of available credits, there is still a choice on how exactly will be the evolution of credit usage with time. This choice is now available on CBR. The way on how credits are used is decided according to a 'transmission policy'. For example, the server can wait for the bucket



to fill and then transmit at PCR; or it can directly use all credits by transmitting at PCR and then continue at SCR (the bucket's filling rate).

As the speed at which the transmission buffer empties is now constant –, as was the case for CBR – there is no direct relationship between buffer level and transmission delay. This implies a certain interaction between the transmission policy and the coder generation controller in order to satisfy the delay requirements. Therefore, a rather complex controller should be designed which takes into account not only the state of the transmission buffer but also of the traffic contract (the bucket level in this case) and the transmission policy. The problem of finding out an optimum controller becomes rather difficult.

However, it will be shown in the next paragraphs that the controller can be decoupled from the transmission policy. Also, the actual transmission buffer level and the credit bucket level can be put into a single variable called a 'virtual buffer level'. This virtual buffer level has a fixed maximum value, which ensures that the delay requirements are met. The controller described in the preceding sections can be used in this case, by observing the virtual buffer level instead of the actual transmission buffer level.

In order to prove the above assertions, the terminology used in the remaining of the section is as follows:

- PCR is the peak cell rate.  $br(PCR)$  means the bit rate corresponding to PCR.
- SCR is the sustainable cell rate.  $br(CBR)$  is the corresponding bit rate.

- TB is the transmission buffer level in bits, MTB is its maximum.
- B is the bucket depth in bits.
- C are the used credits, in bits. Therefore, B-C are the available credits. The bucket is referred to as 'full' when C=0.
- MBS is the maximum burst size.  $MBS [1-SCR/PCR]=B$  holds for the fluid approach.
- Delay is the portion of total delay assigned for buffering (e.g. 80ms). Other delays correspond to coding, propagation, etc. will not be considered in the following sections.

Consider, the data generated by the coder as being fed simultaneously into two transmission buffers: one buffer served by the SBR mechanism with its credit bucket, and another buffer served at a constant rate SCR (Fig. 3). Call this second buffer the 'virtual buffer', whose level is VB.

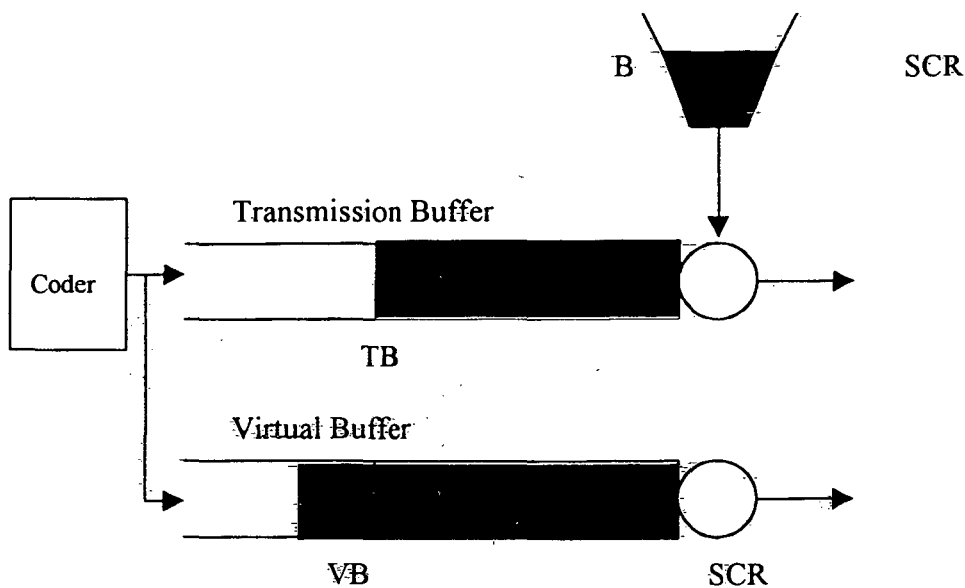


Fig.3 Real and virtual buffers

Since both servers are conservative, there is a simple relation between both buffer levels:  $VB=TB+C$ .

One way to see this is assuming that data arrives to both buffers at rate SCR. VB will be stable. TB will be stable if data is transmitted at SCR and therefore the bucket level and C are constant. If C increases by 1, it means that 1 bit from the buffer has been transmitted in addition to the SCR traffic, so TB will be reduced by 1. Notice also that the virtual buffer empties only when the bucket gets full.

The following lemmas will give more meaning to the variable VB.

***Lemma 1: Maximum transmission buffer***

Transmission buffering delay is bounded by Delay. Given the actual number of credits used (C), the maximum amount of data that can be transmitted in Delay is the number of credits available at that moment plus the number of credits which will arrive during this time. This gives the maximum level the transmission buffer can have at any moment:  $\max(TB)=MTB=B-C+Delay \cdot SCR$

Notice that both MTB and C are functions of time.

Therefore, if TB never reaches MTB, transmission is possible in a time less than Delay ♦

Assuming that the coder puts data into the buffer at a speed much higher than actual transmission speeds, the controller may safely allow the coder to generate up to (MTB-TB) bits. This is a rather conservative approach because if the coder is known

to produce data at a maximum rate, data in the buffer can be actually transmitted while the coder is producing new data. In this case, this bound on generation could be made tighter (i.e. generate more than simply  $MTB-TB$ ).

***Lemma 2: The virtual buffer (VB) and its maximum (MVB)***

The variable VB introduced above has the intuitive interpretation of being the level of a CBR smoothing buffer, and is actually easy to keep track of without even knowing the bucket level: VB increases with coder data generation and decreases at a constant rate  $br$  (SCR). The coding system easily keeps track of time since the time between frames or slices is constant. So recalculating VB at each slice interval is a trivial task.

The usefulness of VB is now presented below.

From Lemma 1, it is seen that  $TB+C \leq \text{Delay} \cdot br + B$  for the delay bounds to be met. Therefore, introducing the constant 'maximum virtual buffer'  $MVB = \text{Delay} \cdot br + B$ , the delay bounds are summarised in the simple expression:  $VB \leq MVB$  ♦

So, if the controller described in section 2 is given VB as a variable to control rather than TB, with a maximum of MVB rather than MTB, data can be delivered on time, independently of the transmission policy used.

When  $VB=0$ , the bucket is full of credits and there is no data to transmit.  $VB=MVB$  indicated for example that MVB bits are present in the buffer and all credits are available; or that no credits are available and there are  $\text{Delay} \cdot br$  bits to transmit.

Recalling that for a CBR transmission at rate SCR, the buffer was limited to  $\text{Delay} \cdot \text{br}(\text{SCR})$  bits, the ability to send bursts is equivalent to an increase of B bits in the buffer size, even when the mean data rate ( and mean quality) remains the same. A larger buffer allows for a better quality consistency: scene transients can be coded at a lower qp without risk of buffer overflow.

With respect to the connection parameters,  $\text{MVB} = \text{Delay} \cdot \text{br}(\text{SCR}) + \text{MBS} (1 - \text{SCR}/\text{PCR})$ . So virtual buffering capacity decreases for lower PCR values.

***Restriction 1: The Delay imposes a PCR – MBS relationship***

The value for PCR is not arbitrary, even if this parameter does not appear explicitly in Lemma 1. An underlying assumption is that TB can effectively be transmitted in a time less than Delay, given the finite value of PCR. Therefore, the condition  $\text{TB}/\text{br}(\text{PCR}) < \text{Delay}$  should be valid for any possible level of TB given in lemma 1:  $\text{Delay} \cdot \text{br}(\text{SCR}) + \text{B} - \text{C} \leq \text{Delay} \cdot \text{br}(\text{PCR})$ . This restriction depends on the used credits C. Its worst case gives  $\text{B} \leq \text{Delay} [\text{br}(\text{PCR}) - \text{br}(\text{SCR})]$ , or equivalently,  $\text{MBS} \leq \text{Delay} \cdot \text{br}(\text{PCR})$ .

This means that given a value for PCR, the useful burst size (or bucket depth) is limited. A bigger bucket is useless because extra credits will be never evacuated in time. The useful region lies below the straight lines shown in Figure 1, which are drawn for two different delay values (80 and 120 ms). Higher delay values correspond to higher slopes.

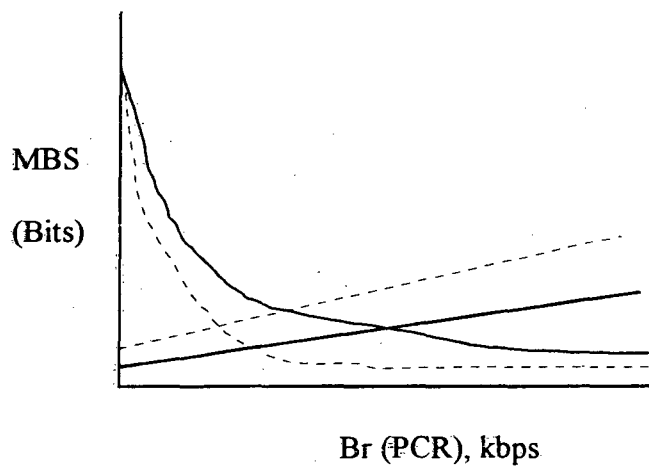


Fig. 4 PCR – MBS relationship.

***Restriction 2: A fixed video quality imposes a PCR-MBS relationship***

The response to transients is better if the controller allows a big virtual buffer excursion. This is tightly related to MVB, which is maximum of this buffer. Restriction 2 shows the relation between MBS and PCR that correspond to a given video quality.

From Lemma w,  $MVB = \text{Delay} \cdot br (SCR) + MBS (1 - SCR/PCR)$ .

Give MVB and SCR, different (MBS, PCR) couples will give the same transient quality, provided that  $MBS (1 - SCR/PCR) = \text{constant} = MVB - \text{Delay} \cdot br (SCR)$ .

The hyperbolas in Fig. 4 are constant quality curves. The minimum PCR and maximum MBS are obtained at the intersection of the line and the hyperbola.

corresponding to a certain Delay value. In this case,  $MBS_{max} = MVB$  (independent of Delay), and  $br(PCR)_{min} = MVB/Delay$ .

Only the low right part of each hyperbola is of interest, where the MBS-PCR balance is meaningful. Note that an increase in PCR provides only a very slight decrease in the maximum burst size.

## ***25 TRANSMISSION POLICIES: HANDLING OF BUFFER AND BUCKET***

The main problem in video transmission is having the data stored on time at the reception buffer for when it is required by decoder. An SBR type transmission can be thought of as a CBR transmission with the ability of sending data in advance to the reception buffer. The amount of data sent in advance is exactly the value of C used previously. Therefore, during periods of low generation rate, credits can be accumulated thus lowering C, so when a high generation rate transient appears, up to B bits can be sent 'in advance' to the receiver letting the remaining data for transmission at the normal rate (SCR). This gives a rather intuitive idea of how SBR can be more performing in terms of quality consistency when low delays are required.

A transmission policy can be defined as a set of rules for delivering data into the network. This includes the management of the available credits and of the transmission buffer.

It has been shown in previous sections that the coding process could work independently of the transmission policy used. However, this policy cannot be chosen arbitrarily. For example, a policy could be to transmit at SCR, never using extra credits for emitting bursts. This CBR type policy is conservative, but certainly some data will not be delivered in time (this occurs to all data which find the transmission buffer level above  $\text{Delay}_{br}$  (SCR)). Therefore, a well-behaved policy should emit bursts according to the transmission buffer level to ensure in time delivery of data into the network.

Two examples of possible transmission policies will be presented below, each producing traffic with different burstiness characteristics.

In the remaining of the section, the following terminology will be used:

- $N$  is the total delay measured in video units (for instance, slices or frames).
- When coding the  $n$ th unit, the  $(n-N)$ th unit is to be presented at the receiver.
- $RB$  and  $TB$  are the transmission and reception buffer levels when  $n$  has just been coded.
- $C$  are the used credits from the bucket, and  $B$  the bucket depth ( $0 < C < B$ ).
- $B_{pu}$  (SCR) and  $b_{pu}$  (PCR) are bits per video unit at mean and peak rates.
- $g(n)$  are the bits generated by  $n$ th unit.



Note that for any transmission scheme, the number of video units stored at the transmission and reception buffers add to a constant value (N). Video units are assumed to be stored and extracted from the buffers instantaneously by the coder and decoder.

### **Transmission policy 1 – Save as many credits as possible**

With this policy, the steady state situation corresponds to a full credit bucket ( $C=0$ ) and the transmission buffer at the target level. So transmission is normally done at SCR until TB reaches the level  $\text{Delay}_{br}(\text{SCR})$ . Whenever TB goes beyond this level, transmission at a higher rate should be triggered. The precise implementation of this policy is quite complex regarding when to resume transmission at SCR after a burst is sent.

When the coder's generation rate is lower than SCR, first the bucket is allowed to fill, then data are transmitted in advance in order not to waste credits.

Assume that the  $(n-N-1)$ th unit has just been decoded. Therefore the reception buffer, if  $RB > 0$ , contains some part of the  $(n-N)$ th unit. The remaining of the  $g(n-N)$  bits should be present in the reception buffer in the next time unit. Therefore,  $g(n-N) - RB$  bits have to be transmitted by this time (if this value is positive). As the policy is to save as many credits as possible, the volume to be transmitted is  $\max [g(n-N) - RB, \text{bpu}(\text{SCR}) - C, 0]$ .

The second term corresponds to transmission being conservative: credits that cannot be used to further fill the bucket are used to transmit data in advance. This policy will produce rare but considerable bursts. In fact, as the bucket is normally full, the potential burst size is the maximum allowed by the leaky bucket size.

#### **Transmission policy 2 – Transmit as early as possible**

Now the target situation is to have a certain number of used credits and the transmission buffer empty. The reception buffer will contain in average at least the controller's VB target level.

This policy is very simple to implement: each time there are credits available, data is transmitted at PCR. It will produce frequent short bursts (tens of cells) much like the coder's output. Some credits are normally used, so even the largest bursts will be smaller than MBS.

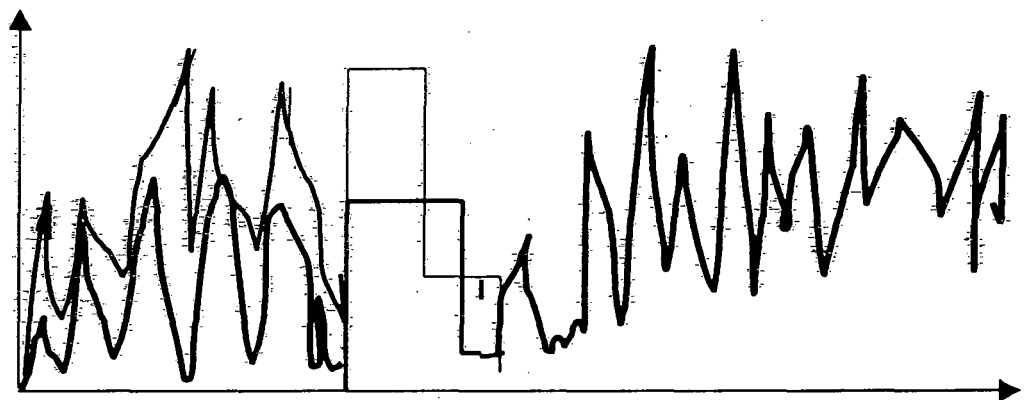
These small bursts should be multiplexed more effectively at the network level with others of the same type. So this policy is easier to implement, and not aggressive to the network (bursts are smaller).

The transmitted amount during one time unit is  $\min \{ \text{bpu (PCR)}, \text{TB (n)} - g(n), \text{bpu (SCR)} + \text{MBS} - C \}$ .

This means that the transmitter is delivering as much as it can, limited by the most restrictive of: PCR, the total amount of information present in transmission buffer, or the credits present in the bucket.

Recall that in all cases the transmitted signal is the same and the real time requirements are met.

Fig. 5 shows the traffic in cells per second when the second policy is used corresponding to mean rates  $br$  (SCR) of 512 kbps and 320 kbps,  $PCR=3 SCR$ , and  $B=167$  cells (64,000 bits). The test sequence has a scene cut in frame 79. After the cut the sequence images are more complex at the bottom. Indeed, the generation pattern during each frame exhibits a peak around the bottom slices of each frame. During the steady state, as transmission is immediate, the transmission rate coincides with the generation rate. Note that in this phase bursts are very short. In fact, they correspond to a video slice time, which is  $40 \text{ ms}/18=2 \text{ ms}$  and contain about 7 cells per slice when  $br$  (SCR)=512 kbps, or 4 cells when  $br$  (SCR)=320 kbps. It is expected that a set of sources of this kind could be multiplexed using moderate buffers in the network.



**Fig. 5 Traffic including a scene cut.**

During the scene cut the cell rate rises to  $PCR$  until the credits are exhausted. At this moment transmission continues rate  $SCR$  until the large units are delivered. Note that the  $SCR$  phase is longer in the 512 kbps case. This is because the common  $B$  value (167 cells) is relatively much higher when  $br$  (SCR)=320 kbps, therefore resulting almost sufficient for emitting all of the transient information units.

## **2.6 SBR – DBR COMPARISON**

Simulations were performed on 25 frame per second sequences using mean rates in the range 320 to 512 kbps for the video signal, speeds which are adequate for video conference or high quality video telephony. The delay assigned for transmission buffering was set to 80 ms, equivalent to two frames, leaving sufficient margin for other delays (coding, propagation, switching, image scanning) which are not considered here.

The DBR option is compared with SBR regarding several aspects of the global service quality.

**Same signal, same mean rate  $\Rightarrow$  Delay?**

For the same video quality and mean rate, if the DBR capability is used, transmission should be done at CBR (with  $PCR=br$  (SCR)). Here the maximum delay is given by the size of the smoothing buffer, which has been shown to be MVB (section 4). Numerical values are given in Table 1. The CBR figures are too high for a real time service, since buffering delay is only part of the global delay.

**Table 1 Same signal, same mean rate**

<b>Mean rate (kbps)</b>	<b>Delay SBR (ms)</b>	<b>Delay CBR (ms)</b>
320	80	213
512	80	163

**Same signal, same delay  $\Rightarrow$  Negotiated peak rate?**

If the same coded signal, and therefore the same mean data rate, and the same buffering delays are used, a DBR connection can be used but with a higher rate. In this case, since this rate is rarely used, statistical multiplexing is compromised. Using the data obtained in the working examples, the minimum DBR to be negotiated is about 2 times the mean rate for 512 kbps, and 2.66 times for 320 kbps, giving a very low utilisation factor. The rates to be negotiated in SBR and DBR in this case are shown in Table 2. A complete comparison should consider the number of DBR and SBR channels that may be multiplexed.

**Table 2. Same signal, same delay**

<b>Mean rate (kbps)</b>	<b>Negotiated mean SBR (kbps)</b>	<b>Negotiated rate, DBR (kbps)</b>
320	320	853
512	512	1043

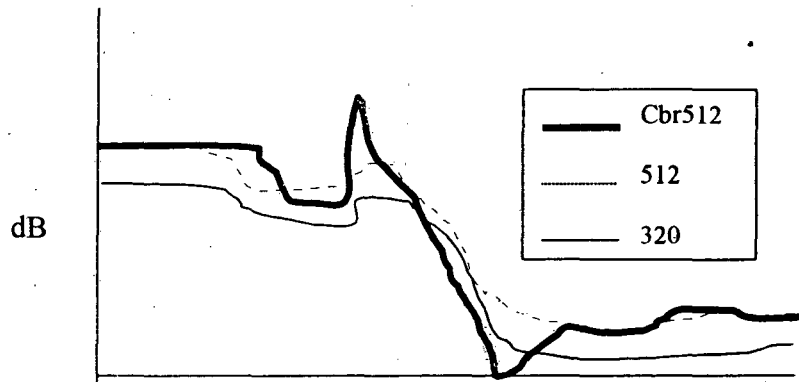
**Same mean rate, same delay  $\Rightarrow$  Signal quality?**

Finally, when mean rate and delay are maintained, the signal quality is compared giving error images and luminance signal to noise ration (LSNR) during the sequence. In this case, for the DBR channel the signal is coded at mean rate SCR, using a short buffer to achieve the delay.

Global image quality is affected. Fig. 6 shows the LSNR for DBR and SBR. A scene cut occurs in frame 79. In steady state, both DBR and SBR at 512 kbps give similar results. The error images obtained for DBR and SBR transmission strategies. Error images were obtained by subtracting original and decoded images, taking absolute values, and magnifying by a factor 8. The parameters used were Delay=80 ms (i.e 2 frames), br (SCR)=512 kbps, PCR =3SCR and MBS=17 cells. Other parameters are summarised in Table 3. The expected burst is calculated as  $MBS \cdot (1 - SCR/PCR) \cdot \text{target}$ , because transmission policy 2 was used, so target credits are normally used. Note that the traffic is not very bursty. In fact, the expected bursts are not large.

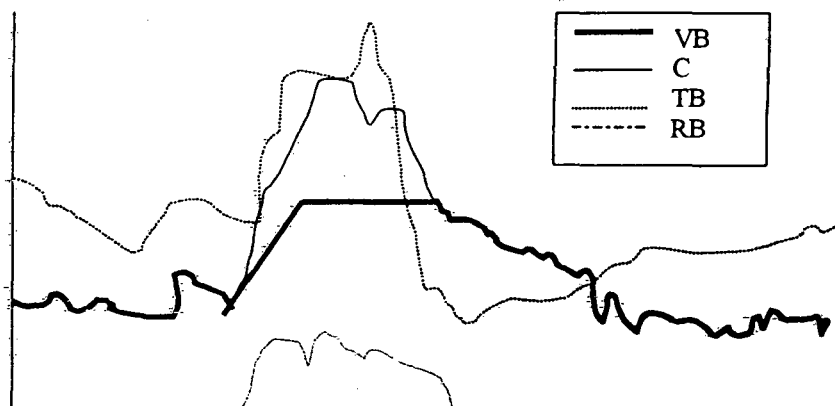
**Table 3 Same delay and mean rate. Simulation values overview**

Br(SCR) (kbps)	PCR (kbps)	MVB (bits)	control target (Bits)	expected burst (cells)
320	3 SCR	68 267	13 653	76
512	3 SCR	83 626	16 725	67
512 (CBR)	512	40 960	8196	-



**Fig. 6 LSNR during a scene cut**

A simulation using  $br(SCR)=320$  kbps is also included. In these conditions, a CBR connection is unable to respect the imposed delay, even with the coarsest quantisation level, whereas the SBR image is fairly good. Naturally, the steady state LSNR is lower when the mean rate is  $br(SCR)=320$  kbps. Subjective evaluation of the coded sequences show that the transient of the CBR signal (at 512 kbps) is quite noticeable, whereas the SBR signal, even at 320 kbps, has consistent quality and the distortion is not annoying.



**Fig. 7 Transmission (TB) and Reception (RB) buffers, Virtual buffer (VB) and used credits (C) evolution.**

The behaviour of buffers and bucket in the SBR scenario, corresponding to the given SCR, PCR and MBS and to the transmission policy 2 is shown in Fig. 7. In steady state the reception buffer contains an amount of data corresponding to the delay and some credits are normally used. The transmission buffer is usually empty.

During the scene cut credit usage grows at PCR, and stays at B during some frames. This is where the transmission buffer (TB) accumulates data. TB first grows because generation is faster than PCR. Then it grows as no more credits are left and transmission is limited to SCR. The burst transmitted during the scene cut causes an increase in the reception buffer level. Then this level goes below the steady state level due to the decoding of the huge frame 79. Anyway, this buffer is always far from starvation.



## Chapter-3

# MPEG Encoding

---

### *3.1 MPEG Encoder*

Since an uncompressed digital video stream needs hundreds of Mbits/sec to be transmitted, video compression has been studied a lot and several coding algorithms have been proposed to reduce the required bandwidth. For instance, ITU adopted H.261 for video telephone and video teleconference. Another attempt resulted in standardization of MPEG-I algorithm for storage of moving pictures. However, due to the flexibility of its algorithm, it can be used for various applications such as in multimedia workstations, video communication and so on.

The MPEG coding algorithm uses three different types of frames; Intraframe (I), Predictive (P) and Interpolative (B) frames. Initially, an I frame contains a two dimensional  $8 \times 8$  discrete cosine transform (DCT) of the original image. Secondly the coefficients of DCT are fed to a quantizer and a  $8 \times 8$  matrix obtains the quantization step size for of each DCT coefficient. Some psycho physics experiments have defined the values of this matrix. (ISO/IEC, 1994) Low frequency DCT coefficients are quantised with more accuracy than the high frequency ones. As a result a significant number of quantized coefficients will have zero value. Then the algorithm codes this block ( $8 \times 8$ ) of coefficients using run length and modified Huffman coding. The MPEG standard the

coding uses a fixed work following by the length of zero and the non-zero value of DCT coefficient (ISO/IEC, 1994).

In P frames macro blocks are coded with or without motion estimation. The algorithm searches a square area around a each macro block of the previous I or P frame in order to find a motion vector that minimizes the absolute difference between the current macro block and the chosen one in the previous frame. If the absolute difference is less than a threshold the motion vector is coded and transmitted. Then a DCT transform is applied to the prediction error of each macro block. The coefficients of transformation are quantized with constant step size instead of intra frame coding. The quantized coefficients are coded as an intra frame mode. If the absolute difference is greater than the above threshold the motion estimation cannot be used and the coding of the current macro-block is the same as the intra frame. B frames coding are similar to the procedure that has been previously described for (P ones). The only difference is that the motion vector can be estimated with respect to the previous I, P frames or the following I,P frames or an interpolation between them.

Since the MPEG coding algorithm has a fixed quantizer matrix (I,P,B). It can scale the quantization level using a parameter  $q$ . When MPEG video tries to maintain the output rate constant (CBR Mode) it is necessary to vary the  $q$  parameter dynamically, therefore the quality of coded moving pictures. If  $q$  remains constant the quality does not change but the output rate becomes to depend on the video activity. In a MPEG coding

there is a constant number  $N$  of frames where I, P and B are repeated periodically. The most typical cyclic frame pattern is IBBPBBPBBPBBBI.

The pattern period consists of 1 I, 3 P and 8 B frames. We call the mean value of this cyclic frame pattern Group of Picture (GoP). That means that the GoP value represents the average size of I, P and B frames within one pattern.

In intra frame mode the size of I frames has greater value than the size of P and B frames due to lack of motion estimation. However the volume of P and in particular B frames is not always very small. As we have stated above it is possible for the coding of P and B frames not to use the motion compensation for all macro blocks. This occurs when the prediction error is greater than a defined threshold. Therefore it is anticipated that a high activity causes real sizes of P and B frames.

We observe that the fluctuation of P and B frames is much greater than the I. B frames have the highest ratio of maximum to minimum frame size. Another important characteristic is the ratio  $\text{mean}(I)/\text{mean}(\text{Total})$  because it affects the aggregate traffic behavior. Since I frames have a mean value 3 or 4 times greater than the average total frames' size (including I P B). It is very difficult to achieve without multiplexing gain, low loss probabilities with small buffer size and for utilization close to 0.75 or above. Although the small size of P and B frames it seems that they play a major role to loss probabilities due to their large fluctuations (see sec. Traffic behaviour). Therefore it is not correct to ignore them so as to study the characteristic of MPEG video.

Moreover due to the motion estimation there is significant correlation between the I, P and B frames (Heyman , 1994). Let  $B_i$  ( $i=1$  to 8) and  $P_i$  ( $i=1$  to 3) be the I B and P frames respectively in a cyclic frame pattern. According to the MPEG algorithm  $B_1$  and  $B_2$  can have different values only if a video scene change occurs between them. In addition  $P_1$  is related to I frame and  $P_2$  to  $P_1$  and so on. The same relation has been noticed among  $B_1$  and  $P_1$ ,  $B_2$  and  $P_1$  or  $B_1$ ,  $B_2$  and I frames. These relations are a consequence of the motion compensation that the MPEG algorithm performs. In particular it is observed that the strongest dependency is between B frames while the most weak between P,I and B, I ones. It is therefore difficult to find appropriate models that can fit accurately with a real video MPEG source.

### ***3.2 THE DISTRIBUTION OF AN MPEG SOURCE***

Modeling of an MPEG sequence is a very useful work for designing a telecommunication network especially due to the great evolution of multimedia services along with the advances in VLSI technology that made possible the realization of the real-time MPEG systems. It can be used to determine the loss probabilities or the required network resources (Bandwidth, Buffers) Without having to witness the real video data we examine statistical models which characterize the properties of an MPEG video. Our analysis is based on a long real video sequence (approximately 27 min) With scene changes and camera zooming that are considered as the most difficult sequences for the MPEG to handle. As we mentioned above any MPEG video sequence consists of

three types of frames intra frame(I), predictive frame (P) and interpolative (B). The log normal and the gamma distributions fit well with the experimental data.

### ***Gamma Distribution***

The probability density function for gamma distribution is given by

$$f_x(x) = \frac{a^p}{\Gamma(p)} x^{(p-1)} e^{-x} \dots\dots\dots(1)$$

where  $\Gamma(p)$  is the known function defined as  $\Gamma(p) = \int_0^{\infty} x^{(p-1)} e^{-x} dx$

It can be proved easily that the mean and the variance of gamma distribution are related to the parameters a, p as following

$$E_x(x) = \frac{p}{a} \quad \text{and} \quad V_x(x) = \frac{p}{a^2} \dots\dots\dots(2)$$

In order to estimate the parameters of gamma df which fit well with the real data we use the method of moments. If MEAN and VAR are the mean and the variance of an MPEG video sequence then the estimated parameters are calculated as

$$\hat{a} = \frac{MEAN}{VAR} \quad \text{and} \quad \hat{p} = \frac{MEAN^2}{VAR} \dots\dots\dots(3)$$

### ***Lognormal distribution***

The pdf of log normal distribution is

$$f_L(x) = \frac{1}{\sqrt{2\pi\sigma^2}} x^{-1} \exp\left\{-\frac{1}{2\sigma^2}(\ln(x) - \mu)^2\right\} \dots\dots\dots(4)$$

where its mean and variance are given by  $E_{LN}(x) = e^{\mu + \frac{\sigma^2}{2}}$  and  $V_{LN}(x) = e^{2\mu + 2\sigma^2}(e^{\sigma^2} - 1)$ .

The estimated parameters which are calculated by the method of moments are given by the following formulas.

$$\sigma^2 = \ln\left\{\frac{VAR}{\exp(2\ln(MEAN))} + 1\right\} \text{ and } \hat{\mu} = \ln(MEAN) - \frac{\sigma^2}{2} \dots\dots(5)$$

Based on the equations (3) and (5) we can find the appropriate estimated parameters of gamma and log normal pdf's which fit well with real video data.

### ***3.3 AR Modeling of MPEG***

The models that we have described in the previous section cannot approximate the traffic behavior of a VBR MPEG video source. This occurs because each sample that is generated according to the probability density functions is independent of others. Therefore no correlation among them exists. However they are used for some other models (see DAR Models which require to know the PDF of the source. In this section we describe linear models based on the correlation function (or auto covariance) of an MPEG sequence.

**Analysis:**

Let  $x(n)$  be a stochastic process with  $n=0,1,\dots,N$ . A  $k$  order autoregressive model for  $x(n)$  is defined as

$$x(n) = -\sum_{i=1}^k a_i x(n-i) + b e(n) \quad \dots\dots\dots (6)$$

Where  $e(n)$  is an independent and Identically Distributed (i.i.d) variable with mean  $m_e$ , variance 1, and  $a_i, b$  contents. In order to minimize the square value of error, that is the  $\min\{E(e(n)^2)\}$  we conclude to the Yate-Walker equations

$$R \cdot a = -r_k \quad \dots\dots\dots (7)$$

Where  $R$  is a toeplitz matrix with elements the autocovariance values  $x(n)$  which is denoted as  $r(j) = E\{x(n)-m_x)(x(n-j)-m_x)\}$  where  $a$  the vector of the unknown parameters  $a_i$  and  $r_k = [r(1), r(2), \dots, r(k)]^T$ .

The mean value of error  $m_e$  and the unknown parameter  $b$  are given by

$$m_e = \frac{-m_x + \sum_{i=1}^k a_i m_x}{b} \text{ and } b^2 = r(0) + a r_k \quad \dots\dots\dots (8)$$

The autocovariance function of an AR model for  $l > k$  is

$$r(l) = -\sum_{i=1}^k a_i r(i) \quad \dots\dots\dots (9)$$

*In case of  $k = 1$  it can be shown that  $r(n) = \frac{b}{(1 - a_1^2)} a_1^n$  meaning auto covariance*

*has exponential behaviour.*

If  $k > 1$  we can find that the autocovariance consists of the sum of exponential functions.

## Modeling and simulation Results

---

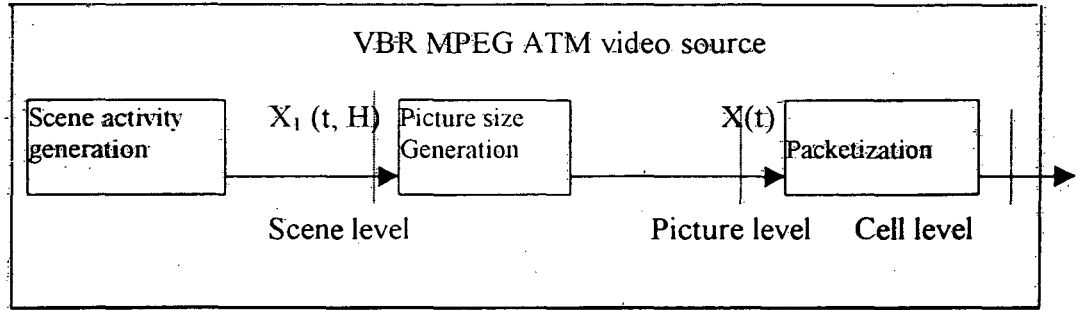
### *4.1 Modeling of VBR MPEG Video source*

The picture sizes of the VBR MPEG video data streams are modeled by the hierarchical source model presented in (see Figure) that is based on the GOP-periodic exponential transformation of the fast fractional Gaussian noise random process (ffGN). This model takes the first and second order statistical properties and the short and long-term correlation characteristics of the VBR MPEG video data streams into account, including long-range dependence (LRD) effects. The model can be adapted to represent any MPEG encoded video sequence by seven parameters that can be derived e. g. from an empirical sequence. In the following, the basic equations necessary to implement the ffGn process for the scene activity generation and the transformation to get the picture sizes will be presented. The ATM packetization process used for the simulations will be described shortly.

At the scene-level, an ffGn process is used to approximate the LRD behavior of video sources at that time-scale. The ffGn process  $X_f(t, H)$ ,  $t \in \mathbb{N}$ ,  $H \in [0.5, 1]$



generates a sequence of  $N(0,1)$ -distributed random variables with an autocorrelation function that approximates the autocorrelation function



**Figure 2. VBR MPEG ATM-video source model**

$$r\Delta B_H(\tau)\Delta_H(\tau) = \begin{cases} 1 & \text{for } \tau = 0 \\ \frac{1}{2} [(\tau+1)^{2H} - 2\tau^{2H} + (\tau-1)^{2H}] & \text{for } \tau > 0 \end{cases} \quad (1)$$

of these discrete fractional Gaussian noise random process  $\Delta B_H$ , that is exactly second order self-similar. The ffGn process is constructed as the sum of a low frequency term

$$X_l(t, H) = \sum_{k=1}^{N(n_{ffGn})} W_k X_{MG}(t, r_k) \quad (2)$$

and a high frequency term  $X_h(t, H)$ . The Markov Gauss processes  $X_{MG}(t, r_k)$  are defined as

$$X_{MG}(t, r_k) = \begin{cases} G_k(t) & \text{for } t = 1 \\ r_k X_{MG}(t-1, r_k) + \sqrt{1-r_k^2} G_k(t) & \text{for } t > 1 \end{cases} \quad (3)$$

Where  $G_k(t)$ ,  $k = 1, \dots, N(n_{ffGn})$ ,  $t \in \mathbb{N}$  denote sequences of independent standard normally distributed random variables. The log-1 covariance  $r_k$  is defined as

$$r_k = e^{-B^{-1}} \quad (4)$$

and the weight factors  $W_k$  are determined by

$$W_k^2 = \frac{H(2H-1)(B^{1-H} - B^{H-1})}{\Gamma(3-2H)} B^{-2k(1-H)} \quad (5)$$

The number  $N_{(\text{ffGn})}$  of Markov-Gauss processes depends on the number  $n_{\text{ffGn}}$  of consecutive random numbers that shall exhibit the LRD property. It is defined by

$$N(n_{\text{ffGn}}) = \lceil \ln(Q \cdot n_{\text{ffGn}}) / \ln(B) \rceil \quad (6)$$

The ffGn process is, besides the time  $t$  and the Hurst parameter  $H$ , depending on two additional parameters, then base  $B$  and the quality  $Q$ . These parameters define the accuracy of the approximation of the dfGn random process and its autocorrelation function that is achieved by the ffGn random process. As  $B \rightarrow 1$  and  $Q \rightarrow \infty$  the approximation will improve. Reasonable ranges for the parameters  $B$  and  $Q$  are  $B \in [1, 1.2, 0]$  and  $Q \in [10, 20]$  to achieve accurate LRD behaviour. The high frequency term so that the random variables of the ffGn random process are  $N(0, 1)$ -distributed. It has to be chosen according to

$$X_{\text{ff}}(t, H) = \sqrt{1 - \frac{H(2H-1)B^{H-1}}{\Gamma(3-2H)}} G(t) \quad (7)$$

Table 2 LRD accuracy of the ffGn random process ( $B=2$ ,  $Q = 10$ ,  $n_{\text{ffGn}} = 3000000$ )

Desired H	0.6	0.7	0.8	0.95
Measured HRS	0.58	0.68	0.79	0.91

Where again  $G(t)$  is a sequence of independent standard normally distributed random variables. Table 2 shows the resulting Hurst parameter  $H_{RS}$  that is measured using the RS-analysis [Ma Wa 69a, Ma Wa 69b] in comparison with the desired  $H$  that was used as an input parameter for the fFGn process.

Based on the fFGn process  $X_f(t, H)$ , the picture sizes  $X(t)$  are generated using the transformation

$$X(t) = \exp \left[ \sqrt{\ln \left( 1 + \frac{\text{Var}[X_k]}{E[X_k]^2} \right)} X_f(t, H) + \ln E[X_k] - \frac{1}{2} \ln \left( 1 + \frac{\text{Var}[X_k]}{E[X_k]^2} \right) \right] \quad (8)$$

The random variables  $X_k, k \in \{I, P, B\}$  denote the I, P and B picture sizes. For the picture size of the picture with number  $t, t \in \mathbb{N}$  the parameter  $k$  is chosen according to the MPEG GOP pattern. The probability density function  $f_X(x)$  of the picture sizes  $X(t)$  is

$$\text{given by } f_X(x) = \sum_{k \in \{I, P, B\}} p_k f_{X_k}(x) \quad (9)$$

Where  $f_k(x), k \in \{I, P, B\}$  denote the lognormal probability density functions of the three picture types and  $p_k$  is the probability of finding a picture of type  $k$  within the GOP pattern.

The pictures are packetized separately into ATM cells that are assumed to have a payload of 47 byte. All the cells of a picture have to be transmitted in a burst at a specified burst bit rate  $R_x$ , on (the burst bit rate is increased for a single picture if it is too large for the specified rate) or equally spaced over the available 40 ms (in the following this is the default case if no burst bit rate is specified).

#### 4.2 Modeling of Multiplexer

In order to investigate the feasibility and efficiency of the transmission of VBR video traffic in future ATM systems, understanding the behavior of a multiplexer and its performance is essential. The ATM statistical multiplexer is modeled as a queue with deterministic service time  $D$  corresponding to a link bit rate of  $R_L$  Mbit/s and a maximum queue size of  $S$  cells. It is fed by  $N_{Mux}$  VBR MPEG ATM video sources

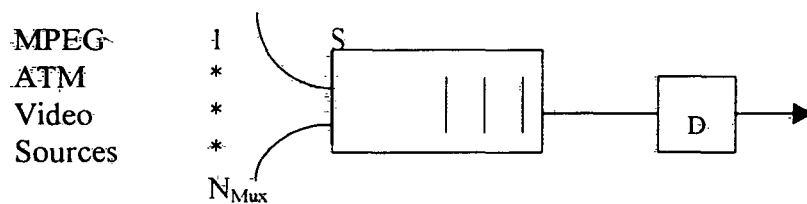


Figure 3 ATM multiplexer model

#### 4.3 CONVERGENCE OF SIMULATION RESULTS

For the stationary evaluation of a technical system via discrete event simulation, it is essential to assess the convergence of the system behavior to a steady-state. The convergence rate is depending on the system architecture and the correlation structure of

the stochastic process used to model randomly performing system components. In the case of an ATM multiplexer with deterministic service time, the random processes used to model the sources decisively determine the convergence rate.

The level of convergence can be described by confidence intervals that give the interval in that the real value of a variable is located with probability (e.g. 95% or 99%), based on the measurements taken during the simulation runs. For a discrete time stochastic process (sequence of random variables)  $\{X_t\}$ ,  $t \in \mathbb{N}$ , the convergence of its mean

$$X_t^{**} = \frac{1}{2} \sum_{i=1}^t X^i \quad (10)$$

is of basic interest. Thus, in the following the convergence of the mean  $X_t^{**}$  of a sequence of standard normally distributed random variables  $\{X_t\}$  with zero mean and unit variance with different correlation structures will be evaluated in terms of its confidence interval size, since this type of process is used to model the scene level within the VBR MPEG video source model. The confidence interval

$$\left[ -z_{\frac{1+\gamma}{2}}, z_{\frac{1+\gamma}{2}} \right] \quad (11)$$

is delimited by the  $\frac{1+\gamma}{2}$  quantile  $z_{\frac{1+\gamma}{2}}$  with

$$z_{\frac{1+\gamma}{2}} = \Phi^{-1}\left(\frac{1+\gamma}{2}\right) \quad (12)$$

where  $\Phi(x)$  denotes the distribution function of the standard normal distribution.

If  $\{X_t\}$  is a sequence of independent identically  $N(0,1)$  distributed random variables,  $X_t^{**}$  is distributed according to  $N(0, 1/t)$  and the confidence interval is given by

$$\left[ -\frac{1}{\sqrt{t} \frac{z_{1+\gamma}}{2}}, \frac{1}{\sqrt{t} \frac{z_{1+\gamma}}{2}} \right] \quad (13)$$

For an exactly second order self-similar standard normally distributed random process, the variance of its mean  $X_t^{**}$  is given by [Cos 84] as  $\text{VAR} [X_t^{**}] = t^{-2(1-H)}$  and so its resulting confidence interval is

$$\left[ -\frac{1}{t^{1-H} \frac{z_{1+\gamma}}{2}}, \frac{1}{t^{1-H} \frac{z_{1+\gamma}}{2}} \right] \quad (14)$$

The mean  $X_f^{**}(t, H)$  of a standard normally distributed fFGn-process is calculated according to

$$X_f^{**}(t, H) = \sum_{k=1}^{n_{fFGn}} W_k \frac{1}{t} \sum_{i=1}^t X_{MG}(i, r_k) + \frac{1}{t} \sum_{i=1}^t X_h(i, H) \quad (15)$$

The random variables of the Markov Gauss process  $X_{MG}(t, r_k)$  (see equation 3) can be restated as a moving average process

$$X_{MG}(t, r_k) = \begin{cases} G_k(1) \\ r_k^{t-1} G_k(1) + \sqrt{1-r_k^2} \sum_{i=2}^t r_k^{i-1} G_k(i) \end{cases} \text{ for } t > 1 \quad (16)$$

Therefore, their sum can be written

$$\sum_{i=1}^t X_{MG} = \sum r_k^{i-1} G_k(i) + \sqrt{1-r_k^2} \sum_{j=i=1}^{t-1} \sum_{j=i+1}^t r_k^{i-1} G_k(j+1) \quad (17)$$

Since the random variables  $G_k(t)$  are mutually independent and  $N(0,1)$ -distributed,

the random variable  $\frac{1}{t} \sum_{i=1}^t X_{MG}(i, r_k)$  is  $N(0, \sigma_{MG}^2(t, r_k))$  distributed

$$\sigma_{MG}^2(t, r_k) = \frac{1}{t} + 2 \sum_{i=1}^{t-1} \frac{t-i}{t^2} r_k \quad (18)$$

Consequently, the weighed sum of all Markov Gauss processes is normally distributed with zero mean and variance  $\sigma_{MG}^2(t)$  with

$$\sigma_{MG}^2(t) = \sum_{k=1}^{n_{MG}} W_k^2 \sigma_{MG}^2(t, r_k) \quad (19)$$

The high frequency term  $X_h(t, H)$  is a sequence of independent identically standard normally distributed random variables with zero mean and variance

$$\sigma_h^2 = 1 - \frac{I(2H-1)B^{H-1}}{\Gamma(3-2H)} \quad (20)$$

Hence  $\frac{1}{t} \sum X_h(i, H)$  is  $N\left(0, \frac{\sigma_h^2}{2}\right)$  - distributed. Finally,  $X_f^{**}(t, H)$  is normally

distributed with

$$E\left[X_f^{**}(t, H)\right] = 0 \text{ and} \quad (21)$$

$$\text{VAR}\left[X_f^{**}(t, H)\right] = \sigma_{MG}^2(t) + \frac{\sigma_h^2}{t} \quad (22)$$

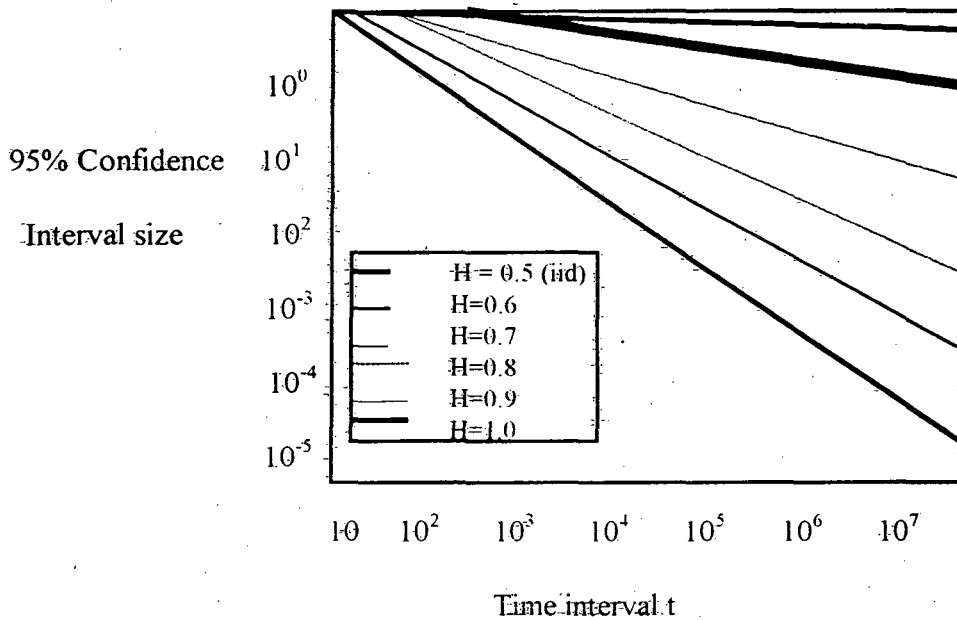
and its confidence interval to the confidence level  $\gamma$  results in

$$\left[ -\sqrt{\sigma_{MG}^2(t) + \frac{\sigma_h^2}{t} z_{\frac{1+\gamma}{2}}}, \sqrt{\sigma_{MG}^2(t) + \frac{\sigma_h^2}{t} z_{\frac{1+\gamma}{2}}} \right] \quad (23)$$

In Figure 4, the 95% confidence interval sizes for standard normally distributed exactly second order self-similar processes with different Hurst parameters  $H$  are depicted according to equation 14. With increasing  $H$  the confidence interval size decreases much slower compared with a sequence of uncorrelated random variables ( $H = 0.5$ ). Consequently, to achieve a desired confidence interval size for the simulation result, the simulation a desired confidence interval size for the simulation results, the simulation program has to run considerably long. E.g. to achieve a confidence-interval size of 10-1 the simulation takes more than four orders of magnitude longer for  $H = 0.8$  compared to  $H = 0.5$ . For  $H \rightarrow 1$  the confidence interval size even stays constant, independent of the



simulation duration. This is an indication for the fact that the boundary of the stationary region of the random process is reached.



**Figure 4. Confidence interval size of  $X_t$  of second order self similar stochastic processes for varying  $H$**

Since the ffGn random process is approximately self-similar, the degree of self-similarity and thus the convergence rate of its mean random process depends on the choice of its parameters  $B$ ,  $Q$  and  $n_{ffGn}$ . Figure 5 shows the confidence interval sizes for an ffGn process with  $H = 0.8$  in comparison with those of an exactly second order self-similar stochastic process and a sequence of uncorrelated  $N(0,1)$ -distributed random variables. The larger the number of desired sample  $n_{ffGn}$  is chosen, the better is the ability of the ffGn process to model the self-similar behavior. Therefore, for larger the slope of the

exactly self-similar random process before it starts to decay as fast as the confidence interval size of the in-correlated random variables.

## 4.5 ATM MULTIPLEXER PERFORMANCE EVALUATION

The performance of an ATM statistical multiplexer is evaluated at the cell, picture and scene level to demonstrate the impact of the source parameters at the different time scales on the multiplexer behavior and the QoS perceived by the video data stream.

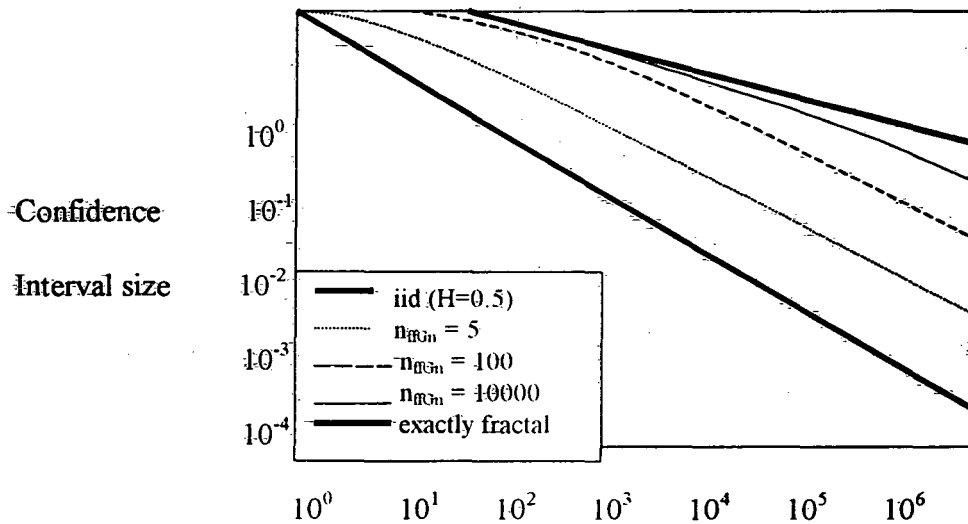
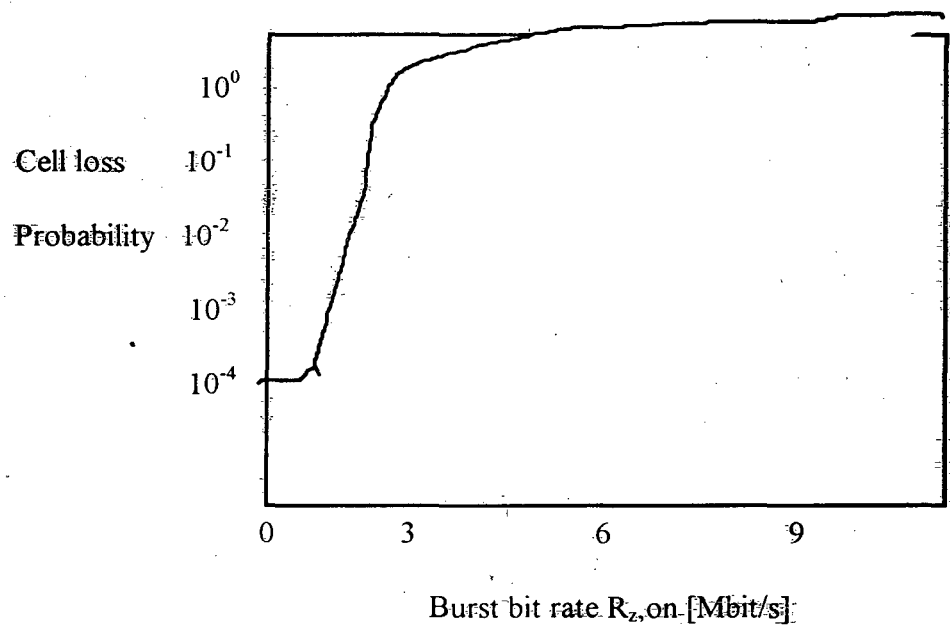


Figure 5 Confidence interval size of the ffGn mean value process.

$$X^{**}(t, H) \quad (H = 0.8, B = 1.1, Q = 20)$$

### 4.4.1 Cell Level

At the cell level, the way in that the ATM cells of the consecutive pictures are sent determines the short-term characteristics of the cell stream. Thus, the packetization



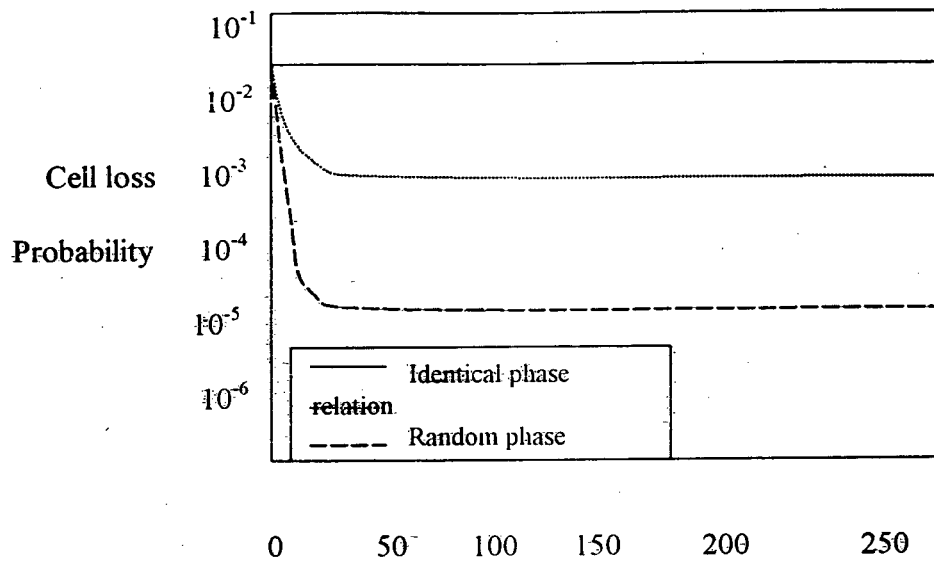
**Figure .6**

**Cell loss probability  $V_z$  depending on burst bit rate**

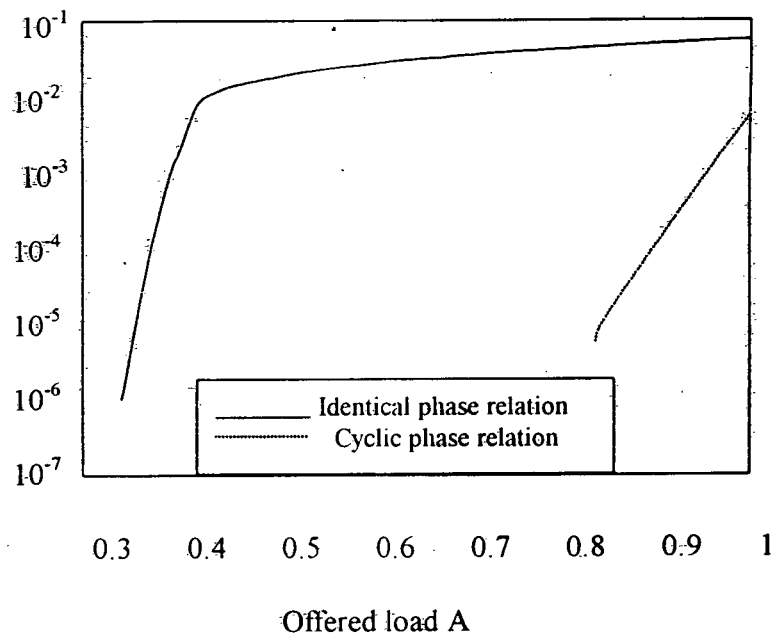
**$R_{z,on}(A=0.82, S= 50, N_{Mux}=40, R_L = 100 \text{ Mbit/s})$**

process has a strong impact on the performance of a multiplexer with a small buffer.

Figure 6 illustrates the influence of the burst bit rate  $R_{z,on}$  on the cell loss probability  $V_z$ ,



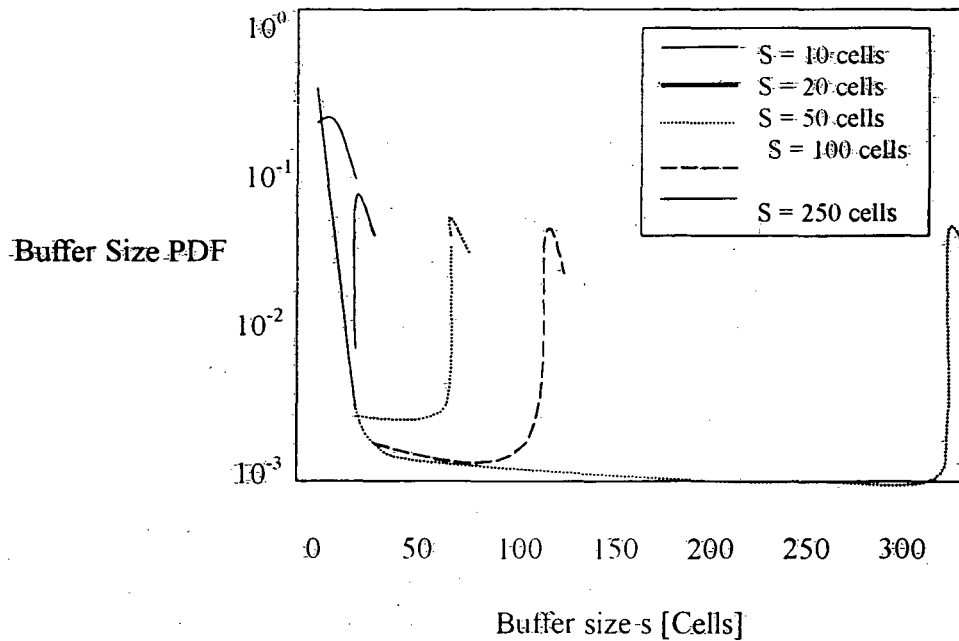
Buffer size S (Cells) (a) Offered load  $A = 0.82$



(b) Buffer Size  $S=100$  Cells

Figure.7 Cell loss probability  $V_2$  depending on the phase Relation of the VBR MPEG video data streams ( $R_L=100\text{Mbit/s}$ ,  $H = 0.5$ )

of a multiplexer with a buffer capacity of 50 ATM cells. It clearly shows the increase in the cell loss rate due to the increase of the on-off characteristic of the source cell streams when most of the frames are sent with a burst bit rate larger than the bit rate necessary for an equally spaced play out.



**Figure 8. Buffer size probability density function (PDF)**

$F_{S_A}(s)$  at cell arrivals ( $A = 0.93$ )

#### 4.4.2 Picture Level

Since the MPEG encoding algorithm uses three different picture types in a periodic GOP pattern, the resulting ATM cell stream reflects the periodic bit rate variations. Therefore, the phase relations of the VBR MPEG ATM cell streams with each other influence the characteristics of the overall input cell stream to the multiplexer and thus its performance. The worst case is that all sources are in phase (identical phase relation), i.e. the start of the GOP patterns of all sources are aligned. The smoother aggregated cell

stream results from a cyclic starting pattern for the GOPs of the different sources (cyclic phase relation).

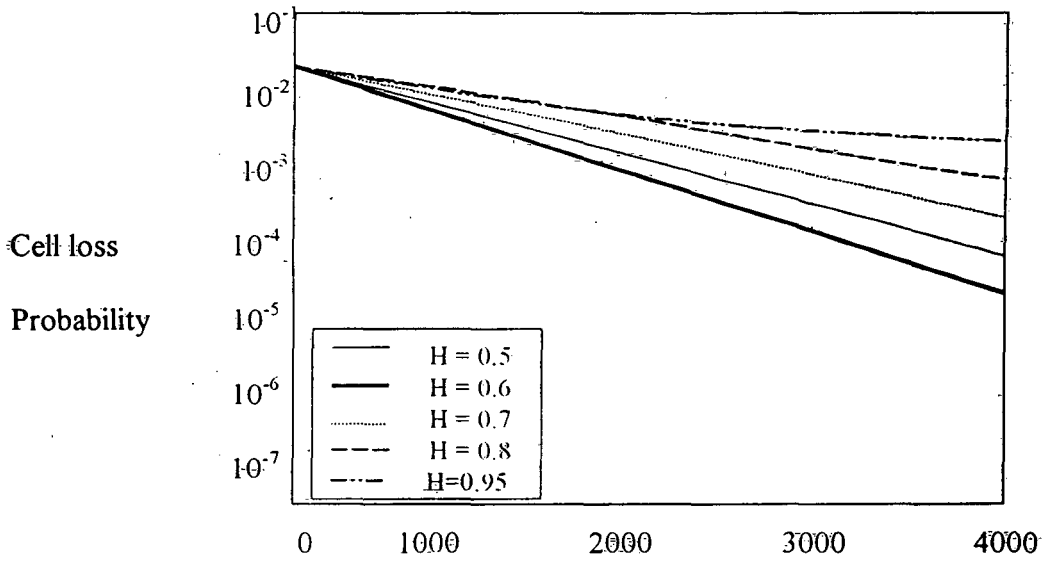
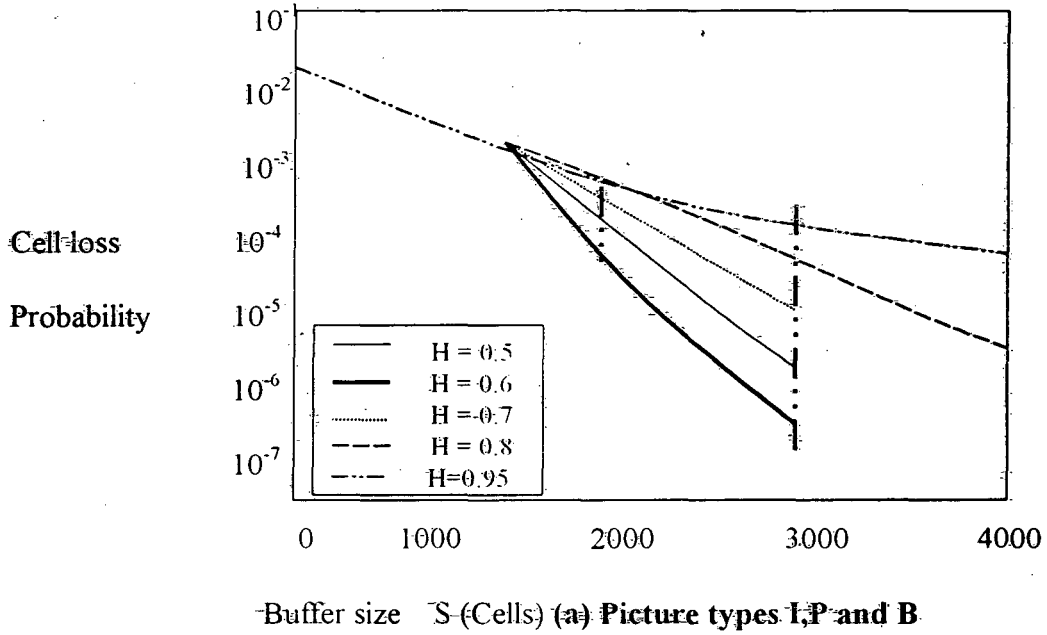
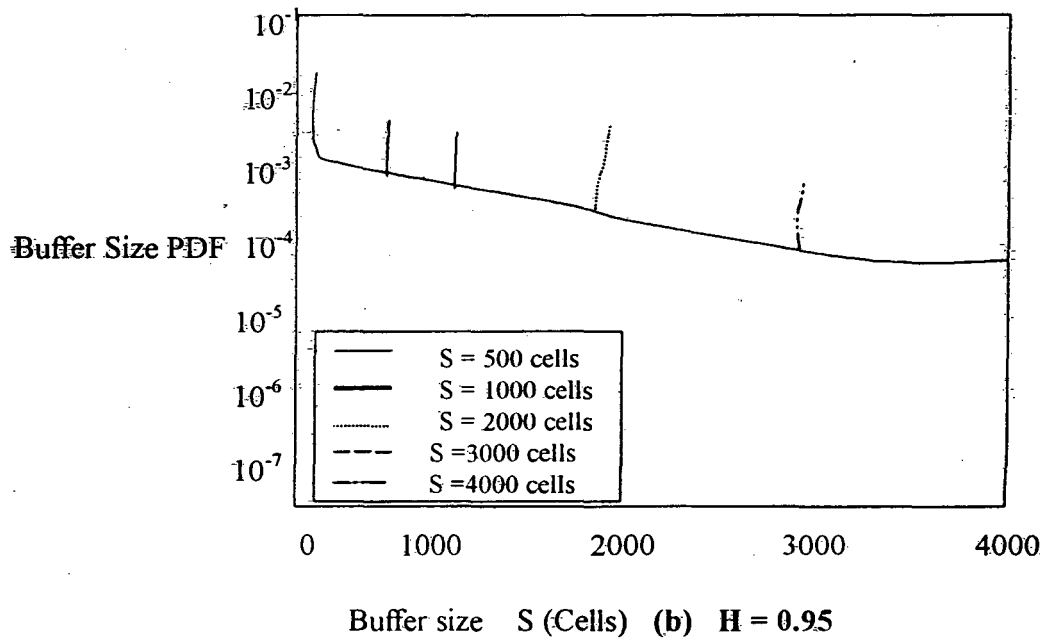
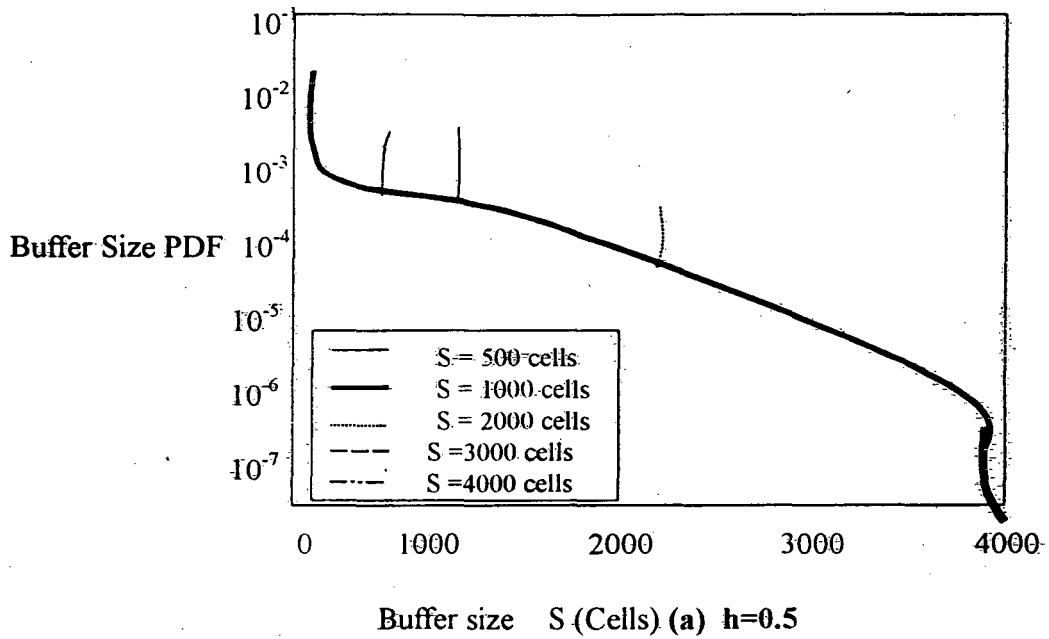


Figure 9 Cell loss probability  $V_L$  depending on the multiplexer Buffer size  $S$  and the Hurst parameter  $H$  ( $A=0.93$ ,  $N_{Mux}=40$ )

Figure 7(a) shows the dependence of the cell loss probability  $V_z$  on the buffer size  $S$  for different phase relations. If it is possible to choose a cyclic phase relation, e.g. by a video on demand system, it is possible to lower the cell loss ratio at the picture level by more than four decades. Even compared with the case of a random phase selection the cyclic phase relation achieves a gain of two decades. To achieve a certain cell loss probability the cyclic phase relation increases the admissible load of the multiplexer by a factor of about 2.4 compared with the worst case relation. The Hurst parameter  $H$  has no effect on the cell loss ratio for buffer-sizes up to several hundred cells. Figure 8 further emphasizes that the Hurst buffer sizes are used. The probability density functions  $f_{S_a}(s)$  of the number of cells  $S_a$  that are already waiting in the buffer when a new cell arrives have a pot-like shape and are almost indistinguishable for a Hurst parameter  $H$  belongs  $[0.5, 1.0]$ . Therefore, the mean cell delay and the cell delay variation as well as the cell loss ratio are independent of the Hurst parameter for medium buffer sizes.

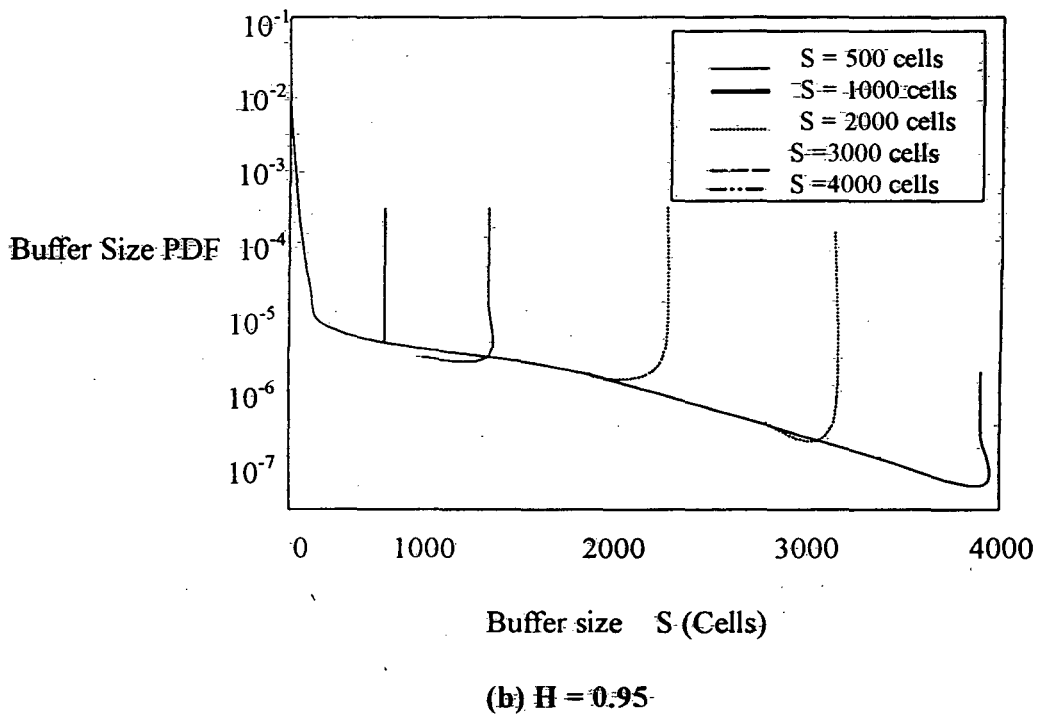
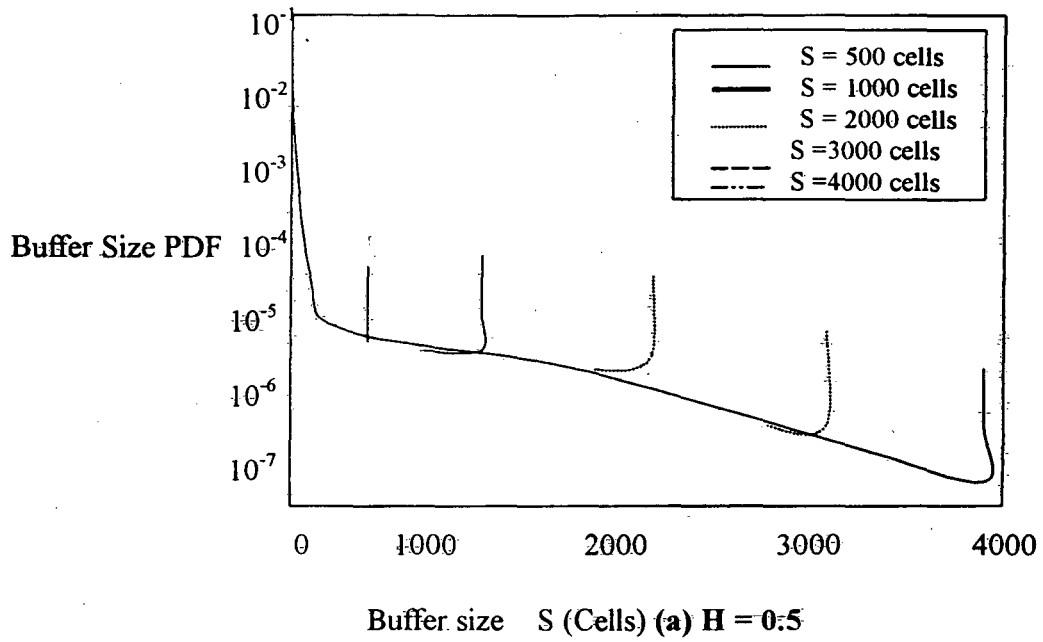
#### 4.4.3 Scene level

The behavior of VBR MPEG video data streams at the scene level is characterized by the Hurst parameter  $H$  that is mainly depending on the video contents. Fig. 9 displays the influence of video data streams with identical probability distribution functions for the picture size but different Hurst-parameters  $H$  on the cell loss probability  $Y$ . For Fig. 9(a) three different pictures types are used whereas in Fig. 9(b) a single lognormal PDF is used with the parameters of all pictures from Table 1. At large buffer-sizes the impact of the Hurst parameter is clearly visible. Video data streams with a high Hurst parameter will cause a cell loss ratio that is orders of magnitude higher than that of stream with a Hurst



**Figure 10: Multiplexer buffer size PDF  $f_{S_A}(s)$  depending on the maximum multiplexer buffer size  $S$  for the picture Types I, P, and B ( $A=0.93$ ,  $N_{Mux}=40$ )**





**Figure.11 Multiplexer buffer size PDF  $f_{S_s}(s)$  depending on the Maximum multiplexer buffer size S for picture with identical PDFs ( $A = 0.93$ ,  $N_{Max} = 40$ )**

parameter close to 0.5. Figure 9 also shows that the use of I, P, and B pictures gives an additional drop in the cell loss ratio and the slope of this decay is steeper compared to case (b).

The corresponding buffer size PDFs  $f_{s_a}(s)$  are depicted in the Fig. 10 and 11 for  $H = 0.5$  and  $H = 0.95$ . They have basically the same pot like.

## Chapter-5 Conclusion

---

There are two alternatives to determine the statistical behavior of a technical system.

They are

- (1) Analysis.
- (2) Simulation.

Analytical techniques are difficult to apply to complex systems or source model. So many performance evaluation studies are conducted using discrete event simulation. In this dissertation we studied the impact of complex stochastic process, that are necessary to realistically characterise VBR MPEG video data streams, on the behaviour of an ATM multiplexer and its evaluation via discrete event simulation. A hierarchical VBR MPEG source model was used to capture the behaviour of such sources at the scene, picture and cell level. The LRD correlation structures of the stochastic processes used to model the scene level of the video data streams have a significant influence on the simulation duration. In the case of second order self-similar random processes the convergence rate of the mean is directly related to the Hurst parameter. The higher the Hurst parameter of a stochastic process is, the slower is the decay of the confidence interval size of its mean. For the approximately self-similar fGn random process, the convergence rate additionally depends on the time span for that the approximation is intended.

The multiplexer behavior in terms of the quality of service parameters perceived by the video data streams mirrors the three levels of the video source model. The cell loss rate decreases at three distinct slopes with increasing buffer size according to the fluctuation of the interarrival time introduced by the packetization process, the picture size variation and the long term scene level activity fluctuations. With small buffers, the cell loss rate increases drastically when the ATM segmentation process clusters the cell of individual pictures in bursts instead of smoothing their transfer as good as possible over the picture duration. At medium buffer sizes, the phase relation of the multiplexed VBR MPEG video data stream is essential. If it is possible to influence the mutual phase relations, the cell loss rate can be lowered by several orders of magnitude. Finally, when using large buffers, the degree of long-range dependence present in the VBR MPEG video data streams determines the cell loss rate. In any cases the cell delay PDFs have a pot-like shape and their asymmetry is depending on the choice of the source model parameters and the multiplexer buffer size.

## References

1. [Enss94] J.Enssle, *Modelling and statistical Multiplexing of VBR MPEG compressed video in ATM networks*, proceedings of the 4th open workshop on high speed networks, Brest, France, September 7-9,1994, pp.59-67.
- 2.[Hurst 51] H.E. Hurst, *long-Term stor age capacity of reservoirs*, trans. Amer.Soc. Civil Eng., Vol.116, pp.770-799, 1951.
3. [kSH 95] M.Krunz, R.Sass, H.hughes, *statistical characteristics and multiplexing of MPEG Streams*, Proceedings of IEEE INFOCOM '95, Boston, April 4-6,1995;pp.455-62.
4. [HTL 94] D.P. Heyman, A.Tabatabai, T.V. Lakshman, *statistical analysis of MPEG-2-Coded VBR Traffic*,proceedings of the sixth international Workshop on Packet Video, Portland, Oregon, September 26-27, 1994,Paper B2.
5. [Enss 95] J.Enssle, *Modelling of short and long term properties of VBR MPEG Compressed video in ATM networks*, Proceedings of the 1995 silicon Valley Networking Conference&Exposition, San Jose, CA, April 5-7,1995,pp. 95-107.

6.MPEG-2, *International Standard ISO/IEC 13818*, ISO/IEC, Switzerland, 1996.

7.S.Dixit and P.Skelly, "MPEG-2 overATM for Video Dial-Tone Networks: Issues and Strategies," *IEEE Network*, Special issue on Digital Interactive Broadband Video-Dial Tone Networks, Sep./oct. 1995, pp.30-40.

8.S.J.Wee, M.O. polley, and W.f. Schreiber, "A Generalized framework for scalable Video Coding."